

ACL 2026

BioNLP 2026 - Shared Tasks

**Proceedings of the 25th Workshop on Biomedical Language
Processing (Shared Tasks)**

July 3-4, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-435-4

Introduction

Deepak Gupta and Dina Demner-Fushman

The BioNLP workshop has a long-standing tradition of offering shared tasks to advance biomedical natural language processing through thorough, novel, and timely topics, evaluation, and community engagement. These shared tasks foster the development of robust and reproducible methods, evaluation frameworks, and datasets, as well as clinically relevant research challenges for the community. This year, we invited SIGBioMed members to submit a description of a shared task proposal. We reviewed the proposals and selected four shared tasks to include in BioNLP 2026. The selected tasks cover a diverse and timely set of research directions, including detecting and labeling medical decisions in ICU discharge summaries (MedExACT), detecting and classifying psychological defense mechanisms in emotional support dialogues (PsyDefDetect), multimodal biomedical question answering with grounded attribution (MedGenVidQA), and clinical skill understanding and continuous perception from clinical video frames (ClinicalSkillQA). Overall, these shared tasks foster the development and benchmarking of innovative methods for multimodal information retrieval, information extraction, question answering, clinical psychology, and mental health in biomedicine.

A total of 39 papers were submitted across the four shared tasks, with participants detailing a range of novel approaches and systems. We accepted 35 papers for inclusion in the proceedings. Each shared task is briefly described below.

MedExACT The first MedExACT shared task focused on detecting and labeling medical decisions in ICU discharge summaries. A medical decision is defined by the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM), which covers ten categories, including Defining Problem, Therapeutic Procedure, and Evaluation. The MedDec dataset contains over 56k expert-annotated decision spans drawn from de-identified ICU discharge summaries in MIMIC-III, supplemented with patient demographics and ten phenotypes (Cancer, Heart Disease, Lung Disease, Chronic Neurologic Dystrophies, Chronic Pain, Alcohol Abuse, Substance Abuse, Obesity, Psychiatric Disorders, Depression). Systems were evaluated for accuracy and robustness at both span and token levels, including stratified analyses by sex, race, English proficiency, and disease type. Baseline models such as RoBERTa indicated the complexity of the task, and participants were supported through expedited access to MedDec via PhysioNet, a public leaderboard, and a Python starter kit.

PsyDefDetect PsyDefDetect invited researchers to tackle a novel challenge at the intersection of Clinical Psychology and Natural Language Processing: detecting and classifying psychological defense mechanisms in emotional support dialogues. Grounded in the clinically validated Defense Mechanism Rating Scales (DMRS) framework, this shared task aimed to advance the understanding of unconscious defensive functioning in text. The benchmark addressed the challenge of capturing subtle linguistic cues of deep-seated psychological mechanisms within highly informal and context-dependent emotional dialogues. This initiative supported research at the intersection of clinical psychology and NLP, aiming to operationalize complex psychological constructs for computational analysis. Participating systems were ranked using Accuracy, Precision, Recall, and F1-score.

MedGenVidQA The recent surge in the availability of online videos has changed the way information and knowledge are acquired. Many people prefer instructional videos to learn how to accomplish a particular task effectively and efficiently through a series of step-by-step procedures. This need is not limited only to general audiences. In professional settings such as healthcare, instructional videos are widely used by physicians and other professionals to learn, review, and standardize procedural workflows. In addition, consumers increasingly seek step-by-step visual explanations to better understand medical procedures and clinical practices. With the advancement of generative models, the medical domain

has also seen progress in medical video understanding, clinical decision support, etc. Toward this, the MedGenVidQA shared task focused on developing systems that utilized generative models to retrieve relevant multimodal (textual and visual) sources and to localize visual answers within medical videos in response to consumer and healthcare professional medical queries. Following earlier editions of medical question answering tasks—MedVidQA 2023, MedVidQA 2024, BioGen 2024, and BioGen 2025—this shared task expanded medical video question answering for both professionals and consumers, with a focus on generative approaches to solving these tasks.

ClinicalSkillQA Multimodal large language models (MLLMs) have the potential to support clinical training and assessment by assisting medical experts in interpreting procedural videos and verifying adherence to standardized workflows. Reliable deployment in these settings requires evidence that models can continuously interpret students' actions during clinical skill assessments, which underpins MLLMs' understanding of clinical skills. Systematically evaluating and improving MLLMs' understanding of clinical skills and their continuous perception in clinical skill assessment scenarios is therefore essential for building reliable and high-impact AI systems for medical education. To address this need, the shared task on medical question answering targeted clinical skill assessment scenarios.

Organizing Committee

MedExACT

Mohamed Elgaar, University of Massachusetts Lowell, USA
Jiali Cheng, University of Massachusetts Lowell, USA
Nidhi Vakil, University of Massachusetts Lowell, USA
Mehrnaz Sadrolashrafi, Beth Israel Deaconess Medical Center, USA
Mitra Mohtarami, Saint Anselm College, USA
Adrian Wong, Beth Israel Deaconess Medical Center, USA
Hadi Amiri, University of Massachusetts Lowell, USA
Leo A. Celi, Massachusetts Institute of Technology, USA

PsyDefDetect

Hongbin Na, University of Technology Sydney, Australia
Zimu Wang, Xi'an Jiaotong-Liverpool University, China
Zhaoming Chen, University of Utah, USA
Yining Hua, Harvard University, USA
Rena Gao, The University of Melbourne, Australia
Kailai Yang, The University of Manchester, United Kingdom
Ling Chen, University of Technology Sydney, Australia
Wei Wang, Xi'an Jiaotong-Liverpool University, China
Shaoxiong Ji, ELLIS Institute Finland & University of Turku, Finland
John Torous, Harvard University, USA
Sophia Ananiadou, The University of Manchester, United Kingdom

MedGenVidQA

Deepak Gupta, National Library of Medicine, National Institutes of Health (NIH), USA
Dina Demner-Fushman, National Library of Medicine, National Institutes of Health (NIH), USA

ClinicalSkillQA

Xiyang Huang, Wuhan University, China
Yihuai Xu, Wuhan University, China
Zhiyuan Chen, Wuhan University, China
Keying Wu, Wuhan University, China
Jiayi Xiang, Wuhan University, China
Buzhou Tang, Harbin Institute of Technology, China
Renxiong Wei, Zhongnan Hospital of Wuhan University, China
Yanqing Ye, Zhongnan Hospital of Wuhan University, China
Jinyu Chen, Zhongnan Hospital of Wuhan University, China
Cheng Zeng, Wuhan University, China
Min Peng, Wuhan University, China
Qianqian Xie, Wuhan University, China
Sophia Ananiadou, The University of Manchester, United Kingdom

Program Committee

Chairs

Deepak Gupta, National Library of Medicine, NIH
Dina Demner-Fushman, National Library of Medicine, NIH

Reviewers

Hadi Amiri, University of Massachusetts Lowell
Jiali Cheng, Northeastern University
Mohamed Elgaar, University of Massachusetts Lowell
Deepak Gupta, National Library of Medicine, NIH
Yining Hua, Harvard University
Xiyang Huang, Wuhan University
Nidhi Vakil, University of Massachusetts Lowell
Hilmi Demirhan, University of North Carolina Wilmington
Gulustan Dogan, University of North Carolina Wilmington
Basil Ebinesar, Purdue University Northwest
Pedram Golnari, Case Western Reserve University
Watcharitpol Sermsrisuwan, Department of Biomedical Engineering, Faculty of Engineering, Mahidol University
Hongbin Na, University of Technology Sydney

Table of Contents

<i>NM at CRF Filling 2026: A Two-Stage LLM Pipeline for Clinical CRF Population</i> Niccolò Morabito	1
<i>VerbaNexAI at ClinicalSkillQA: From Visual Evidence to Procedural Order A Two-Stage Generative Vision-Language Framework for ClinSkillQA</i> Andrea Menco Tovar, Jairo E. Serrano, Edwin Puertas and Juan Carlos Martinez-Santos.....	6
<i>zzucs at PsyDefDetect: Bridging Long-Tail Imbalance and Clinical Rubrics for DMRS Defense-Level Detection</i> Bin Huang, Liuyuan Su, Kaixuan Yuan, Guanghui Zhao, Shixin Zhang and Kunli Zhang	13
<i>zzunlp at ClinicalSkillQA: Perceive-and-Plan with Decomposed In-Context Learning and Saliency-Guided Perception for Clinical Skill Keyframe Reordering</i> Bin Huang, Yi Luo, Zhontian Hua, Guanghui Zhao, Kaixuan Yuan and Kunli Zhang.....	24
<i>DLNLP at ClinicalSkillQA: EvidenceFlow for Structured Zero-Shot Clinical Keyframe Ordering</i> Kexin Li, Zhekun Wang, Yiran Wang and Di Zhao.....	33
<i>UTS at PsyDefDetect: Multi-Agent Councils and Absence-Based Reasoning for Defense Mechanism Classification</i> Dima Galat and Marian Rizoiu	38
<i>Otter at MedExAct2026: Diverse Encoder Ensemble for Medical Decision Span Detection</i> Lalita Lowphansirikul and Piyalitt Ittichaiwong	47
<i>Eraserhead at PsyDefDetect: Prompt Design and Class Rebalancing for Psychological Defense Mechanism Detection</i> Muhammad Abu Horaira, Mehreen Rahman and Nahian Chowdhury	54
<i>Nürnberg NLP at PsyDefDetect: Multi-Axis Voter Ensembles for Psychological Defence Mechanism Classification</i> Philipp Steigerwald, Eric Rudolph and Jens Albrecht	59
<i>Neural Nexus at PsyDefDetect: Fine-Tuning RoBERTa with Focal Loss and Role-Tagged Dialogue History for Defense Level Detection</i> Subhrajyoti Basu	66
<i>ELiRF-UPV@MedExACT 2026: Dynamic Section Conditioning for Medical Decision Span Detection in Discharge Summaries</i> Vicent Ahuir, Lluís Hurtado and María Castro-Bleda	71
<i>VISHC at PsyDefDetect: Mitigating Data Scarcity in Psychological Defense Classification with Context-Aware Synthetic Augmentation</i> Hoang-Thuy-Duong Vu, Quoc-Cuong Pham and Huy-Hieu Pham	77
<i>Diverse Transformer Ensemble with Majority Voting for Medical Decision Extraction at MedExACT 2026</i> Rishik Kondadadi	87
<i>FBK-NLP at ClinSkill QA 2026: Improving Temporal Reasoning via Keypoint-Augmented Inputs</i> Pedro Gabriel Campana, Alberto Lavelli and Bernardo Magnini	92

<i>transformer_1376 at PsyDefDetect: A QLoRA-Based Generative Framework for Context-Aware Psychological Defense Mechanism Detection</i>	
Pritha Saha, Shuvodwip Saha and Anik Mahmud Shanto	99
<i>Explainers at PsyDefDetect: Hierarchical Prompting and Representation-Based Classification for Psychological Defenses</i>	
Liudmila Babakova, Christopher Luongo-Vazquez and Ilia Stepin	104
<i>PerceptionLab at PsyDefDetect: Overcoming Extreme Response Bias in LLMs via Rubric-Grounded Retrieval and Supervised Clinical Reasoning Distillation for Fine-Grained Ordinal Classification</i>	
Tamjid Fahim, Syed Johan and Saad Bin Maksud	109
<i>LinguUTics at PsyDefDetect: Iterative Imbalance-Aware Fine-tuning of Qwen3-8B for Psychological Defense Mechanism Classification</i>	
Shefayat Adib, Ahmed Sani, Md Hasibur Alif and Ajwad Abrar	123
<i>TONI-NLP at PsyDefDetect: Defense Mechanism Detection via LLM-based Ensemble Methods</i>	
Durjoy Paul, Arshitha Basavaraj, Callum Chan, Veronica Perez-Rosas, Diana Inkpen, Francisco Pereira and Juan Antonio Lossio-Ventura	132
<i>Zero-Shot, Fine-Tuned, and Retrieval-Augmented Extraction of Clinical Decisions with Corpus Boundary Diagnostics</i>	
Mohammed Alliheedi, Robert Mercer, Anemily Machina, Sudipta Roy, Yetian Wang and Xindi Wang	141
<i>CASPAR: A Context-Aware Span Refinement Approach for Decision Support</i>	
Jing Tao, Amir Eskandari and Farhana Zulkernine	146
<i>KCL-Cogstack at PsyDefDetect: A Hierarchical Approach to Detecting Defense Mechanisms in Supportive Dialogue</i>	
Shubham Agarwal, Thomas Searle and Richard Dobson	155
<i>DAL Team at PsyDefDetect: From Supervised Encoders to Hierarchical LLM-RAG for Psychological Defense Detection</i>	
Anh Chu, Luong Tran, Dat Do, Phuong Mai, Quynh Le and Cat Can	164
<i>CUAMC @ MedExACT 2026: Robust Ensemble Voting for Fair Medical Decision Extraction</i>	
William Baumgartner and Lisa Schilling	171
<i>LAMAR at MedExACT 2026: Agreement-Driven Large Language Model Ensembles for Clinical Decision Extraction from Discharge Summaries</i>	
Monrada Chiewhawan, Keetawan Limaroon and Titipat Achakulvisut	179
<i>CS_Metro at PsyDefDetect: Detecting Psychological Defense Mechanisms in Mental Health Dialogues with Summarization-Enhanced Transformer Ensembles</i>	
Oarisa Rebayet, Radiul Walee, Symom Hossain Shohan, Kawsar Ahmed and Mohammed Moshiul Hoque	191
<i>Sparse Category Routing and Fairness-Aware Optimization for Medical Decision Extraction</i>	
Ahmed Elshehaby, Mohamed Abdalla and Youssef Mohamed	201
<i>AlienAnnotators at PsyDefDetect: What Lies Between the Lines: Probing Lightweight Open-Source LLMs for Psychological Defense Mechanism Detection</i>	
Siam Karip and Nahid Hossain	213

<i>Team Aurum at MedExACT 2026@ACL: Data Augmentation and Clinical Longformer Fine-Tuning for Medical Decision Extraction</i>	
Jyoti Kumari, Vinay Ulli and Anindita Mondal	224
<i>NJUST-KMG at MedGenVidQA 2026: Cascade Multi-modal Alignment with Gaussian Soft Priors for Medical Visual Answer Localization</i>	
Jinglong Li and Yang Yang	229
<i>LAMAR-2 at MedGenVidQA 2026: Visual Answer Localization in Medical Videos via Multimodal LLM and Context-Augmented Prompting</i>	
Watcharitpol Sermsrisuwan, Nopporn Lekuthai, Seksan Yoadsanit and Titipat Achakulvisut .	233
<i>Varja-Dominators at MedGenVidQA 2026: Hybrid Video and Document Retrieval using PubMedBERT, T5 Query Expansion, and Cross-Encoder Re-Ranking</i>	
Pratik Dhaktode, Suhani Bighane and Anupama Phakatkar	243
<i>Pride-Boiler at MedGenVidQA 2026: LLM-Augmented BM25 Retrieval with Corrective Self-Verification for Biomedical Evidence Retrieval</i>	
Basil Ebinesar, Keyuan Jiang, Charansai Maddineni and Ashok Raja	248
<i>Seahawk at MedGenVidQA 2026: LLM Segment-Range Selection for Medical Visual Answer Localization</i>	
Xiaotian Tian and Gulustan Dogan	257
<i>UNCC at MedGenVidQA 2026: Structured Temporal Grounding for Medical Video Question Answering</i>	
Hilmi Demirhan and Wlodek Zadrozny	262

Program

Saturday, July 4, 2026

09:00 - 09:10 *Opening remarks*

09:10 - 09:20 *MedExACT Overview recapitulation*

09:20 - 09:50 *Shared Task Papers*

CUAMC @ MedExACT 2026: Robust Ensemble Voting for Fair Medical Decision Extraction

William Baumgartner and Lisa Schilling

Otter at MedExAct2026: Diverse Encoder Ensemble for Medical Decision Span Detection

Lalita Lowphansirikul and Piyalitt Ittichaiwong

LAMAR at MedExACT 2026: Agreement-Driven Large Language Model Ensembles for Clinical Decision Extraction from Discharge Summaries

Monrada Chiewhawan, Keetawan Limaroon and Titipat Achakulvisut

09:50 - 10:00 *Talk by Moustafa Yehia Hassan — The Divergence Hypothesis: Unmasking Lexical Interference and Label Bias in Mental Health NLP*

10:00 - 10:10 *Talk by Sanya Bathla Taneja — CNET: Context Engineering Framework for Normalization of Clinical Trial Procedures*

10:10 - 10:20 *Talk by Niloofar Arazkhani — KALIMBA: Knowledge-Assisted Literature Mining for Biological Interaction Analysis*

10:20 - 10:30 *Talk by Krushil Bhojani — BioRAG: A Systematic Ablation Study of Retrieval Strategies for Biomedical Question Answering*

10:30 - 11:00 *Coffee Break*

11:00 - 11:10 *MedGenVidQA Overview recap*

Varja-Dominators at MedGenVidQA 2026: Hybrid Video and Document Retrieval using PubMedBERT, T5 Query Expansion, and Cross-Encoder Re-Ranking

Pratik Dhaktode, Suhani Bighane and Anupama Phakatkar

Pride-Boiler at MedGenVidQA 2026: LLM-Augmented BM25 Retrieval with Corrective Self-Verification for Biomedical Evidence Retrieval

Basil Ebinesar, Keyuan Jiang, Charansai Maddineni and Ashok Raja

Saturday, July 4, 2026 (continued)

Seahawk at MedGenVidQA 2026: LLM Segment-Range Selection for Medical Visual Answer Localization

Xiaotian Tian and Gulustan Dogan

UNCC at MedGenVidQA 2026: Structured Temporal Grounding for Medical Video Question Answering

Hilmi Demirhan and Wlodek Zadrozny

11:50 - 12:00 *Talk by Kadir Bulut Ozler — Analyzing Prompt Design Choices in Biomedical Information Extraction for Low-Resource Languages*

12:00 - 12:10 *Talk by Santiago Martínez and Lina Gomez — MeSHClass-ES and AnatEM-ES: Open Resources for Spanish Biomedical NLP*

12:10 - 12:20 *Talk by Lana Yeganova — BioTopicXplor: A Web Tool for Interactive Exploration of PubMed Literature through Reproducible Topics*

12:20 - 12:30 *Talk by Won Kim — Learning to Combine AI Annotations for Improved Biomedical Relevance Labeling*

12:30 - 14:00 *Lunch*

14:00 - 14:10 *PsyDefDetect Overview recap*

14:10 - 14:40 *Shared Task Papers*

UTS at PsyDefDetect: Multi-Agent Councils and Absence-Based Reasoning for Defense Mechanism Classification

Dima Galat and Marian Rizoiu

KCL-Cogstack at PsyDefDetect: A Hierarchical Approach to Detecting Defense Mechanisms in Supportive Dialogue

Shubham Agarwal, Thomas Searle and Richard Dobson

Nürnberg NLP at PsyDefDetect: Multi-Axis Voter Ensembles for Psychological Defence Mechanism Classification

Philipp Steigerwald, Eric Rudolph and Jens Albrecht

14:40 - 14:50 *Talk by Olga Pelloni — Evaluation of Multilingual Text Simplification for the Mental Health Domain: Exploring Small Language Models*

14:50 - 15:00 *Talk by Vishwaa Shah — A Multi-View Framework for Cross-Domain Nutrition Misinformation Detection in Social Media*

Saturday, July 4, 2026 (continued)

15:00 - 15:10 *Talk by Vishwaa Shah — Small LLMs for Biomedical Claim Verification: Cost-Effective Fine-Tuning, Structural Dataset Shortcuts, and Cross-Domain Generalization*

15:10 - 15:20 *Talk by Francois Remy — Diagnosable ColBERT: Debugging Late-Interaction Retrieval Models Using a Learned Latent Space as Reference*

15:20 - 15:30 *Talk by Hui Wang — CrossDDI: Cross-Source Evidence-Grounded Drug-Drug Interaction Verification*

15:30 - 16:00 *Coffee Break*

16:00 - 16:10 *ClinicalSkillQA Overview recap*

16:10 - 16:30 *Shared Task Papers*

zzunlp at ClinicalSkillQA: Perceive-and-Plan with Decomposed In-Context Learning and Saliency-Guided Perception for Clinical Skill Keyframe Reordering
Bin Huang, Yi Luo, Zhontian Hua, Guanghui Zhao, Kaixuan Yuan and Kunli Zhang

FBK-NLP at ClinSkill QA 2026: Improving Temporal Reasoning via Keypoint-Augmented Inputs
Pedro Gabriel Campana, Alberto Lavelli and Bernardo Magnini

16:30 - 16:40 *Talk by Yixian Ma — VERICITE: Evaluating Sentence-Level Citation Faithfulness in Retrieval-Augmented Medical Question Answering*

16:40 - 16:50 *Talk by Rez Samantha Floresca — Forgotten Words: Benchmarking NeoBERT for Dementia Detection in Low-Resource Conversational Filipino and English Speech*

16:50 - 17:00 *Talk by Brandon Colelough — Towards Grounded Hallucination Definitions for Biomedical Question Answering with Reproducible Examples*

17:00 - 18:00 *Poster presentation by Mahule Roy — Towards Unified Factuality Evaluation for Biomedical QA and Summarization: Aligning Metrics with Clinical Use-Cases*

17:00 - 18:00 *Poster presentation by Yikun Han — When Evidence Conflicts: Uncertainty and Order Effects in Retrieval-Augmented Biomedical Question Answering*

17:00 - 18:00 *Shared Task Posters*

VerbaNexAI at ClinicalSkillQA: From Visual Evidence to Procedural Order A Two-Stage Generative Vision-Language Framework for ClinSkillQA
Andrea Menco Tovar, Jairo E. Serrano, Edwin Puertas and Juan Carlos Martinez-Santos

Saturday, July 4, 2026 (continued)

zzucs at PsyDefDetect: Bridging Long-Tail Imbalance and Clinical Rubrics for DMRS Defense-Level Detection

Bin Huang, Liuyuan Su, Kaixuan Yuan, Guanghui Zhao, Shixin Zhang and Kunli Zhang

zzunlp at ClinicalSkillQA: Perceive-and-Plan with Decomposed In-Context Learning and Saliency-Guided Perception for Clinical Skill Keyframe Reordering

Bin Huang, Yi Luo, Zhontian Hua, Guanghui Zhao, Kaixuan Yuan and Kunli Zhang

DLNLP at ClinicalSkillQA: EvidenceFlow for Structured Zero-Shot Clinical Keyframe Ordering

Kexin Li, Zhekun Wang, Yiran Wang and Di Zhao

UTS at PsyDefDetect: Multi-Agent Councils and Absence-Based Reasoning for Defense Mechanism Classification

Dima Galat and Marian Rizoiu

Otter at MedExAct2026: Diverse Encoder Ensemble for Medical Decision Span Detection

Lalita Lowphansirikul and Piyalitt Ittichaiwong

Eraserhead at PsyDefDetect: Prompt Design and Class Rebalancing for Psychological Defense Mechanism Detection

Muhammad Abu Horaira, Mehreen Rahman and Nahian Chowdhury

Nürnberg NLP at PsyDefDetect: Multi-Axis Voter Ensembles for Psychological Defence Mechanism Classification

Philipp Steigerwald, Eric Rudolph and Jens Albrecht

Neural Nexus at PsyDefDetect: Fine-Tuning RoBERTa with Focal Loss and Role-Tagged Dialogue History for Defense Level Detection

Subhrajyoti Basu

ELiRF-UPV@MedExACT 2026: Dynamic Section Conditioning for Medical Decision Span Detection in Discharge Summaries

Vicent Ahuir, Lluís Hurtado and María Castro-Bleda

VISHC at PsyDefDetect: Mitigating Data Scarcity in Psychological Defense Classification with Context-Aware Synthetic Augmentation

Hoang-Thuy-Duong Vu, Quoc-Cuong Pham and Huy-Hieu Pham

Diverse Transformer Ensemble with Majority Voting for Medical Decision Extraction at MedExACT 2026

Rishik Kondadadi

Saturday, July 4, 2026 (continued)

FBK-NLP at ClinSkill QA 2026: Improving Temporal Reasoning via Keypoint-Augmented Inputs

Pedro Gabriel Campana, Alberto Lavelli and Bernardo Magnini

transformer_1376 at PsyDefDetect: A QLoRA-Based Generative Framework for Context-Aware Psychological Defense Mechanism Detection

Pritha Saha, Shuvodwip Saha and Anik Mahmud Shanto

Explainers at PsyDefDetect: Hierarchical Prompting and Representation-Based Classification for Psychological Defenses

Liudmila Babakova, Christopher Luongo-Vazquez and Ilia Stepin

PerceptionLab at PsyDefDetect: Overcoming Extreme Response Bias in LLMs via Rubric-Grounded Retrieval and Supervised Clinical Reasoning Distillation for Fine-Grained Ordinal Classification

Tamjid Fahim, Syed Johan and Saad Bin Maksud

LinguUTics at PsyDefDetect: Iterative Imbalance-Aware Fine-tuning of Qwen3-8B for Psychological Defense Mechanism Classification

Shefayat Adib, Ahmed Sani, Md Hasibur Alif and Ajwad Abrar

TONI-NLP at PsyDefDetect: Defense Mechanism Detection via LLM-based Ensemble Methods

Durjoy Paul, Arshitha Basavaraj, Callum Chan, Veronica Perez-Rosas, Diana In-ken, Francisco Pereira and Juan Antonio Lossio-Ventura

Zero-Shot, Fine-Tuned, and Retrieval-Augmented Extraction of Clinical Decisions with Corpus Boundary Diagnostics

Mohammed Alliheedi, Robert Mercer, Anemily Machina, Sudipta Roy, Yetian Wang and Xindi Wang

CASPAR: A Context-Aware Span Refinement Approach for Decision Support

Jing Tao, Amir Eskandari and Farhana Zulkernine

KCL-Cogstack at PsyDefDetect: A Hierarchical Approach to Detecting Defense Mechanisms in Supportive Dialogue

Shubham Agarwal, Thomas Searle and Richard Dobson

DAL Team at PsyDefDetect: From Supervised Encoders to Hierarchical LLM-RAG for Psychological Defense Detection

Anh Chu, Luong Tran, Dat Do, Phuong Mai, Quynh Le and Cat Can

CUAMC @ MedExACT 2026: Robust Ensemble Voting for Fair Medical Decision Extraction

William Baumgartner and Lisa Schilling

Saturday, July 4, 2026 (continued)

LAMAR at MedExACT 2026: Agreement-Driven Large Language Model Ensembles for Clinical Decision Extraction from Discharge Summaries

Monrada Chiewhawan, Keetawan Limaroon and Titipat Achakulvisut

CS_Metro at PsyDefDetect: Detecting Psychological Defense Mechanisms in Mental Health Dialogues with Summarization-Enhanced Transformer Ensembles

Oarisa Rebayet, Radiul Walee, Symom Hossain Shohan, Kawsar Ahmed and Mohammed Moshiul Hoque

Sparse Category Routing and Fairness-Aware Optimization for Medical Decision Extraction

Ahmed Elshehaby, Mohamed Abdalla and Youssef Mohamed

AlienAnnotators at PsyDefDetect: What Lies Between the Lines: Probing Lightweight Open-Source LLMs for Psychological Defense Mechanism Detection

Siam Karip and Nahid Hossain

Team Aurum at MedExACT 2026@ACL: Data Augmentation and Clinical Long-former Fine-Tuning for Medical Decision Extraction

Jyoti Kumari, Vinay Ulli and Anindita Mondal

NJUST-KMG at MedGenVidQA 2026: Cascade Multi-modal Alignment with Gaussian Soft Priors for Medical Visual Answer Localization

Jinglong Li and Yang Yang

LAMAR-2 at MedGenVidQA 2026: Visual Answer Localization in Medical Videos via Multimodal LLM and Context-Augmented Prompting

Watcharitpol Sermsrisuwan, Nopporn Lekuthai, Seksan Yoadsanit and Titipat Achakulvisut

Varja-Dominators at MedGenVidQA 2026: Hybrid Video and Document Retrieval using PubMedBERT, T5 Query Expansion, and Cross-Encoder Re-Ranking

Pratik Dhaktode, Suhani Bighane and Anupama Phakatkar

Pride-Boiler at MedGenVidQA 2026: LLM-Augmented BM25 Retrieval with Corrective Self-Verification for Biomedical Evidence Retrieval

Basil Ebinesar, Keyuan Jiang, Charansai Maddineni and Ashok Raja

Seahawk at MedGenVidQA 2026: LLM Segment-Range Selection for Medical Visual Answer Localization

Xiaotian Tian and Gulustan Dogan

UNCC at MedGenVidQA 2026: Structured Temporal Grounding for Medical Video Question Answering

Hilmi Demirhan and Wlodek Zadrozny

Saturday, July 4, 2026 (continued)

NM at CRF Filling 2026: A Two-Stage LLM Pipeline for Clinical CRF Population

Niccolò Morabito

morabito.niccolo@gmail.com

Abstract

This paper describes our participation in the CRF Filling Shared Task 2026, which aims to automatically populate a predefined Case Report Form (CRF) from clinical notes describing patients with dyspnea. We propose a two-stage pipeline based on large language models (LLMs). In the first stage, a few-shot prompted LLM extracts candidate CRF fields from the clinical note and outputs them in a structured JSON format. In the second stage, a separate LLM verifies each extracted field against the original note and removes predictions that are not supported by explicit textual evidence. This verification step aims to reduce false positives generated during extraction. Experiments on the development set show that the verification stage significantly reduces unsupported predictions while preserving most correct extractions, resulting in improved macro F1. On the official test set, the proposed system achieves a macro F1 score of 0.56 for both English and Italian. These results indicate that separating extraction and verification can balance recall-oriented extraction with precision-oriented validation in CRF population tasks.

1 Introduction

Automatically converting unstructured clinical narratives into structured data is an important step for enabling large-scale clinical research and clinical decision support systems. In many medical workflows, relevant patient information is recorded in free-text notes, making large-scale analysis difficult without structured representations.

In this context, the CRF Filling Shared Task (Ferrazzi et al., 2026b) focuses on populating a predefined Case Report Form (CRF) from clinical notes describing patients presenting with dyspnea.

To address this problem, we propose a two-stage pipeline combining extraction with conservative verification. In the first stage, a few-shot large language model (LLM) extracts candidate CRF fields

directly from the clinical note. In the second stage, a separate LLM verifies each extracted field against the original note, removing predictions that are not supported by explicit textual evidence. This design aims to balance coverage and precision by combining the generative capabilities of LLMs with a verification step that reduces unsupported predictions.

2 Task Description

The shared task consists of automatically populating a CRF from free-text clinical notes describing patients presenting with dyspnea. The CRF contains a fixed set of 134 clinical items, each corresponding to a specific variable such as symptoms, clinical findings, or laboratory measurements. For each item, systems must select one value from a predefined list of valid options based solely on the information contained in the clinical note. The same CRF structure is used for all notes.

A key challenge of the task is the extreme sparsity of annotations. In many cases, the clinical note does not contain information relevant to most CRF items. When an item cannot be populated from the note, it is assigned the value *unknown*. As a result, approximately 95% of the CRF fields are labeled as *unknown* across the dataset, requiring systems to extract information only when explicit textual evidence is present.

The training data combines multiple sources of supervision, including a small set of 10 gold-standard pairs of clinical notes and filled Dyspnea CRFs (Kaczmarek et al., 2026), a collection of semi-automatically annotated CRFs for other medical conditions (Ferrazzi et al., 2025), and a set of 2,667 unannotated clinical notes describing patients with dyspnea (Ferrazzi et al., 2026a). In our experiments, we only use the 10 gold-standard pairs as few-shot examples for prompting the extraction model (see 4.1).

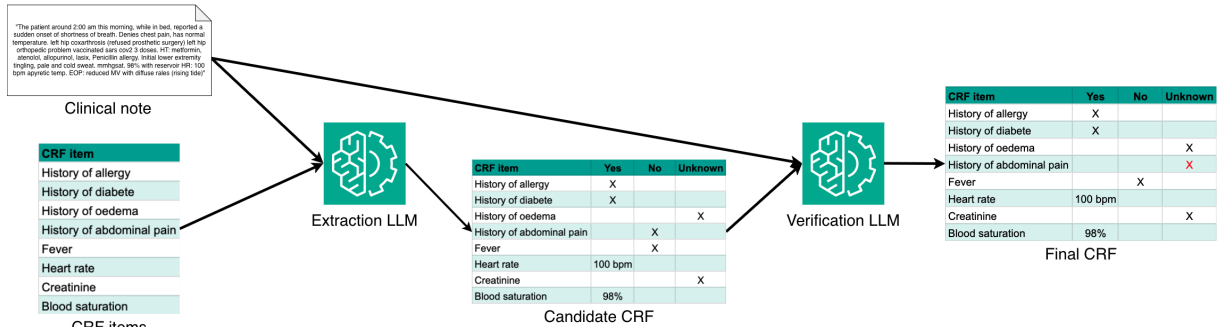


Figure 1: Overview of the two-stage pipeline for CRF extraction and verification. The first stage extracts candidate CRF fields from clinical notes using a few-shot LLM, and the second stage verifies each field against the original note to remove unsupported predictions to generate the final CRF.

System	Model	TP	FP	FN	F1
Extraction	Gemini 3 Flash	278	420	83	0.584
Extraction	LLaMA4 Maverick 17B	179	131	170	0.578
Extraction + Verification	Gemini 3 Flash + LLaMA4 Maverick 17B	274	373	88	0.612

Table 1: Performance comparison on the English development set.

For evaluation, the organizers provide a development set of 80 gold-standard note-CRF pairs and a test set of 200 gold-standard pairs (Kaczmarek et al., 2026). The task is released in two languages, English and Italian, with parallel datasets available for both languages.

3 System Overview

Our system follows a two-stage pipeline designed to balance recall and precision when extracting structured information from clinical notes. The first stage performs CRF extraction using a few-shot prompted LLM. Given a clinical note, the model generates a candidate CRF containing only the CRF fields that are explicitly supported by the text (see Figure 1).

Preliminary experiments showed that this extraction stage tends to over-predict fields, producing a relatively high number of false positives. To mitigate this issue, we introduce a second verification stage. In this step, a different LLM receives both the original clinical note and the CRF generated in the first stage and verifies whether each predicted field is supported by explicit textual evidence. Fields that cannot be directly justified by the text are removed from the output of the final CRF.

This two-stage design allows the system to combine the strong extraction capabilities of large generative models with a more conservative verifica-

tion mechanism aimed at reducing unsupported predictions.

4 Method

This section describes the two stages of the proposed pipeline in detail.

4.1 CRF Extraction with Few-Shot LLM

The extraction stage formulates CRF completion as a structured information extraction problem solved through few-shot prompting of a large language model. Few-shot learning allows a model to adapt to a new task by conditioning on a small number of input–output examples provided directly in the prompt.

The shared task provides a very small set of gold-standard annotated examples (10 note–CRF pairs). These examples can be used as demonstrations in the prompt to guide the model toward the desired behavior. In our experiments, we include three examples in the prompt, which provided a good balance between task guidance and prompt length.

Given a clinical note, the model receives a system prompt describing the extraction task and specifying strict constraints on how CRF fields should be produced. The model is prompted once per clinical note to generate a structured JSON object containing all the CRF fields that can be populated from the text, rather than generating each field independently. In particular, the prompt instructs the model to extract only fields that are supported by

System	Model	TP	FP	FN	F1
Extraction	Gemini 3 Flash	289	385	56	0.599
Extraction	GPT-OSS 120B	231	190	121	0.575
Extraction + Verification	Gemini 3 Flash + GPT-OSS 120B	276	372	78	0.604

Table 2: Performance comparison on the Italian development set.

clear and explicit evidence in the clinical note and to avoid any form of inference or interpretation beyond the literal text. The full prompt used in the extraction stage is reported in Appendix A.1.

The 3-shot examples included in the prompt consist of pairs of clinical notes and their corresponding CRF representations, demonstrating the expected input–output format.

For this stage we use the Gemini 3 Flash model (Pichai, 2025), which showed the strongest extraction performance among the models evaluated during preliminary experiments. These experiments consisted of running different LLMs under the same prompting setup on the development set and comparing their macro F1 scores. In addition to Gemini 3 Flash, we tested open-weights models such as LLaMA4 Maverick 17B, Gemma3 27B, Qwen3, and GPT-OSS 120B, which achieved lower extraction performance when used alone. For conciseness, these additional results are not reported in detail in this paper, as they do not affect the main findings.

4.2 CRF Verification

Although the extraction stage is explicitly instructed to output only fields supported by clear textual evidence, we observed that the model often produces unsupported predictions, resulting in a relatively high number of false positives (particularly with the Gemini 3 Flash model).

To address this issue, we introduce a second verification stage. In this step, a separate LLM receives two inputs: the original clinical note and the CRF generated by the extraction stage. The task of this model is to verify whether each predicted field is directly supported by the text.

The verification prompt instructs the model to examine each populated CRF field and retain it only if the clinical note contains clear and unambiguous evidence supporting the assigned value. If the evidence is absent, ambiguous, or implicit, the field must be removed from the output. Importantly, the model is explicitly prohibited from adding new fields or modifying the values of existing predic-

tions. Its role is strictly limited to filtering unsupported predictions. The complete prompt used in this stage is provided in Appendix A.2.

For this stage we selected, among the models evaluated in the preliminary experiments mentioned in Section 4.1, those that exhibited the most conservative prediction behavior, i.e., producing fewer false positives and fewer overall predictions when used for extraction. This selection was performed separately for each language based on development set performance. In particular, we used the LLaMA4 Maverick 17B model for the English language and GPT-OSS 120B for Italian (OpenAI et al., 2025), as these models showed the strongest tendency to refrain from assigning CRF labels in the absence of clear textual evidence in their respective settings.

The final CRF prediction corresponds to the filtered JSON output produced by this verification stage.

5 Results

We evaluate our approach on both the development and test sets provided by the shared task. Performance is measured using macro F1, the official evaluation metric of the competition. The development set allows a more detailed analysis of system behavior, including the number of true positives (TP), false positives (FP), and false negatives (FN), while for the test set only the final macro F1 score is available.

Tables 1 and 2 report the results on the development set for the English and Italian datasets, respectively. We compare the extraction performance of two models used independently with the final two-stage pipeline.

The Gemini 3 Flash model achieves the best extraction performance (also compared to other models tested in preliminary experiments, which are not shown here for brevity). However, it produces a relatively high number of false positives in both English and Italian.

In contrast, on the English dataset (Table 1), LLaMA4 Maverick 17B produces fewer false pos-

itives but suffers from a higher number of false negatives due to its more conservative prediction behavior. The proposed pipeline combines these complementary characteristics by using Gemini for the extraction step and LLaMA4 for the verification step. This second stage filters unsupported predictions while preserving most of the correct extractions. As a result, the verification stage significantly reduces the number of false positives produced by the extraction model without substantially decreasing the number of true positives, leading to a noticeable improvement in macro F1.

For the Italian dataset, a different model was used in the verification stage. Preliminary experiments showed that GPT-OSS 120B achieved slightly better performance in the extraction+validation pipeline. More generally, we observed a similar behavior between LLaMA4 Maverick 17B and GPT-OSS models: compared to the models used for extraction, they tend to produce fewer predictions and adopt a more conservative strategy. While this behavior may reduce recall when used alone, it proves beneficial in the verification step, where the goal is to filter unsupported predictions produced during extraction (Table 2).

Table 3 reports the final performance of the pipeline on the test set for both languages. As expected, the overall macro F1 score is lower than the one observed on the development set. This difference is common, as the test set represents unseen data and may contain different distributions of CRF mentions and textual patterns.

Despite this decrease, the results confirm the effectiveness of the proposed two-stage approach. The pipeline maintains competitive performance across both languages while benefiting from the verification stage. This suggests that separating extraction and validation allows the system to balance recall-oriented extraction with precision-oriented verification, leading to more reliable CRF predictions overall.

System	Eng F1	Ita F1
Extraction + Verification	0.56	0.56

Table 3: Final macro F1 scores of the proposed system on the test set.

6 Conclusion

In this paper we presented a two-stage LLM-based pipeline for automatically populating clinical Case

Report Forms from free-text clinical notes. The system separates the extraction of candidate CRF fields from their verification, allowing different models to specialize in complementary behaviors. The extraction stage focuses on maximizing coverage of potentially relevant fields, while the verification stage adopts a more conservative strategy to filter predictions that are not supported by explicit textual evidence.

Experimental results show that this design helps reduce the number of false positives produced by the extraction model while preserving most correct predictions, leading to improved macro F1 on the development set. The final system achieves a macro F1 score of 0.56 on both English and Italian test sets.

Overall, our results suggest that combining generative extraction with explicit verification is a promising direction for structured information extraction from clinical narratives.

References

- Pietro Ferrazzi, Mattia Franzin, Alberto Lavelli, and Bernardo Magnini. 2026a. [Small LLMs for Medical NLP: a Systematic Analysis of Few-Shot, Constraint Decoding, Fine-Tuning and Continual Pre-Training in Italian](#). *Preprint*, arXiv:2602.17475. ArXiv preprint.
- Pietro Ferrazzi, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. 2026b. Overview of the CRF 2026 Shared Task on Clinical Case Report Forms filling. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma, Mallorca (Spain). ELRA.
- Pietro Ferrazzi, Alberto Lavelli, and Bernardo Magnini. 2025. [Converting Annotated Clinical Cases into Structured Case Report Forms](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 307–318, Vienna, Austria. Association for Computational Linguistics.
- Gabriela Anna Kaczmarek, Pietro Ferrazzi, Lorenzo Porta, Vicky Rubini, and Bernardo Magnini. 2026. [Toward automatic filling of case report forms: A case study on data from an Italian emergency department](#). *Preprint*, arXiv:2602.23062. ArXiv preprint.
- OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. [gpt-oss-120b & gpt-oss-20b Model Card](#). *Preprint*, arXiv:2508.10925.

Sundar Pichai. 2025. [A new era of intelligence with gemini 3.](#)

A Appendix: Prompt Details

A.1 Extraction Stage Prompt

The system prompt used for the CRF extraction stage is shown below:

```
You are a medical information extraction system.
Task:

From a single clinical note, select ONLY the
CRF fields that are supported by clear,
explicit, and unambiguous evidence in the
text, and assign them a value.

Critical rules:
- If a field is not EXPLICITLY and CLEARLY
  stated, DO NOT output that field.
- Absence of information, negations, hypotheses
  , suspicions, plans, history without
  confirmation, vague wording or doubt must
  result in the field being completely
  omitted.
- Only output a field when the text contains a
  direct statement that uniquely determines
  its value, and choose the value that
  exactly matches the literal meaning of the
  text.
- Never infer, guess, normalize, or interpret
  beyond the literal text.

Output rules:
- Output a valid, minified JSON object
  containing ONLY the selected fields.
- Each output value must be exactly one of the
  allowed values for that field.
- All values are strings.
- Do not include explanations, comments, or
  extra text.
```

- Do NOT change values.
- Do NOT normalize or reinterpret.
- Do NOT infer from medical knowledge.
- Judge ONLY based on what is explicitly written in the note.

Output:

- Return ONLY a valid, minified JSON object.
- Include ONLY the fields that are fully supported.
- No explanations, comments, or extra text.

A.2 Verification Stage Prompt

The system prompt used for the verification stage is shown below:

```
You are a clinical data verification system.
Task:
Given
1) a clinical note
2) a JSON object containing ONLY populated CRF
  fields
Your job is to VERIFY each field in the JSON
  against the clinical note.

Rules:
- For EACH field in the input JSON, check
  whether the clinical note contains clear,
  explicit, and direct evidence supporting
  the assigned value.
- If the evidence is sufficient and unambiguous
  , KEEP the field exactly as is.
- If evidence is absent, weak, implicit,
  ambiguous, contradictory, or based on
  inference, REMOVE the field entirely from
  the output.
- Do NOT add new fields.
```

VerbaNexAI at ClinicalSkillQA: A Two-Stage Generative Vision-Language Framework for Procedural Frame Ordering

Andrea Menco-Tovar and Jairo E. Serrano and Edwin Puertas
and Juan Carlos Martínez-Santos

Universidad Tecnológica de Bolívar
Cartagena, Colombia

amenco@utb.edu.co, jserrano@utb.edu.co, epuerta@utb.edu.co
jcmartinezs@utb.edu.co

Abstract

This work addresses the temporal ordering task of clinical frames in the Basic Life Support (BLS) subset of ClinSkillQA. A two-stage hybrid pipeline based on Qwen2-VL-2B-Instruct in a zero-shot configuration is proposed. In Stage 1, each image is processed independently to extract factual visual evidence, which is then transformed, using deterministic rules, into a structured representation. In Stage 2, ordering is formulated as an ordinal scoring task over procedural stages, with ties broken using PCA applied to multimodal embeddings. Evaluation followed the official benchmark protocol, considering Task Accuracy, Pairwise Accuracy, and BERTScore. In the test phase, the system achieved Task Accuracy = 0.17, Pairwise Micro Accuracy = 0.60, and BERT F1 = 0.71, with complete coverage in both predictions and rationales. The results demonstrate an interpretable and reproducible foundation, although challenges in fine-grained temporal discrimination remain.

1 Introduction

Clinical skills assessment is a central component of medical training because it evaluates not only declarative knowledge but also the accurate and sequential execution of procedures according to predefined standards. Instruments such as the Objective Structured Clinical Examination (OSCE) support structured assessment of clinical competence, but their large-scale implementation remains limited by time, infrastructure, and expert availability (Issenberg et al., 2005; Vermylen et al., 2025). In this context, evidence from simulation-based education and deliberate practice has shown that repeated procedural training improves skill acquisition and retention, particularly in critical scenarios such as resuscitation and basic life support (BLS), where both action quality and procedural order are essential (Khanghahi and Azar, 2018; Panchal et al., 2020; McGaghie et al., 2011).

The digitalization of training environments and the growing availability of image and video data have created opportunities for automated support in clinical skills assessment. Large multimodal models are especially relevant because they can integrate visual and textual information for clinical image interpretation, guided analysis, and feedback generation (Li et al., 2023; Liu et al., 2023; OpenAI, 2024). However, most biomedical multimodal benchmarks, including VQA-RAD, SLAKE, and PMC-VQA, focus mainly on visual question answering over static medical images (Lau et al., 2018; Liu et al., 2023). Although these resources have advanced medical multimodal AI, they do not explicitly evaluate temporal progression, procedural state transitions, or the reconstruction of ordered clinical actions. Similar limitations persist in video-based assessment of surgical and procedural skills, where temporal granularity, generalization, and reliable interpretation remain challenging (AlSaad et al., 2024; Hartsock and Rasool, 2024; Liu et al., 2021; Seenivasan et al., 2022; Zhang et al., 2024).

The ClinSkillQA 2026 challenge addresses this gap by formulating clinical-skill understanding as a frame-ordering and explanation task. Given shuffled keyframes extracted from videos of student-performed clinical procedures, the system must reconstruct the correct sequence of actions and provide rationales aligned with expert reasoning (Huang et al., 2026). This setting moves beyond isolated visual recognition, requiring models to infer precedence relationships between procedural states while generating explanations that can support formative feedback. Accordingly, the benchmark evaluates both ordering performance, using sequence-level and pairwise accuracy, and rationale quality, using BERTScore and an LLM-as-judge strategy based on G-Eval (Zhang et al., 2024; Li et al., 2023; Huang et al., 2026).

In this work, we present our participation in ClinSkillQA through a two-stage multimodal frame-

work for ordering BLS clinical keyframes and generating verifiable rationales. The proposed approach combines structured visual evidence extraction, ordinal procedural scoring, and latent-space tie-breaking to handle ambiguities between visually similar states. Beyond reporting challenge performance, this study analyzes the potential and limitations of multimodal models for procedural assessment tasks requiring temporal reasoning, explanatory consistency, and alignment with clinical expert criteria.

2 Background

Clinical skills assessment is a central component of medical education, since professional competence depends not only on declarative knowledge but also on the accurate, timely, and sequential execution of procedures. Evidence from simulation-based education shows that high-fidelity simulation, deliberate practice, competency-based training, and structured observational tools such as Direct Observation of Procedural Skills (DOPS) can support the acquisition and assessment of procedural competence. These findings highlight the need to evaluate not only whether a clinical action is performed, but also whether it follows the appropriate procedural order and fidelity (Issenberg et al., 2005; Khanghahi and Azar, 2018; McGaghie et al., 2011; Vermylen et al., 2025).

Recent advances in multimodal large language models (MLLMs) and vision-language models (VLMs) have expanded the possibilities for automating clinically relevant visual-textual reasoning tasks. In healthcare, these models have been increasingly applied to image understanding, clinical assistance, report generation, visual question answering, and multimodal decision support. Nevertheless, most medical VQA benchmarks and medical VLMs remain centered on single-image reasoning or localized diagnostic contexts, particularly in radiology. Even procedure-oriented datasets, such as those developed for surgical VQA, tend to emphasize scene understanding or question answering rather than reconstructing a coherent temporal sequence of actions from shuffled visual observations. Similarly, conversational biomedical assistants demonstrate the feasibility of open-ended multimodal interaction, but they do not directly address procedural ordering from fragmented visual evidence (AlSaad et al., 2024; Hartsock and Rasool, 2024; Lau et al., 2018; Zhang et al., 2024;

Seenivasan et al., 2022; Li et al., 2023).

ClinSkillQA addresses this gap by formulating clinical-skill understanding as a frame-ordering task over shuffled keyframes. In this setting, models must arrange frames into a coherent sequence of clinical actions and generate explanations that justify the predicted order. The benchmark includes 200 sets of shuffled keyframes from three types of clinical-skills videos, each accompanied by reference orderings and expert-annotated rationales. Its evaluation is explicitly dual, combining ordering metrics, namely Task Accuracy and Pairwise Accuracy, with explanation-quality metrics based on BERTScore and an LLM-as-a-Judge scheme using G-Eval (Huang et al., 2026).

Against this background, a two-stage generative vision-language framework provides a suitable response to the task. By separating per-frame visual evidence extraction from cross-frame procedural reasoning, the proposed approach first captures observable and reusable cues from each image and then uses them to infer the most coherent clinical sequence and generate set-level rationales. This decomposition aligns with the evaluation structure of ClinSkillQA and connects recent progress in multimodal artificial intelligence with the need for scalable, structured, and procedure-faithful clinical skills assessment.

3 System Overview

The proposed system addresses the ClinSkillQA frame-ordering task as a two-stage hybrid pipeline for Basic Life Support (BLS) procedural sequences. Each input sample contains an unordered set of four or six clinical frames, and the objective is to infer the most plausible procedural order from visual evidence. As shown in Figure 1, the system combines frame-level semantic evidence extraction with sample-level ordinal ranking and tie resolution.

The pipeline as shown in Figure 1, begins by loading the benchmark input file, `BLS_input.json`, and resolving the relative image paths to their corresponding disk locations. This step creates a frame-level index containing the `sample_id`, `frame_id`, and image path, allowing frames to be processed independently in the first stage and later regrouped by sample for sequence prediction.

3.1 Stage 1

In the first stage, each frame is analyzed independently using Qwen2-VL-2B-Instruct (Wang et al., 2024). The model is prompted to produce a short factual description limited to visible content, avoiding speculation and explicit temporal reasoning. The prompt requests three to five bullet points describing observable cues such as hand placement, body posture, chest exposure, clothing condition, and visible equipment.

The generated text is converted into a structured intermediate representation through deterministic post-processing rules. This representation includes detected entities, detected actions, a coarse procedural state, key visual evidence, uncertainty flags, and a short caption summary. The procedural state is inferred from lexical cues in the description: for instance, references to CPR or active compression are mapped to `compressions`, hands placed on the chest without clear downward motion to `position_hands`, and clothing opening or thorax exposure to `expose_chest`. When the evidence is insufficient, the frame is labeled as `other_or_unclear`. All outputs are cached at the frame level to avoid redundant inference.

Rather than producing the final order directly, Stage 1 provides an interpretable semantic layer that summarizes the visual content of each frame. This evidence is later used as auxiliary guidance for ranking and as the basis for textual rationales.

3.2 Stage 2

In the second stage, the same vision-language model is used under a constrained ordinal scoring formulation. Each frame receives a single score from 0 to 5, corresponding to the following procedural states: 0=approach, 1=prepare, 2=expose_chest, 3=position_hands, 4=compressions, and 5=check_or_adjust. The model receives the image and a prompt requiring a single-digit answer, with the Stage 1 caption optionally appended as a hint. This converts frame ordering into ordinal state estimation, which is more stable than direct pairwise comparison.

The predicted score is automatically parsed from the model output. If no valid digit is obtained, the system performs a stricter retry; if parsing still fails, a conservative fallback score is assigned. Frames are then grouped by sample and sorted in ascending order according to their predicted procedural score.

Because multiple frames may receive the same

score, the pipeline applies an embedding-based tie-breaking step. For tied frames only, hidden-state representations are extracted from the multimodal model, averaged into one vector per image, and projected onto a one-dimensional axis using Principal Component Analysis (PCA). Frames within the tied group are then ordered according to their PCA projection values. The final sequence is obtained by concatenating the ordered groups from lower to higher procedural stages.

For each sample, the system exports the predicted frame order together with an auxiliary rationale derived from the Stage 1 evidence. In this way, the framework integrates frame-level visual grounding, ordinal procedural ranking, and interpretable evidence traces for each prediction.

4 Experimental Setup

Experiments were conducted on the Basic Life Support (BLS) subset of ClinSkillQA, using the temporal ordering task over shuffled clinical frames. The benchmark was loaded from `BLS_input.json`, and a frame-level index was created with `sample_id`, `frame_id`, and the corresponding absolute image path. The final processed set included 200 samples: 150 samples with four frames and 50 samples with six frames, with no missing images after path verification.

The implementation was executed on Kaggle using PyTorch and Transformers with CUDA support on a Tesla T4 GPU. The base model was Qwen/Qwen2-VL-2B-Instruct, loaded in `float16` when GPU acceleration was available. All experiments were performed in a zero-shot setting, without task-specific fine-tuning, and deterministic decoding was used by setting `do_sample=False`.

In Stage 1, each image was processed independently to generate a factual description of visible clinical evidence. These outputs were then converted through lexical rules into a structured representation containing detected entities, actions, procedural state, key evidence, a short summary, and uncertainty markers. In Stage 2, the same model assigned each frame an ordinal score from 0 to 5, representing the expected progression of the BLS procedure. Frames within each sample were ordered according to this score, while ties were resolved using PCA over multimodal embeddings extracted from the model’s last hidden layer.

For development and debugging, a DEV subset

of 10 samples was used, whereas the final run was performed on the complete BLS set. Intermediate and final outputs were stored in JSONL format, including the predicted sequence for each sample, denoted as `pred_order`, and a textual rationale derived from the Stage 1 evidence.

Following the official ClinSkillQA protocol, evaluation considers both ordering and explanation quality. Ordering performance is assessed using Task Accuracy, which measures exact sequence prediction, and Pairwise Accuracy, which evaluates correctly ordered adjacent pairs. Explanation quality is evaluated using BERTScore and an LLM-as-a-judge scheme based on G-Eval. Thus, the experimental setup produces both sequence predictions and textual rationales aligned with the official challenge format.

5 Results

Table 1 summarizes the visible test-phase leaderboard results for the ClinSkillQA benchmark, including the position achieved by VerbaNexAI Lab. Our system ranked fourth among the eight visible teams, obtaining an overall score of 37.96. At the ordering level, the method achieved a Task Accuracy of 0.17 and a Pairwise Micro Accuracy of 0.60. This indicates that the system was able to capture some local temporal relationships between frames, but still struggled to recover the complete procedural sequence exactly.

System	Metric	Rank
Score		
Highest visible system	71.43	1/8
<i>VerbaNexAI Lab</i>	<i>37.96</i>	<i>4/8</i>
Task Accuracy		
Highest visible system	0.63	1/8
<i>VerbaNexAI Lab</i>	<i>0.17</i>	<i>4/8</i>
Pairwise Micro Accuracy		
Highest visible system	0.86	1/8
<i>VerbaNexAI Lab</i>	<i>0.60</i>	<i>4/8</i>
BERT F1		
Highest visible system	0.79	1/8
<i>VerbaNexAI Lab</i>	<i>0.71</i>	<i>4/8</i>

Table 1: Compact summary of VerbaNexAI Lab performance in the visible test-phase leaderboard.

For explanation quality, the system obtained BERT Precision = 0.74, BERT Recall = 0.68, and BERT F1 = 0.71. In addition, both Predicted Coverage and Rationale Coverage reached 1.0, showing

that the pipeline generated complete predictions and rationales for all evaluated samples. Compared with the highest visible BERT F1 value of 0.79, the semantic quality of the generated explanations was relatively competitive.

The main performance gap was therefore concentrated in temporal ordering rather than in rationale generation. While the best visible systems reached substantially higher Task Accuracy and Pairwise Micro Accuracy, our method showed a larger drop in exact sequence reconstruction. This suggests that the proposed zero-shot ordinal scoring strategy can identify partial procedural progression, but remains limited when distinguishing visually similar or contiguous BLS states.

Overall, the results show that the proposed framework can generate interpretable evidence traces and complete rationales, but accurate frame ordering remains the main bottleneck. Future improvements should therefore focus on strengthening temporal discrimination, reducing ambiguity between adjacent procedural states, and improving the tie-resolution mechanism used when frames receive similar ordinal scores.

6 Conclusion

This work presented a two-stage hybrid architecture for ordering clinical frames in the BLS subset of ClinSkillQA. The framework combines visual evidence extraction, ordinal procedural scoring, and embedding-based tie resolution to generate both sequence predictions and interpretable rationales. The results show that the system produced complete predictions and explanations for all evaluated samples. However, its main limitation was the exact reconstruction of procedural sequences, especially when adjacent clinical states were visually similar. This suggests that isolated visual recognition is not sufficient for robust procedural ordering. Overall, the study highlights the potential of evidence-driven vision-language reasoning for interpretable clinical skills assessment, while also showing the need for stronger temporal modeling to better distinguish fine-grained procedural transitions.

References

Rawan AlSaad, Ala Abd-Alrazaq, Sabah Boughorbel, Arfan Ahmed, Marie A. Renault, Rami Damseh, and Javaid Sheikh. 2024. [Multimodal large language models in health care: Applications, challenges, and](#)

- future outlook. *Journal of Medical Internet Research*, 26:e59505.
- Ian Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: a review. *Frontiers in Artificial Intelligence*, 7:1430984.
- Xiyang Huang, Jiawei Lin, Keying Wu, Jiaxin Huang, Kailai Yang, Renxiong Wei, Jiayi Xiang, Ziyang Kuang, Min Peng, Qianqian Xie, and 1 others. 2026. Siming-bench: Evaluating procedural correctness from continuous interactions in clinical skill videos. *arXiv preprint arXiv:2604.09037*.
- S. Barry Issenberg, William C. McGaghie, E. Ronald Petrusa, David L. Gordon, and Rosemary J. Scalse. 2005. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical Teacher*, 27(1):10–28.
- Mahdiah Ebrahimi Khangahi and Farahnaz Esmaeili Fomani Azar. 2018. Direct Observation of Procedural Skills (DOPS) evaluation method: Systematic review of evidence. *Medical Journal of the Islamic Republic of Iran*, 32(1):45.
- Jessica J. Lau, Shibashis Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Hao Liu, Jian Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36.
- Bo Liu, Li-Ming Zhan, Lin Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.
- William C. McGaghie, S. Barry Issenberg, Elaine R. Cohen, Jeffrey H. Barsuk, and Diane B. Wayne. 2011. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? a meta-analytic comparative review of the evidence. *Academic Medicine*, 86(6):706–711.
- OpenAI. 2024. GPT-4 Technical Report.
- Ashish R. Panchal, Jason A. Bartos, José G. Cabañas, Michael W. Donnino, Ian R. Drennan, Karen G. Hirsch, Peter J. Kudenchuk, Michael C. Kurz, Eric J. Lavonas, Peter T. Morley, Brian J. O’Neil, Mary Ann Peberdy, Jon C. Rittenberger, Alexis J. Rodriguez, Katherine N. Sawyer, and Katherine M. Berg. 2020. Part 3: Adult basic and advanced life support: 2020 american heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation*, 142(16_suppl_2):S366–S468.
- Lakshmi Seenivasan, Mobarakol Islam, A. K. Krishna, and Hongliang Ren. 2022. Surgical-VQA: Visual question answering in surgical scenes using transformer. In *Lecture Notes in Computer Science*, volume 13437, pages 33–43.
- Julie H. Vermynen, Elaine R. Cohen, David A. Cook, William C. McGaghie, S. Barry Issenberg, Jennifer Arnold, Heather Ballard, Mohamed Bayoumi, Michael Beestrum, Ruth Bremner, Sarah Crawford, Naomi Einstein, Christina Mannarino, Anjali Misra, T. M. Tomita, Hannah Waldron, Francesca Yanko, and Deborah O. Kessler. 2025. Competency-based simulation training for procedural skills: A systematic review and meta-analysis. *Simulation in Healthcare*.
- Peng Wang, Shuai Bai, Shidong Tan, Shunqiang Wang, Zhihao Fan, Jing Bai, Kai Chen, Xian Liu, Jie Wang, Wei Ge, Yu Fan, Kai Dang, Meng Du, Xin Ren, Rui Men, Dong Liu, Chen Zhou, Jingren Zhou, and Ji Lin. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie. 2024. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277.

1.1 Fine-grained Error Analysis

A qualitative inspection of the ordering behavior suggests that the main difficulty was associated with visually adjacent BLS states rather than with completely unrelated procedural phases. This interpretation is consistent with the difference between Pairwise Micro Accuracy and Task Accuracy: the model captured some local frame relationships, but local ambiguities accumulated and affected exact sequence recovery.

The most challenging transitions in BLS are those in which adjacent procedural states share similar visual cues. Early frames may show the rescuer approaching or preparing the scene without clear evidence of a specific procedural action. Similarly, chest exposure and hand positioning may overlap visually, since the chest can become visible before, during, or immediately after hand placement. The transition between hand positioning and active compressions is also difficult in static keyframes because the presence of hands on the chest does not necessarily indicate whether compression has already started. Likewise, compression and adjustment phases may share similar body posture when the rescuer pauses, corrects hand position, or resumes the maneuver.

These ambiguities help explain why the system could generate complete rationales while still fail-

ing to recover the exact sequence. The extracted visual evidence was often sufficient to describe relevant elements in the scene, but not always sufficient to discriminate fine-grained temporal transitions between adjacent BLS states. Therefore, the main limitation of the current framework lies in converting local visual evidence into a globally correct procedural order. This finding also supports the role of the PCA-based tie-breaking mechanism as a local ambiguity-resolution step for frames assigned to similar or tied procedural stages, rather than as a complete solution to temporal reasoning.

.2 Limitations of the Approach

The proposed approach has several limitations. First, the system was evaluated in a zero-shot setting, without task-specific fine-tuning, which likely constrained its ability to distinguish visually similar frames belonging to nearby procedural states. Second, the Stage 1 representation depends on rule-based extraction from short generated descriptions, so relevant cues may be missed when the model expresses the same evidence using unexpected wording. Third, the Stage 2 formulation reduces temporal reasoning to a discrete ordinal scale, thereby improving stability but potentially oversimplifying subtle transitions between adjacent clinical actions. In addition, the PCA-based tie-breaking strategy is only an approximate solution, since a one-dimensional projection does not explicitly model temporal dependencies among frames. Finally, the current experiments focused only on the BLS subset, so the generalizability of the proposed pipeline to other clinical procedures remains to be validated.

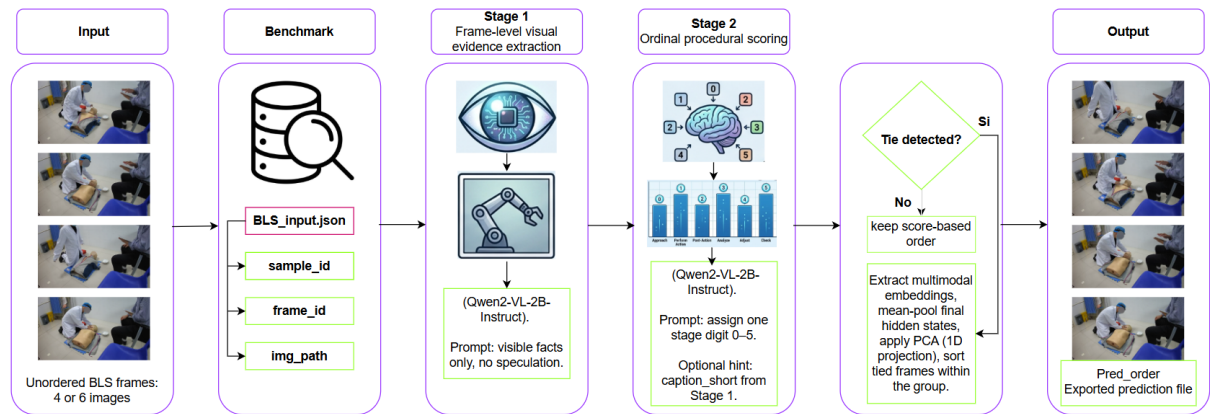


Figure 1: Outline of the proposed model.

zzucs at PsyDefDetect: Bridging Long-Tail Imbalance and Clinical Rubrics for DMRS Defense-Level Detection

Bin Huang¹ and Liuyuan Su¹ and Kaixuan Yuan¹ and
Guanghui Zhao¹ and Shixin Zhang¹ and Kunli Zhang^{1,*}

¹School of Computer and Artificial Intelligence, Zhengzhou University
Zhengzhou, China

{1084893712, 2805476399, 2285986836}@qq.com, suliuyuan@gs.zzu.edu.cn,
17837272557@163.com, ieklzhang@zzu.edu.cn

*Corresponding author

Abstract

Detecting DMRS defense levels in emotional support dialogues is challenging due to severe class imbalance and fine-grained clinical distinctions between adjacent levels, issues well documented in psychotherapy-oriented NLP surveys (Na et al., 2025). We present **zzucs** for PsyDefDetect at BioNLP 2026 (Na et al., 2026a), adopting a **data-supervision co-design** strategy. **SCCR** applies stratified resampling to balance support across nine defense levels. **CoR-QLoRA** encodes clinical rubrics, including task contracts, taxonomy definitions, and boundary cues, into static prompts for 8B model fine-tuning. Ablations show SCCR improves macro-F1 by **4.9 points** over random oversampling. Our system from team **zzucs**, submitted on CodaBench under the display name *sly_zzu* with submission ID 652647, achieves **0.3585 macro-F1** on the **official blind-test leaderboard LB1**. It ranks **6th of 21** registered teams with official submissions and surpasses all published 8B baselines by **4.4 F1 points** over the strongest 8B comparator, Ministral-8B. The code has been released at https://github.com/jackssdd/zzucs_psydefdetect_code.

1 Introduction

Psychological defenses, broadly defined as automatic strategies for managing distress, shape how individuals disclose emotions and respond to therapeutic support (Vaillant, 1992; Perry and Henry, 2004). The Defense Mechanism Rating Scales (DMRS) operationalize these constructs into a validated clinical taxonomy comprising three defensive categories, seven functional levels, and thirty specific mechanisms (Perry and Henry, 2004; Di Giuseppe and Perry, 2021). Despite their centrality in clinical theory, computational detection of defensive functioning in conversational settings remains largely unaddressed in current emotional support dialogue (ESD) systems (Liu et al., 2021;

Rashkin et al., 2019).

The **PSYDEFCONV corpus** (Na et al., 2026b) introduces the first conversational dataset annotated with DMRS defense levels, establishing a nine-way classification task over help-seeker utterances. The PsyDefDetect shared task at BioNLP 2026 formalizes blind leaderboard evaluation on this setting (Na et al., 2026a). Recent surveys also contextualize LLMs in psychotherapy (Na et al., 2025). Although this appears to be a standard multi-class problem, it poses two concrete challenges that limit conventional supervised learning:

1. **Severe distributional imbalance:** Level 7 (high-adaptive) defenses dominate 51.8% of seeker turns, while Level 1 (action) and Level 5 (neurotic) each constitute merely 2.6–5.8% (Na et al., 2026b). Standard training is biased toward majority classes, leading to poor coverage of clinically important but rare defense levels.
2. **Clinical granularity:** Adjacent DMRS levels encode subtle distinctions. For example, *disavowal* at Level 3 and *neurotic* defenses at Level 5 both involve avoidance, but they differ in awareness and displacement patterns that remain hard to capture with vanilla classification objectives, even with strong pre-trained representations.

We argue that simply scaling model size is insufficient for this long-tailed, clinically nuanced setting. While instruction tuning broadly improves LLM task performance (Zhang et al., 2023), severely imbalanced clinical taxonomies still require targeted data and supervision design rather than a larger backbone (Na et al., 2025). We therefore adopt two complementary strategies:

- **Data-centric rebalancing:** stratified class-conditional resampling with replacement, which preserves within-class diversity while equalizing class support.

- **Rubric-based prompting:** encoding DMRS taxonomy definitions and pairwise boundary cues as a static input prefix, providing structured clinical supervision without inference-time generation.

Contributions. Our work makes three contributions: (i) **SCCR**, a stratified resampling method for class-imbalanced clinical dialogue data, which improves macro-F1 by 4.9 points over random oversampling and 9.5 points over no resampling (Table 1). (ii) **Chain-of-Rubric (CoR) prompting**, a structured prompting method that encodes DMRS taxonomy definitions and boundary cues directly in the input prefix. (iii) Results on the official LB1 blind-test leaderboard (Na et al., 2026a) showing our 8B QLoRA system outperforms all published fine-tuned 8B baselines. Due to the system paper page limit, **related work** is deferred to Appendix A.

2 Related Work

Na et al. (2026a) describe PsyDefDetect at BioNLP 2026: nine-way DMRS defense-level detection with blind leaderboard metrics over PSYDEFCONV (Na et al., 2026b). Na et al. (2025) survey LLM applications in psychotherapy and identify challenges around data sparsity and evaluation, which motivates structured clinical supervision for DMRS-level tasks. Most emotional support dialogue (ESD) resources emphasize supporter strategies and empathy ratings (Liu et al., 2021; Rashkin et al., 2019). In contrast, PSYDEFCONV targets *seeker-side* DMRS defense levels as a distinct clinical classification problem. Long-tailed text classification is tackled via reweighting, such as focal loss (Lin et al., 2017), resampling (Chawla et al., 2002), or quality-aware synthesis (Peng et al., 2024; Zhou et al., 2024). Our SCCR uses stratified resampling without generative augmentation to protect DMRS validity. Parameter-efficient adaptation (LoRA/QLoRA (Hu et al., 2022; Dettmers et al., 2023)) and structured or chain-style prompts (Wei et al., 2022; Wu et al., 2024) inform our CoR design. **Appendix A** provides an expanded discussion with additional positioning and citations.

3 Methodology

3.1 Task formulation

Each instance comprises dialogue history d , current seeker utterance u , and label $y \in \mathcal{Y} = \{0, \dots, 8\}$ indexing DMRS levels plus *no defense* (0) and

needs more information (8), as defined for PsyDefDetect (Na et al., 2026a,b). We predict \hat{y} from contextualized encoding of composed input z .

3.2 Module I: Stratified Class-Conditional Resampling (SCCR)

Motivation. PSYDEFCONV exhibits extreme distributional skew (Appendix Fig. 2, left): Level 7 comprises 1,211 instances (51.8%), while Level 1 (action defenses) and Level 5 (neurotic) comprise merely 61–136 instances (2.6–5.8%) each. These rare classes represent clinically important defensive patterns (e.g., impulsive acting-out at Level 1 and repressed conflict at Level 5) that standard training tends to ignore due to majority-class bias.

Formalization. Partition training set \mathcal{D} into per-class subsets \mathcal{D}_c with sizes $n_c = |\mathcal{D}_c|$. The **class cardinality target** matches the largest class:

$$N^* = \max_{c \in \mathcal{Y}} n_c \quad (1)$$

For each c , we build multiset $\tilde{\mathcal{D}}_c$ by **sampling with replacement** from \mathcal{D}_c until $|\tilde{\mathcal{D}}_c| = N^*$, then form $\tilde{\mathcal{D}} = \text{Shuffle}(\bigcup_c \tilde{\mathcal{D}}_c)$. Pseudocode is given as Algorithm 1 in Appendix C.

Unlike *random oversampling*, which duplicates instances drawn uniformly from the union \mathcal{D} until an aggregate balance heuristic is met, SCCR sets an explicit **per-class target** N^* equal to the majority count and resamples *within each stratum* \mathcal{D}_c independently. This equalizes effective supervision across all nine labels while keeping duplicates drawn only from clinically homogeneous buckets.

The result is a balanced training set (Appendix Fig. 2, right) that preserves within-class diversity through replacement sampling. Importantly, we avoid synthetic utterance generation, which could distort DMRS clinical validity.

Validation protocol. Validation uses the same stratified holdout construction as training splits (at least five labeled examples per class), but we *do not* apply SCCR oversampling there: frequencies follow the natural validation distribution so macro-F1 tracks generalization rather than artificial balance. Checkpoint selection maximizes macro-F1 on this naturally distributed validation set.

3.3 Module II: Chain-of-Rubric (CoR) Prompting with QLoRA

Figure 1 summarizes the full pipeline. Appendix Fig. 3 illustrates the three-channel rubric layout in detail.

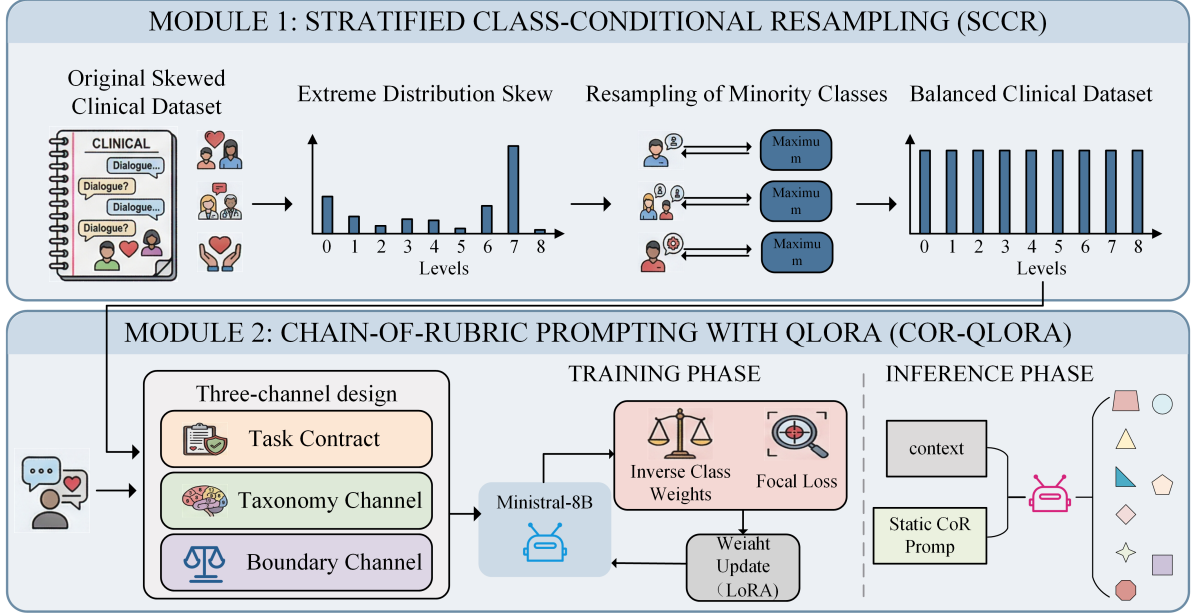


Figure 1: **System architecture.** **Module I** (SCCR) transforms imbalanced training data through stratified over-sampling with replacement. **Module II** (CoR-QLoRA) constructs three-channel rubric prompts (task contract, taxonomy definitions, boundary cues) for injection into a quantized 8B backbone with LoRA adapters. Training employs imbalance-aware focal loss with inverse-frequency class weighting.

Three-channel rubric design. We construct instruction prefix π as **channel concatenation**:

1. **Task Contract (Channel a):** Output space \mathcal{Y} , role restriction (seeker-only), evidence boundary (context up to current turn).
2. **Taxonomy Channel (Channel b):** One-sentence definitions for each DMRS level (0–8) based on the clinical taxonomy, giving the model explicit conceptual anchors.
3. **Boundary Channel (Channel c):** Pairwise disambiguation cues for adjacent level pairs that commonly cause confusion:
 - Level 2 vs 4: extreme versus milder image distortion
 - Level 3 vs 5: defensive avoidance versus indirect neurotic displacement
 - Level 6 vs 7: rigid intellectualization versus flexible adaptive coping

Rationale. While chain-of-thought prompting (Wei et al., 2022) improves reasoning through intermediate generation, this incurs inference-time overhead and can introduce hallucinated reasoning chains. By encoding the rubric directly in the input prefix, we guide the model toward clinically meaningful distinctions without any inference-time generation cost.

Input serialization.

$$z = \underbrace{\pi}_{\text{rubric}} \parallel \underbrace{\text{META}(d, u)}_{\text{metadata}} \parallel \underbrace{\text{CTX}(d)}_{\text{context}} \parallel \underbrace{u}_{\text{target}} \quad (2)$$

QLoRA adaptation. We encode z with Ministral-8B-Instruct (4-bit NF4 quantization (Detmers et al., 2023)), attach a 9-way classification head, and fine-tune LoRA adapters ($r=16$, $\alpha=32$, dropout 0.05) on attention and MLP projection paths (Hu et al., 2022).

3.4 Imbalance-aware optimization

Combining inverse-frequency class weights $\alpha_y \propto 1/n_y$ with focal loss (Lin et al., 2017):

$$\mathcal{L}_{\text{focal}} = -\frac{1}{|\tilde{\mathcal{D}}|} \sum_{(z,y) \in \tilde{\mathcal{D}}} \alpha_y (1 - p_{\theta}(y|z))^{\gamma} \times \log p_{\theta}(y|z) \quad (3)$$

where $\gamma=1.5$ and α_y are renormalized to sum to $|\mathcal{Y}|$. Checkpoints are selected by macro-F1 on stratified validation sets with class-wise evaluation.

4 Experiments

4.1 Experimental setup

Data We train on the official stratified PsyDefDetect training partition of PSYDEFCONV (Na et al.,

Configuration	Acc.	Macro-P	F1
<i>Resampling:</i>			
No resampling	.523	.301	.289
Random oversampling	.561	.342	.335
SCCR (stratified)	.608	.468	.384
<i>Prompt design:</i>			
CoR v1: minimal	.579	.422	.351
CoR v2: + taxonomy	.608	.468	.384
CoR v3: + boundary	.612	.471	.388
<i>Loss functions:</i>			
Standard CE	.589	.446	.371
Class weighting only	.598	.459	.380
Focal ($\gamma=1.5$) + weight	.608	.468	.384

Table 1: Controlled ablations on **stratified validation sets** (not blind test). All experiments use SCCR+focal+CoR v2 as base, varying one factor. Note: Validation F1 is higher than blind test (0.3585 in Table 2) due to distribution similarity with training.

2026b,a): 2,336 seeker utterances with 9-way labels. Validation follows the protocol in §3.2, using a stratified holdout with natural class frequencies and at least five examples per class.

Implementation. Base model: Ministral-8B-Instruct-2410. QLoRA uses 4-bit NF4. Max sequence length is 384. Training: 5 epochs, LR 2×10^{-5} , warmup 5%, batch size 1 with gradient accumulation (effective batch 4–8). See Appendix D for complete hyperparameters.

4.2 Ablation studies

Table 1 summarizes single-factor ablations on stratified validation. Expanded interpretation is in Appendix F. **Resampling:** SCCR outperforms both no resampling and random oversampling (+4.9–9.5 F1), consistent with stratified class-conditional replacement sampling. **Prompt design:** Taxonomy (v2) and boundary cues (v3) improve over v1. Version v3 attains the best F1 despite a higher eval loss than v2. **Loss:** Focal loss ($\gamma=1.5$) with inverse-frequency weighting beats standard CE and weighting-only variants.

4.3 Comparison to published baselines

Table 2 contextualizes our result against published benchmarks tabulated by Na et al. (2026b), following the PsyDefDetect evaluation framing and leaderboard metrics summarized by Na et al. (2026a). This aligns with observations in Na et al. (2025) that clinical NLP tasks benefit from structured supervision beyond generic LLM scaling. Our primary metric is **official LB1 macro-F1 =**

System	Acc.	P	R	F1
<i>Zero-shot LLMs</i> (from (Na et al., 2026b)):				
GPT-5	.528	.276	.166	.195
Gemini 2.5 Pro	.564	.275	.261	.260
DeepSeek-V3.2	.557	.297	.275	.262
<i>Fine-tuned 8B</i> (from (Na et al., 2026b)):				
Llama3.1-8B	.629	.332	.301	.305
GLM-4-9B	.629	.301	.295	.286
Qwen3-8B	.614	.301	.289	.284
InternLM3-8B	.640	.335	.299	.305
Ministral-8B	.648	.340	.305	.315
Ours (zzucs)	.644	.397	.352	.359[†]

Table 2: Baseline numbers reproduced from Table 5 of Na et al. (2026b) (not our re-implementations). [†]**Official LB1** blind-test macro-F1 for team **zzucs** is 0.3585. Rank is **6/21** among registered teams with official submissions. Submission ID is 652647, and the CodaBench display name is *sly_zzu*.

0.3585 on the blind test (Na et al., 2026a). Table 1 shows development-set ablations. Our 8B QLoRA system achieves superior Macro-P (0.3969) and Macro-F1 (0.3585), outperforming the strongest listed 8B full fine-tuning baseline (Ministral-8B at 0.3148 F1) by **4.4 F1 points**.

4.4 Error analysis

On our stratified validation split, errors disproportionately map to Level 7, accounting for about 54% of mistakes. Adjacent-level confusion remains concentrated on 3 vs. 5 and 6 vs. 7 despite boundary cues. Rare Levels 1 and 5 also remain recall-limited after SCCR. Appendix E reports the full diagnostic breakdown and discussion.

5 Conclusion

We described a data and supervision co-design approach for DMRS defense-level detection, combining stratified resampling (SCCR) with rubric-based prompting (CoR) and quantized LoRA fine-tuning. SCCR improves macro-F1 by 4.9 points over random oversampling, and adding boundary cues to the prompt yields further gains despite a higher validation loss, suggesting that structured clinical knowledge acts as a useful regularizer. Our 8B QLoRA system outperforms the strongest published 8B fine-tuning baseline, showing that targeted data and prompt engineering can compensate for limited model scale on clinically fine-grained, long-tailed tasks.

Future work includes dynamic rubric retrieval for ambiguous utterances, quality-preserving syn-

thetic augmentation for rare classes, and cross-lingual transfer to Chinese and Spanish settings where DMRS frameworks may differ. **Ethical note:** Automated defense labels are research tools for analyzing supportive dialogue, not diagnostic substitutes.

Limitations

Language scope: English-only. DMRS taxonomy transfer to other languages requires validation.

Evaluation variance: Single submitted run. Seed and hyperparameter sensitivity remain uncharacterized.

Modality restriction: Text-only. Prosodic cues such as hesitation and affect are unavailable.

Rubric design: Hand-crafted boundary cues. Learned or retrieved rubric augmentation may improve adaptability.

Clinical validity: Synthetic rare-class augmentation must preserve DMRS clinical semantics, an open research challenge.

Acknowledgments

We thank the PsyDefDetect organizers for the benchmark and annotation guidelines. This work was supported by the Natural Science Foundation of Henan Province (Grant No. 252300421877).

References

- Renad M. Alzghoul, Abdulrahman Tabaza, Aya Abdelhaq, and Ahmad Altamimi. 2024. [CLD-MEC at MEDIQA-CORR 2024 task: GPT-4 multi-stage clinical chain of thought prompting for medical errors detection and correction](#). In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 537–556, Mexico City, Mexico. Association for Computational Linguistics.
- Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. [Learning imbalanced datasets with label-distribution-aware margin loss](#). In *Advances in Neural Information Processing Systems*, volume 32.
- Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. [SMOTE: Synthetic minority over-sampling technique](#). *Journal of Artificial Intelligence Research*, 16:321–357.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Mariagrazia Di Giuseppe and J. Christopher Perry. 2021. [The hierarchy of defense mechanisms: Assessing defensive functioning with the defense mechanisms rating scales q-sort](#). *Frontiers in Psychology*, 12:718440.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. [Focal loss for dense object detection](#). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.
- Yuhan Liu, Jun Gao, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. [Empathetic response generation with state management](#). *Preprint*, arXiv:2205.03676.
- Zeyu Liu, Souvik Kundu, Anni Li, Junrui Wan, Lianghao Jiang, and Peter A. Beerel. 2024. [AFLoRA: Adaptive freezing of low rank adaptation in parameter efficient fine-tuning of large models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 161–167, Bangkok, Thailand. Association for Computational Linguistics.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you

- only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.
- Letian Peng, Yi Gu, Chengyu Dong, Zihan Wang, and Jingbo Shang. 2024. [Text grafting: Near-distribution weak supervision for minority classes in text classification](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3741–3752, Miami, Florida, USA. Association for Computational Linguistics.
- J. Christopher Perry and Melissa Henry. 2004. [Studying defense mechanisms in psychotherapy using the defense mechanism rating scales](#). In Uwe Hentschel, Gudmund Smith, Juris G. Draguns, and W. Ehlers, editors, *Defense Mechanisms: Theoretical, Research and Clinical Perspectives*, volume 136 of *Advances in Psychology*, pages 165–192. Elsevier.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Pengjie Ren, Chengshun Shi, Shiguang Wu, Mengqi Zhang, Zhaochun Ren, Maarten de Rijke, Zhumin Chen, and Jiahuan Pei. 2024. [MELoRA: Mini-ensemble low-rank adapters for parameter-efficient fine-tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3052–3064, Bangkok, Thailand. Association for Computational Linguistics.
- George E. Vaillant. 1992. *Ego Mechanisms of Defense: A Guide for Clinicians and Researchers*. American Psychiatric Press, Washington, DC.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35.
- Zhaolong Wu, Abul Hasan, Jinge Wu, Yunsoo Kim, Jason P.Y. Cheung, Teng Zhang, and Honghan Wu. 2024. [KnowLab_AIMed at MEDIQA-CORR 2024: Chain-of-thought \(CoT\) prompting strategies for medical error detection and correction](#). In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 353–359, Mexico City, Mexico. Association for Computational Linguistics.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. [Instruction tuning for large language models: A survey](#). *Preprint*, arXiv:2308.10792.
- Yuhang Zhou, Jing Zhu, Paiheng Xu, Xiaoyu Liu, Xiyao Wang, Danai Koutra, Wei Ai, and Furong Huang. 2024. [Multi-stage balanced distillation: Addressing long-tail challenges in sequence-level knowledge distillation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3315–3333, Miami, Florida, USA. Association for Computational Linguistics.

A Related Work (expanded)

Emotional support dialogue and defensive functioning. ESConv (Liu et al., 2021) established strategy-grounded supportive interaction corpora, with subsequent work extending multi-strategy turn planning and empathetic response evaluation (Rashkin et al., 2019; Liu et al., 2022). However, existing ESD benchmarks focus on supporter strategy selection or empathy ratings, and largely ignore seeker-side defensive functioning, i.e., how distressed individuals manage psychic pain through DMRS defense mechanisms (Na et al., 2026b). Our work addresses this gap by targeting PSYDEFCONV under the PsyDefDetect protocol (Na et al., 2026a). The challenges we observe are consistent with broader psychotherapy LLM issues surveyed by Na et al. (2025).

Long-tailed recognition in text classification. Natural class imbalance in real-world text corpora motivates diverse mitigation strategies. *Reweighting* approaches (focal loss (Lin et al., 2017), label-distribution-aware margins (Cao et al., 2019)) adjust loss contributions. *Resampling* approaches oversample minorities or undersample majorities, with naive random oversampling risking duplicate overfitting (Chawla et al., 2002). Recent work emphasizes *quality-aware* synthesis. Text Grafting (Peng et al., 2024) mines near-distribution templates for minority class augmentation. BalDistill (Zhou et al., 2024) dynamically selects representative head examples while synthesizing tail domain instances for knowledge distillation. SCCR differs by pursuing **pure resampling without synthesis**, leveraging the clinical specificity of PSYDEFCONV utterances where synthetic preservation of DMRS validity is challenging.

Parameter-efficient adaptation and structured prompting. LoRA and its quantized variant QLoRA enable efficient adaptation of large LMs with minimal trainable parameters (Hu et al., 2022; Dettmers et al., 2023). Concurrent work explores adaptive freezing (AFLoRA (Liu et al., 2024)) and mini-ensemble adapters (MELoRA (Ren et al., 2024)) for improved capacity–efficiency tradeoffs.

For classification under limited supervision, *chain-of-thought* prompting (Wei et al., 2022) exposes intermediate reasoning. Similar gains can also come from *schema-based* instructions that encode task structure without generation overhead. Clinical NLP applications demonstrate structured prompting efficacy: MEDIQA-CORR systems employ multi-stage clinical chain-of-thought for medical error detection (Wu et al., 2024; Alzghoul et al., 2024). We extend this line of work by encoding pairwise boundary cues for adjacent DMRS levels directly into the static prompt, providing clinical knowledge supervision without inference-time generation.

B Official benchmark portal

The PsyDefDetect organizers designate **LB1 (macro-F1 on blind test)** as the official shared-task metric (Na et al., 2026a). Numbers should be verified against the official results sheet¹.

Team **zzucs** achieves **LB1 macro-F1 = 0.3585**, ranking **6th of 21** registered teams that submitted to official evaluation.

The public **CodaBench** leaderboard (results tab) aggregates all uploaded entries. The organizers note that it contains 64 entries overall, while the official shared-task ranking is computed over the 21 registered teams with official submissions: <https://www.codabench.org/competitions/12124/#/results-tab>.

C Supplementary figures (SCCR and CoR)

For readability, high-resolution plots referenced from §3.2–3.3 are placed here.

D Implementation details

Hardware. NVIDIA RTX 3090 (24GB) and A100 (40GB) GPUs are used. Training time is approximately 45 minutes per run.

¹https://docs.google.com/spreadsheets/d/1fMDIwC4sisOguLkM3yMd2uMrC4_KjmAiHuWm_QQu1d8/edit?gid=0#gid=0

Algorithm 1 Stratified Class-Conditional Resampling (SCCR)

Require: Training set \mathcal{D} , label space \mathcal{Y}

Ensure: Balanced dataset $\tilde{\mathcal{D}}$

```
1: Partition  $\mathcal{D}$  into per-class subsets  $\mathcal{D}_c$ 
2:  $N^* \leftarrow \max_c |\mathcal{D}_c|$ 
3: for each  $c \in \mathcal{Y}$  do
4:    $\tilde{\mathcal{D}}_c \leftarrow \{\}$ 
5:   while  $|\tilde{\mathcal{D}}_c| < N^*$  do
6:      $x \sim \text{Uniform}(\mathcal{D}_c)$ 
7:      $\tilde{\mathcal{D}}_c \leftarrow \tilde{\mathcal{D}}_c \cup \{x\}$ 
8:   end while
9: end for
10:  $\tilde{\mathcal{D}} \leftarrow \text{Shuffle}(\bigcup_c \tilde{\mathcal{D}}_c)$ 
11: return  $\tilde{\mathcal{D}}$ 
```

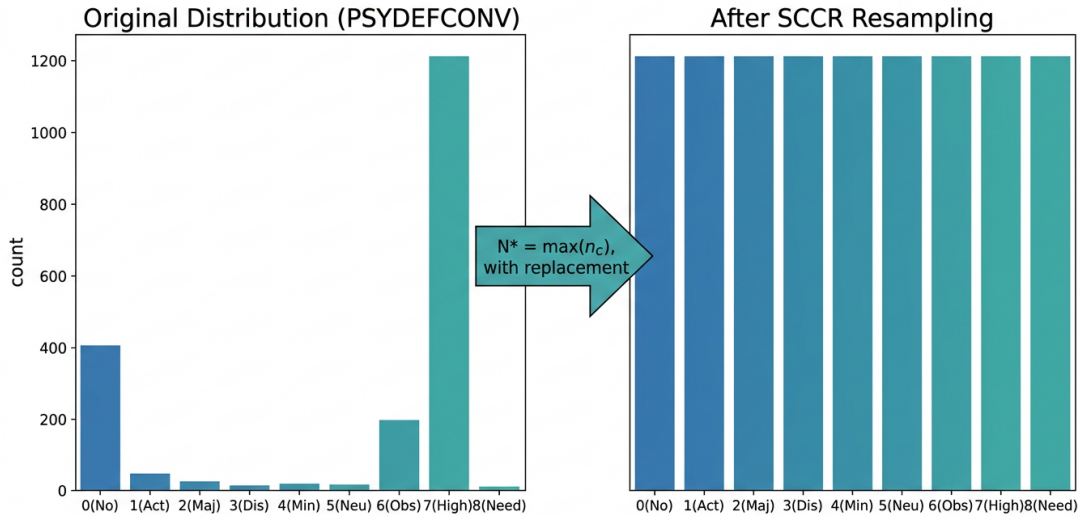


Figure 2: **SCCR**. *Left*: original training distribution (Level 7 majority). *Right*: after resampling, each class has $N^*=1,211$ examples (replacement sampling).

Software stack. PyTorch 2.1.2, Transformers 4.36.2, PEFT 0.7.1, BitsAndBytes 0.41.3, LLaMA-Factory 0.6.0.

Hyperparameters.

- Base model: Ministral-8B-Instruct-2410 (vocab size 32000)
- Quantization: 4-bit NF4, double quantization enabled, compute dtype bfloat16, quant type nf4
- LoRA: $r = 16$, $\alpha = 32$, dropout 0.05, bias none, target modules q_proj,k_proj,v_proj,o_proj,gate_proj,up_proj,down_proj
- Training: 5 epochs, LR 2×10^{-5} constant with warmup, warmup ratio 5%, batch size 1, gradient accumulation steps 8 (effective batch 8)
- Optimizer: AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, weight decay 0.01, max grad norm 1.0
- Focal: $\gamma = 1.5$, class weights $\alpha_y \propto 1/n_y$ renormalized to sum to 9
- Early stopping: patience 5 on macro-F1, evaluation strategy steps, eval steps 80

SCCR implementation. Python/NumPy with fixed seed (42) for reproducibility. Training receives SCCR-balanced multisets as in Algorithm 1. For validation, we hold out a stratified subset with natural

Chain-of-Rubric (CoR) Construction

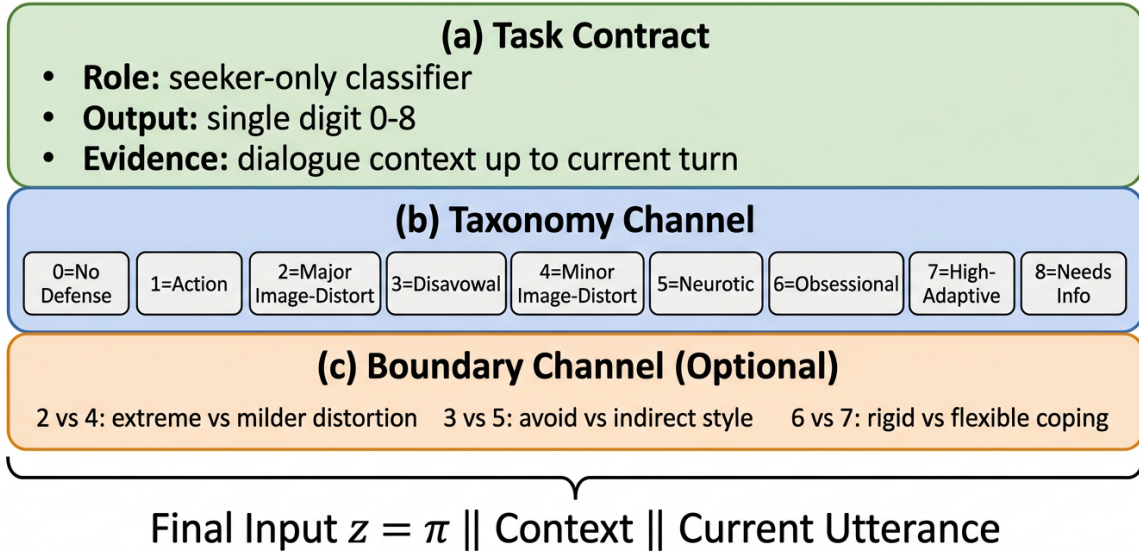


Figure 3: **Chain-of-Rubric (CoR) prompt.** Prefix π : task contract, nine-level taxonomy, pairwise boundary cues.

label frequencies and at least five examples per class where available. If $|n_c| < 5$ in the validation fold, we use all examples of class c without replacement.

E Error analysis (diagnostic)

Level 7 over-prediction. Even with SCCR, models conservatively predict Level 7 for ambiguous utterances (54.2% of errors). This suggests residual *majority bias* from pre-training on general-domain text where adaptive coping is linguistically dominant. Margin-based calibration or decoupled representation learning may further mitigate this.

Boundary confusion. Adjacent levels with subtle distinctions remain challenging despite boundary channel injection. Confusion between Level 3 and Level 5 constitutes 23.7% of errors, while confusion between Level 6 and Level 7 accounts for 18.4%. Manual inspection reveals contextual ambiguity (e.g., sarcasm masking disavowal) defeats explicit rubric cues, suggesting need for *dynamic* rubric conditioning on utterance context.

Rare class recall. Levels 1 (action) and 5 (neurotic) achieve 38.2% and 41.7% recall respectively even with SCCR. This indicates that *data quantity alone is insufficient*. These clinically critical but rare patterns may require specialized synthetic augmentation that preserves DMRS validity, or multi-task pretraining on related clinical constructs such as impulsivity and repression.

F Ablation discussion (expanded)

This section expands the interpretation of Table 1 referenced from §4.2.

Resampling. SCCR yields gains over both no resampling (+9.5 F1) and random oversampling (+4.9 F1). The key difference is per-class targeting to N^* , rather than uniform random duplication, which equalizes class support while keeping each class’s internal diversity intact.

Prompt design. Progressive rubric enrichment improves performance. Version v2 with the taxonomy channel gains +3.3 F1 over v1, and version v3 with boundary cues adds another +0.4 F1. Notably, v3’s best validation loss is higher than v2 (0.6441 vs 0.6219), yet v3 achieves better F1. This suggests that the boundary cues act as a form of regularization, improving generalization even when the model has not fully minimized training loss.

Loss functions. Focal loss with $\gamma=1.5$ improves over standard CE (+1.3 F1) and pure class weighting (+0.4 F1), confirming hard-example mining benefits in this long-tailed setting.

G Complete CoR prompt template (v3)

PSYDEFCONV – DMRS Defense Level Classification for Help-Seeker Utterances

[TASK CONTRACT]

Role: Classify the help-seeker's CURRENT utterance only (not supporter).

Output: Single digit 0-8 representing defense level. No explanations.

Evidence boundary: Use ONLY provided dialogue context up to current turn.

[TAXONOMY CHANNEL – DMRS Level Definitions]

0 No Defense – Phatic/functional utterances with no psychological conflict.

1 Action Defenses – Distress acted impulsively on environment; impulsive, little reflection (e.g., acting out, passive aggression).

2 Major Image-Distorting – Extreme black-white splitting of self/other representations; gross distortion to manage intolerable anxiety.

3 Disavowal – Denial, rationalization, projection, autistic fantasy; refusing to acknowledge unacceptable aspects of reality.

4 Minor Image-Distorting – Milder self-esteem protection than Level 2; devaluation, idealization, omnipotence with less severity.

5 Neurotic – Repression, dissociation, reaction formation, displacement; managing conflict by keeping unacceptable wishes out of awareness.

6 Obsessional – Isolation of affect, intellectualization, undoing; feelings split from facts, rigid, excessive logic, flat affect.

7 High-Adaptive – Affiliation, altruism, anticipation, humor, sublimation; flexible, constructive coping integrating feelings with plans.

8 Needs More Information – Context insufficient for classification.

[BOUNDARY CHANNEL – Pairwise Disambiguation]

If uncertain between:

- Level 2 vs 4: 2 = extreme/gross distortion; 4 = milder self-esteem protection
- Level 3 vs 5: 3 = defensive avoidance/externalizing; 5 = indirect/displaced neurotic style with conflict kept unconscious
- Level 6 vs 7: 6 = rigid affect-intellect split, technical/detached; 7 = flexible integration of feelings with constructive coping

[OUTPUT FORMAT]

Single digit: 0|1|2|3|4|5|6|7|8

H Per-class performance breakdown

Level	Without SCCR			With SCCR		
	Prec.	Recall	F1	Prec.	Recall	F1
0	0.72	0.58	0.64	0.74	0.68	0.71
1	0.12	0.08	0.10	0.38	0.38	0.38
2	0.21	0.15	0.17	0.42	0.41	0.41
3	0.28	0.22	0.24	0.45	0.44	0.44
4	0.25	0.18	0.21	0.41	0.39	0.40
5	0.14	0.09	0.11	0.42	0.42	0.42
6	0.48	0.38	0.42	0.52	0.51	0.51
7	0.68	0.82	0.74	0.64	0.71	0.67
8	0.35	0.28	0.31	0.48	0.46	0.47

Table 3: Per-class precision, recall, and F1 scores on validation set, comparing training without and with SCCR resampling. SCCR dramatically improves rare class (1,2,3,4,5) performance (+27–31 F1 points) while slightly reducing majority class (7) performance (-7 F1), indicating effective rebalancing.

I Negative results and abandoned directions

All macro-F1 figures: **validation**, classes 0–8 (not blind test).

Encoders. BERT-/RoBERTa-/ALBERT-base (CLS head) reached only **0.16–0.18**: short context, imbalance, and adjacent-level DMRS cues favor larger encoders.

Two-LLM few-shot augmentation. G uses label-conditioned few-shot ICL to propose \tilde{x} . S then scores the candidate with threshold τ . Accepted pairs (\tilde{x}, c) are added to \mathcal{D} (Alg. 2).

Algorithm 2 Few-shot LLM generate + score (abandoned)

Require: \mathcal{D} , $\mathcal{Y}_{\text{rare}}$, pools $\mathcal{E}_c^{\text{gen}}$, \mathcal{E}^{scr} , LLMs G , S , τ , K_c

Ensure: \mathcal{D}'

```

1:  $\mathcal{D}' \leftarrow \mathcal{D}$ 
2: for  $c \in \mathcal{Y}_{\text{rare}}$  do
3:   for  $k = 1, \dots, K_c$  do
4:      $\tilde{x} \leftarrow G(\text{GENPROMPT}(\mathcal{E}_c^{\text{gen}}, c))$ 
5:      $s \leftarrow S(\text{SCOREPROMPT}(\mathcal{E}^{\text{scr}}, \tilde{x}))$ 
6:     if  $s \geq \tau$  then
7:        $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{(\tilde{x}, c)\}$ 
8:     end if
9:   end for
10: end for
11: return  $\mathcal{D}'$ 

```

Retraining reached \approx **0.26** macro-F1, which is below stronger non-augmented runs. We therefore use SCCR on real utterances instead (§3).

Ensemble. QLoRA on Qwen2.5-7B + Qwen3-8B with logits averaging reaches \approx **0.28**. This remains well below Ministral-8B + SCCR + Focal, and errors stay correlated on pairs such as 3/5 and 6/7.

Approach	Macro-F1 (val.)
BERT-base (CLS)	0.16
RoBERTa-base (CLS)	0.17
ALBERT-base (CLS)	0.18
Two-LLM few-shot + filter (Alg. 2)	\approx 0.26
Qwen2.5-7B + Qwen3-8B QLoRA ensemble	\approx 0.28

Table 4: Suboptimal explorations (validation macro-F1, classes 0–8).

zzunlp at ClinSkill QA: Perceive-and-Plan with Decomposed In-Context Learning and Saliency-Guided Perception for Clinical Skill Keyframe Reordering

Bin Huang[†] Yi Luo[†] Zhongtian Hua[†] Guanghui Zhao[†]
Kaixuan Yuan[†] Kunli Zhang^{†*}

[†]School of Information Engineering, Zhengzhou University

*Corresponding author

1084893712@qq.com, nancetide@outlook.com, hzt1113@gs.zzu.edu.cn
2285986836@qq.com, 2805476399@qq.com, ieklzhang@zzu.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) show strong medical visual understanding. Their capability for *continuous perception* in procedural clinical workflows remains underexplored. We present **Perceive-and-Plan**, a decomposed in-context learning paradigm for clinical skill keyframe reordering. The method separates visual perception from temporal planning: (1) **structured visual perception** with saliency-guided Picture-in-Picture (PiP) composition that magnifies critical regions (head, chest) as color-coded insets, and (2) **temporal reasoning** with chain-style self-verification via fresh conversation reset and visual-evidence anchoring under BLS Rules R1–R11. With frozen backbone weights, our system achieves **71.43** overall (2nd, ClinSkill QA 2026), 0.86 pairwise accuracy, and 1.0 rationale coverage. Structured prompting with saliency-guided inputs improves procedural understanding in MLLMs. Code is available at <https://github.com/NanceTide/clinskillqa-perceive-and-plan>.

1 Introduction

Multimodal Large Language Models (MLLMs) perform well on visual understanding (Liu et al., 2023; OpenAI et al., 2023), yet applying them to *procedural clinical workflows*, where temporal causality and state transitions matter, remains difficult. Clinical skill assessment requires **continuous perception**: interpreting keyframe sequences, tracking subtle state changes (e.g., clothing exposure, automated external defibrillator pad adhesion), and reconstructing procedural timelines from visual evidence.

The **ClinSkill QA 2026** benchmark formalizes this as a keyframe reordering task: given 4–6 shuffled BLS frames, models must reconstruct the chronological sequence and generate clinically grounded explanations with **visual-evidence anchoring**.

Current *end-to-end prompting* approaches suffer from two limitations: (1) **perceptual ambiguity** (fine-grained details are lost in wide-angle views), and (2) **reasoning interference** (when description and ordering are conflated, models exhibit *temporal hallucination* and favor textbook templates over visual evidence).

We propose **Perceive-and-Plan**, a decomposed prompting paradigm: (1) **perception** (visual analysis with saliency-guided Picture-in-Picture composition), and (2) **planning** (temporal reconstruction in a fresh conversation with chain-style verification). Both stages rely on in-context learning (ICL) with frozen parameters.

Contributions: (1) **PiP composition** that magnifies head and chest regions as color-coded insets, (2) **decomposed ICL** that separates description from ordering via a Visual Anchors protocol, and (3) strong official results on ClinSkill QA 2026: **71.43** overall (2nd), 0.86 pairwise accuracy, and 1.0 rationale coverage.

2 Related Work

Prior work studies temporal multimodal reasoning on generic sequence QA and medical multi-image tasks (Wang et al., 2024; Yu et al., 2025; Huang et al., 2026), with surgical phase recognition (Twinanda et al., 2017) and laparoscopic skill assessment (Liao et al., 2025) as related clinical precedents. Clinical keyframe ordering remains comparatively underexplored.

In-context learning (ICL) supports multimodal adaptation without gradient updates (Dong et al., 2023; Baldassini et al., 2024), yet much of it targets discrete classification or VQA rather than constrained sequence generation. We focus on *clinical procedural ordering* with evidence-grounded rationales, using stage-wise decomposition and saliency-guided perception. Further benchmark-level discussion appears in Appendix A.

3 Methodology

3.1 Overview

Our **Perceive-and-Plan** paradigm decomposes clinical keyframe reordering into two functionally isolated stages. Formally, we define a mapping

$$\mathcal{F} : \mathcal{V} \rightarrow \mathcal{S} \quad (1)$$

where $\mathcal{V} = \{I_1, \dots, I_N\}$ denotes the unordered keyframe sequence and \mathcal{S} denotes the chronologically sorted permutation. As illustrated in Figure 1, the framework factorizes this mapping as:

$$\mathcal{S} = \Psi(\mathcal{V}, \Phi(\mathcal{V}')) \quad (2)$$

where Φ denotes the perception module, Ψ denotes the reasoning module, and \mathcal{V}' represents the set of saliency-enhanced frames.

1. Stage I: Structured Visual Perception (Φ):

Extracts a 4-dimensional clinical state vector \mathbf{d}_i from each keyframe I_i , covering:

- Operator posture
- Hand movement dynamics
- Equipment status
- Patient surface/exposure state

Mitigation of Perceptual Ambiguity: We apply a saliency enhancement function \mathcal{E}_{PiP} defined as:

$$\begin{aligned} I'_i &= \mathcal{E}_{\text{PiP}}(I_i) \\ &= \text{Composite}(I_i, \mathcal{C}_{\text{head}}(I_i), \mathcal{C}_{\text{chest}}(I_i)) \end{aligned} \quad (3)$$

where $\mathcal{C}_{\text{head}}$ and $\mathcal{C}_{\text{chest}}$ denote magnified, color-coded insets of head-neck (red border) and chest-hands (blue border) regions.

2. Stage II: Temporal Planning with Self-Verification (Ψ): Operates under **Conversation Isolation**:

$$\text{Context}(\Psi) \cap \text{Context}(\Phi) = \emptyset \quad (4)$$

The reasoning incorporates original frames \mathcal{V} , pre-computed descriptors $\{\mathbf{d}_i\}$, constraint rules \mathcal{R} , and a Visual Anchors Protocol \mathcal{A} for evidentiary grounding.

Both stages call Qwen3.6-Plus under **ICL**, leveraging long-context multi-image reasoning. Parameters θ are frozen ($\Delta\theta = 0$), and behavior is shaped by prompts and input construction alone.

3.2 Stage I: Structured Visual Perception with PiP Enhancement

Stage I addresses missed fine-grained cues by increasing the effective resolution of salient clinical regions (ROIs). The input construction function \mathcal{E}_{PiP} is:

$$\begin{aligned} I'_i &= \text{Grid}(I_i \oplus \\ &\quad \text{Resize}(\text{Crop}(I_i, \text{bbox}_{\text{head}})) \oplus \\ &\quad \text{Resize}(\text{Crop}(I_i, \text{bbox}_{\text{chest}}))) \end{aligned} \quad (5)$$

where \oplus denotes spatial composition with color-coded borders. The perception module Φ projects this enhanced input into a structured semantic tuple:

$$\mathbf{d}_i = \Phi(I'_i) = \langle \text{Posture, Hands, Equipment, Surface} \rangle_i \quad (6)$$

The prompt conditions Φ on BLS rule library \mathcal{R} as procedural priors, enforcing concise representation for cross-frame comparability. Implementation details appear in Appendix C.3.

3.3 Stage II: Temporal Planning with Self-Verification

3.3.1 Fresh Conversation Reset

Conversation Isolation Protocol: Stage II initializes with context window \mathcal{C}_{Ψ} strictly independent of Stage I:

$$\begin{aligned} \mathcal{C}_{\Psi} &= \{\text{System Prompt}, \mathcal{V}, \{\mathbf{d}_i\}, \mathcal{R}\} \\ &\text{s.t. } \nexists \text{ turn } t \in \mathcal{C}_{\Phi} \end{aligned} \quad (7)$$

This isolation ensures that the ordering process Ψ is not biased by linguistic priors or confidence calibration of description stage Φ .

The planning module Ψ consumes original frames \mathcal{V} , pre-computed descriptors $\{\mathbf{d}_i\}$, and clinical constraints \mathcal{R} to execute a chain-of-verification protocol:

1. **Anchoring:** Enumerate visual evidence \mathcal{E}_v mapping each state transition to a spatial location in I_i
2. **Rationalization:** Generate clinical narrative \mathcal{N} explaining temporal adjacency matrix \mathbf{A}
3. **Serialization:** Output machine-parseable ordered sequence \mathcal{S}

Temperature is reduced in this stage ($\tau = 0.2$) to minimize format drift, contrasting with higher variability in Stage I ($\tau = 0.6$).

Two-Stage Pipeline with Saliency-Guided Perception

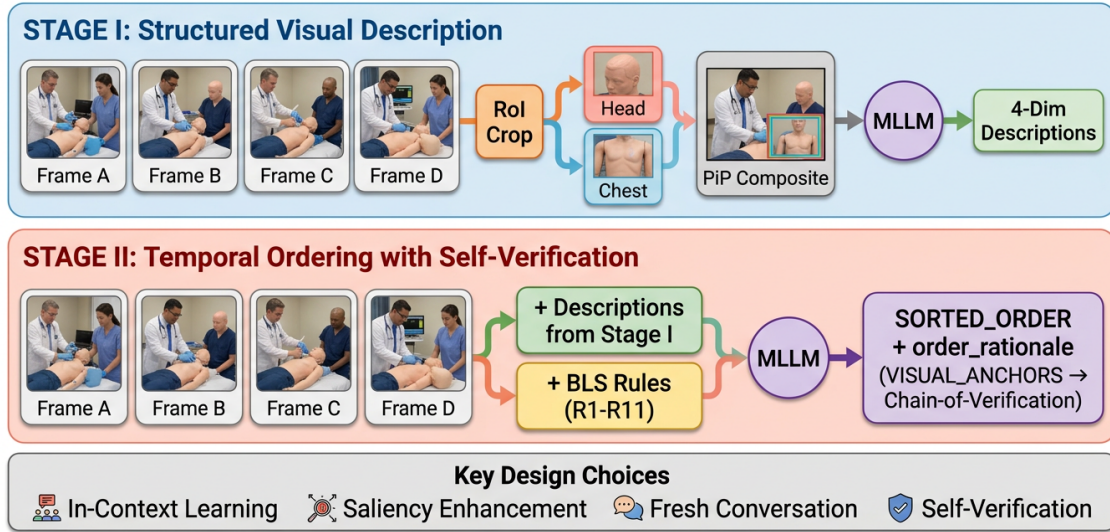


Figure 1: Two-stage pipeline with saliency-guided perception. **Stage I** extracts structured four-dimensional descriptions from ROI-enhanced PiP composites. **Stage II** orders frames with self-verification in a fresh conversation to limit carry-over bias from Stage I. Design elements include ICL with frozen weights, head/chest saliency zoom, a stage-wise conversation reset, and chain-of-verification with visual anchors.

3.4 Implementation Details

We use the Qwen3.6-Plus API with $\tau_1 = 0.6$ (Stage I) and $\tau_2 = 0.2$ (Stage II), without fine-tuning, LoRA, or PEFT. Improvements come from prompts and PiP-style input augmentation (\mathcal{E}_{PiP}). Full prompts, retry logic, and parsing schemas appear in Appendices C–C.5.

4 Experiments

4.1 Dataset and Evaluation

ClinSkill QA 2026 comprises 200 sets of shuffled keyframes from Basic Life Support (BLS) training videos at Zhongnan Hospital, Wuhan University. Each set has 4–6 frames from a continuous procedure (CPR, automated external defibrillator deployment, airway management). Ground truth is expert-annotated chronological order and clinical rationales.

Evaluation metrics:

- **Task Accuracy:** Exact match of predicted order to ground truth
- **Pairwise Accuracy:** Fraction of adjacent frame pairs correctly ordered
- **BERTScore F1:** Similarity between generated and reference rationales

- **Prediction Coverage / Rationale Coverage:** Completeness metrics

4.2 Main Results

Table 1 lists detailed metrics on the official Cod-aBench test set.

Metric	Value	Interpretation
Overall Score	71.43	Primary metric
Task Accuracy	0.63	126/200 exactly correct
Pairwise Micro	0.86	Strong local ordering
BERT Precision	0.79	Semantic precision
BERT Recall	0.79	Semantic recall
BERT F1	0.79	Balanced quality
Prediction Coverage	0.99	198/200 valid predictions
Rationale Coverage	1.0	200/200 with rationale

Table 1: Performance metrics on ClinSkill QA 2026 test set (Team: baovy/zzunlp).

Strengths: Pairwise accuracy 0.86 indicates strong relative ordering even when exact permutations fail. Rationale coverage 1.0 shows that the chain-of-verification protocol reliably yields structured explanations.

4.3 Result Analysis

The gap between task accuracy (0.63) and pairwise accuracy (0.86) reflects correct local relations with residual long-range errors. BERTScore F1 (0.79) indicates solid semantic alignment with references.

Saliency-Guided Input Enhancement: Picture-in-Picture Composition

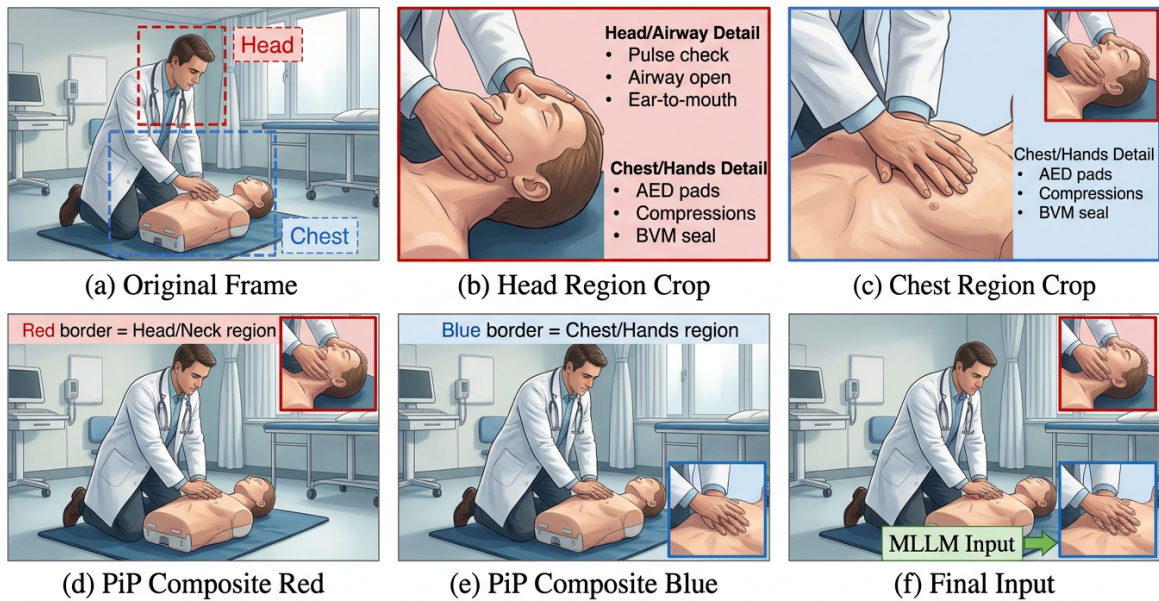


Figure 2: Saliency-guided input enhancement via PiP composition. (a) Original frame with indicated Head (red) and Chest (blue) regions. (b-c) Zoomed crops for head/airway and chest/hands details. (d-e) Individual PiP composites with color-coded borders. (f) Final input to MLLM combining full scene with both magnified insets.

Additional error patterns and qualitative cases appear in Appendix B.

5 Conclusion

We present **Perceive-and-Plan**, a decomposed ICL approach that isolates saliency-guided PiP perception from chain-style temporal planning with self-verification. On ClinSkill QA 2026 it ranks second overall (**71.43**) with frozen model weights. Together, saliency-guided inputs, lightweight domain rules, and explicit verification steps provide a practical approach to procedural reasoning with MLLMs, suited for data-limited and privacy-sensitive educational settings.

Limitations: The system depends on proprietary API access and rate limits. PiP crops are hand-designed for typical BLS manikin layouts and may require retuning for other skills or camera viewpoints. We do not study ensembling or test-time adaptation.

Outlook: We plan to extend the framework to full-video temporal assessment, to explore complementary audio cues (e.g., ventilation sounds, instructor prompts), and to study learned or attention-based region proposals so that saliency cues can transfer beyond fixed BLS-style crops.

Ethics Statement

This study uses the official ClinSkill QA 2026 benchmark, collected under Institutional Review Board (IRB) approval at Zhongnan Hospital, Wuhan University. All footage depicts simulated scenarios with manikins. No real patient data are included. The system is intended for educational assessment only and must not be used for clinical diagnosis or patient care without rigorous validation and regulatory clearance.

Acknowledgments

This work was supported by the Natural Science Foundation of Henan Province (Grant No. 252300421877). We thank the ClinSkill QA organizers for the benchmark and evaluation platform, and the ModelScope team for API access to the Qwen model family.

References

- Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. 2024. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1539–1550.
- Mohamad Ballout, Okajevo Wilfred, Seyedalireza Yaghoubi, Nohayr Muhammad Abdelmoneim, Julius

- Mayer, and Elia Bruni. 2025. [Can you SPLICE it together? A human curated benchmark for probing visual reasoning in VLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11288–11309, Suzhou, China. Association for Computational Linguistics.
- Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. [Benchmarking large language models on answering and explaining challenging medical questions](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, Albuquerque, New Mexico. Association for Computational Linguistics. ArXiv:2402.18060.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2023. [A survey on in-context learning](#). *arXiv preprint arXiv:2301.00234*.
- Sivan Doveh, Shaked Perek, M. Jehanzeb Mirza, Wei Lin, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. 2025. [Towards multimodal in-context learning for vision & language models](#). In *Computer Vision – ECCV 2024 Workshops*, volume 15641 of *Lecture Notes in Computer Science*, pages 250–267. Springer.
- Xiyang Huang, Jiawei Lin, Keying Wu, Jiaxin Huang, Kailai Yang, Renxiong Wei, Cheng Zeng, Jiayi Xiang, Ziyang Kuang, Min Peng, Qianqian Xie, and Sophia Ananiadou. 2026. [Siming-bench: Evaluating procedural correctness from continuous interactions in clinical skill videos](#). *arXiv preprint arXiv:2604.09037*.
- Chuanhao Li, Chenchen Jing, Zhen Li, Mingliang Zhai, Yuwei Wu, and Yunde Jia. 2024. [In-context compositional generalization for large vision-language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17954–17966, Miami, Florida, USA. Association for Computational Linguistics.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. [LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day](#). *arXiv preprint arXiv:2306.00890*.
- Wenqiang Liao, Ying Zhu, Hanwei Zhang, Dan Wang, Lijun Zhang, Tianxiang Chen, Ru Zhou, and Zi Ye. 2025. [Artificial intelligence-assisted phase recognition and skill assessment in laparoscopic surgery: A systematic review](#). *Frontiers in Surgery*, 12.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. [GPT-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Kavya Dasaramoole Prakash, Kiseong Kim, and Youngmahn Han. 2025. [Enhancing clinical reasoning in medical vision-language model through structured prompts](#). *medRxiv*.
- Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. 2017. [EndoNet: A deep architecture for recognition tasks on laparoscopic videos](#). *IEEE Transactions on Medical Imaging*, 36(1):86–97.
- Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. 2024. [Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 416–442, Bangkok, Thailand. Association for Computational Linguistics.
- Suhao Yu, Haojin Wang, Juncheng Wu, Luyang Luo, Jingshen Wang, Cihang Xie, Pranav Rajpurkar, Carl Yang, Yang Yang, Kang Wang, Yannan Yu, and Yuyin Zhou. 2025. [MedframeQA: A multi-image medical VQA benchmark for clinical reasoning](#). *arXiv preprint arXiv:2505.16964*.
- Kangyu Zhu, Ziyuan Qin, Huahui Yi, Zekun Jiang, Qicheng Lao, Shaoting Zhang, and Kang Li. 2025. [Guiding medical vision-language models with diverse visual prompts: Framework design and comprehensive exploration of prompt variations](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11726–11739, Albuquerque, New Mexico. Association for Computational Linguistics.

A Extended Related Work

A.1 Sequence Reasoning Benchmarks

Recent benchmarks (Mementos (Wang et al., 2024), MedFrameQA (Yu et al., 2025), SiMing-Bench (Huang et al., 2026), SPLICE (Ballout et al., 2025)) show that temporal consistency remains difficult for MLLMs across general, medical, and instructional domains. ClinSkill QA additionally requires clinically grounded ordering and rationale generation in one setting.

A.2 Medical and ICL Context

Prior medical vision-language systems often rely on supervised adaptation (Li et al., 2023; Chen et al., 2025) or structured visual prompting (Zhu et al., 2025; Prakash et al., 2025). We instead use decomposed constrained prompting under ICL (Dong et al., 2023; Baldassini et al., 2024; Doveh et al., 2025; Li et al., 2024) to obtain robust structured outputs with frozen backbone weights.

B Extended Error Analysis

We observe three dominant failure modes: (1) **long-range permutation slips**, where most local relations are correct but a single middle transition is misplaced, (2) **partial visibility confusion**, especially for automated external defibrillator pad adhesion and bag-valve-mask sealing under occlusion, and (3) **state granularity mismatch**, where frames with similar global posture but subtle hand-state differences are swapped.

C In-Context Learning Implementation Details

C.1 Full Prompt Templates

C.1.1 Stage I: Structured Visual Description Prompt

Role: Emergency medical training examiner. **Clinical Rules:** Assessment before intervention. Chest exposure progresses from covered to partial to fully exposed. AED phases: package, peel pads, apply to chest, connect to device. BVM sequence: prepare equipment, seal mask, squeeze bag. CPR mandates pulse check before compressions. **BLS Rules (R1–R11):** R1: hand shape and contact site. R2: object and clothing state. R3: space and posture. R4: exposure direction. R5: guideline variants. R6: dependencies. R7: priorities. R8: all people in scene. R9: metadata. R10: visual evidence over theory. R11: continuity constraints. **Output:** Four-dimensional description per image: posture, hands, equipment, surface. Cap at 20 words per dimension and 120 words total.

C.1.2 Stage II: Temporal Ordering Prompt

Role: BLS expert, **Input:** Shuffled frames {A, B, C, D} + descriptions from Stage I, **State Checklist:** (1) EXPOSURE: less-exposed precedes more-exposed, (2) AED PADS: adhered count never decreases, (3) AIRWAY: maneuver precedes sealed BVM, **Chain-of-Verification:** Visual Anchors first (Figure A: [irreversible marker]), then narrative (overview + per-figure paragraphs in chronological order), **Output:** SORTED_ORDER: ["A", "B", ...]

C.2 API Configuration and Retry Logic

- **Model:** Qwen3.6-Plus, accessed via ModelScope API.
- **Stage I:** temperature 0.6, max output 4096 tokens.
- **Stage II:** temperature 0.2, max output 2048 tokens.
- **Timeout:** 300 seconds per request.
- **Retries:** up to 5 attempts with exponential backoff (15, 30, 45, 60, 75 seconds).
- **API Keys:** 20 keys used for rate-limit rotation.

C.3 Picture-in-Picture (PiP) Implementation

ROI Crop Ratios. Head region: (0.05, 0.05, 0.45, 0.55). Chest region: (0.25, 0.40, 0.75, 0.85). **Inset Constraints.** Width: 0.30 of frame width. Height: 0.38 of frame height. **Colors and Borders.** Head border: red (255, 60, 60). Chest border: blue (40, 120, 255). Border thickness: 3 px. **Pipeline.** (1) Load RGB image. (2) Crop head and chest regions. (3) Resize crops with LANCZOS interpolation. (4) Draw color-coded borders. (5) Paste head inset at top-right and chest inset at bottom-right of the original frame. (6) Save composite as JPEG (quality 95).

C.4 Hyperparameter Tuning Log

Hyperparameter sweep on the development set:

Stage I Temperature. Swept 0.0 (too rigid), 0.3 (misses details), 0.6 (selected), 0.9 (too verbose), 1.0 (unreliable). We chose 0.6 as the best balance of detail and conciseness. **Stage II Temperature.** Swept 0.0 (rigid format), 0.2 (selected), 0.5 (occasional errors), 0.7 (inconsistent). We selected 0.2 for reliable format adherence. **Max Tokens.** Stage I: 4096 tokens for four-dimensional descriptions. Stage II: 2048 tokens for rationale and ordering. **Retry Strategy.** Exponential backoff at 15, 30, 45, 60, and 75 seconds, up to 5 retries. Success rate across the full test set was 99.2%.

C.5 Prompt Engineering Iterations

We refined prompts over more than 15 iterations across three weeks:

- **Iterations 1–3 (Structure).** Initial prompting produced irrelevant details. We introduced the four-dimensional description format to focus on clinically relevant content.
- **Iterations 4–6 (Rules).** Injected BLS rules, but the model ignored them. Added explicit “STRICT” markers to enforce compliance.
- **Iterations 7–9 (Format).** Output contained frequent JSON parsing errors. Switched to a SORTED_ORDER line for reliable extraction.
- **Iterations 10–12 (Evidence).** Model engaged in post-hoc rationalization. Introduced Visual Anchors to ground reasoning in observable frame content.
- **Iterations 13–15 (Brevity).** Descriptions were either too sparse or too verbose. Added explicit 20-word-per-dimension caps.

D Sample Predictions and Analysis

D.1 Successful Case: CPR Sequence

Input: [C, A, D, B] (shuffled)

Ground Truth: [D, B, A, C]

Prediction: [D, B, A, C] (Correct)

Rationale: This CPR sequence shows the progression from initial approach to chest compression. Figure D shows the operator approaching the manikin with clothing fully fastened: consistent with the earliest pre-compression stage. Figure B shows partial clothing exposure via unzipping: must follow Figure D and precede full exposure. Figure A shows fully opened clothing with complete chest exposure: immediately before compressions. Figure C shows hands stacked on sternum performing compressions: confirming the final position.

Analysis: The model tracked the monotonic clothing exposure sequence and identified the irreversible compression initiation step.

D.2 Challenging Case: Ambiguous AED Sequence

Input: [A, B, C, D]

Ground Truth: [B, A, D, C]

Prediction: [B, A, C, D] (Incorrect)

Error: Frame C showed one pad partially adhered with unclear visibility of the second pad. The model treated partial adhesion as full completion and placed C before D, illustrating sensitivity to occlusion and ambiguous pad states.

E System Architecture Pseudocode

The PerceiveAndPlan function accepts a sample identifier and a dictionary of image paths. It executes two stages in sequence, each calling the MLLM API with frozen parameters.

Stage I: Structured Visual Perception. For each image, if the saliency-PiP mode is active, the function builds a composite frame with cropped and color-bordered head and chest insets. Each composite is base64-encoded and sent to the MLLM with the Stage I prompt (temperature 0.6). The response is parsed to extract a four-dimensional description (posture, hands, equipment, surface).

Stage II: Temporal Planning. In a fresh conversation, the original frames are base64-encoded and sent alongside the pre-computed Stage I descriptions and the Stage II prompt (temperature 0.2). The response is parsed to extract the sorted ordering and the clinical rationale.

Return Value. The function returns a dictionary mapping the sample ID to the predicted frame order and the corresponding rationale text.

F Detailed BLS Rules Explanation

The BLS Rules Library (R1–R11) encodes emergency medicine practice guidelines, refined through iterative error analysis:

- **R1: Hand and contact disambiguation.** Identify hand shape and contact site. Neck contact indicates pulse check. Forehead or chin contact indicates airway management.
- **R2: Progressive state tracking.** Clothing follows a monotonic sequence: zipper closed, then half open, then fully open. Chest coverage: full, then partial, then exposed. BVM: hovering, then sealed on face, then squeezing. AED pads: 0, then 1, then 2 adhered pads, and the count never decreases.
- **R3: Spatial and posture cues.** Standing back indicates early-stage observation. Kneeling or bent-over posture indicates active intervention. Mask floating near the face suggests preparation, while mask sealed suggests execution.
- **R4: Clinical direction.** Exposure precedes recovery. End-of-scene indicators include examiner writing notes and supplies being stowed.
- **R5: Guideline variants.** Both C-A-B and A-B-C sequences may appear. Assessment precedes intervention unless the scenario involves drowning.
- **R6: Action dependencies.** Airway maneuvers precede rescue breaths. Peeling backing precedes pad adhesion. Pads must be placed before rhythm analysis and shock delivery.
- **R7: Priority rules.** Patient assessment takes priority over hygiene steps and equipment assembly.
- **R8: Multi-person analysis.** Examiner crouching and writing indicates a late-stage frame. Assistant standing back indicates an early-stage frame.
- **R9: Metadata as soft prior.** Sample IDs provide weak ordering hints but must never override direct visual evidence.
- **R10: Evidence over theory.** Prefer the visible timeline in the frames over textbook procedure defaults when they conflict.
- **R11: Continuity constraints.** Adhered pads never decrease in count. Torn packages remain open. Exposed chest stays exposed until the recovery phase.

G Leaderboard Submission Details

Team **baovy** (zzunlp) submitted on 2026-04-07 at 09:07 (submission ID 667353) to the CodaBench platform (<https://www.codabench.org/competitions/14884/>).

G.1 Detailed Performance Metrics

Submission ID 667353, submitted on 2026-04-07 at 09:07 to the CodaBench platform.

Detailed metrics:

- Overall Score: 71.43 (primary metric).
- Task Accuracy: 0.63 (126 out of 200 samples exactly correct).
- Pairwise Micro: 0.86 (strong local ordering).
- BERTScore Precision, Recall, F1: 0.79 (balanced semantic quality).
- Prediction Coverage: 0.99 (198 of 200 predictions valid).
- Rationale Coverage: 1.0 (all 200 samples include an explanation).

H Computational Resources

- Model: Qwen3.6-Plus (via ModelScope API)
- API Calls: 400 requests (200 samples \times 2 stages)
- API Keys: 20 keys for rate limit rotation
- Average Latency: 15–45 seconds per request
- Total Processing Time: 3 hours for full test set
- Development Experiments: 50 hours (hyperparameter tuning)
- Prompt Iterations: 15+ versions over 3 weeks
- Cost: API credits via ModelScope platform

DLNLP at ClinicalSkillQA: EvidenceFlow for Structured Zero-Shot Clinical Keyframe Ordering

Kexin Li, Zhekun Wang, Yiran Wang, Di Zhao*

Dalian Minzu University

zhaodi@dlnu.edu.cn

Abstract

The ClinSkill QA shared task requires models to recover the temporal order of scrambled clinical keyframes and generate explanations. We propose EvidenceFlow, a structured zero-shot framework based on Qwen2.5-VL that decomposes the task into global overview, local evidence modeling, and ordering decision, with two variants: model-led EvidenceFlow-M and rule-guided EvidenceFlow-R. On the official test set, EvidenceFlow-R achieves better ordering performance, while EvidenceFlow-M produces better explanation quality, revealing a trade-off between ordering stability and rationale generation. EvidenceFlow provides an interpretable zero-shot baseline for clinical keyframe ordering.

1 Introduction

Clinical skill assessment is crucial in medical education. The ClinSkill QA task requires models to reorder scrambled clinical keyframes and provide a rationale for the predicted sequence, offering a fine-grained testbed for multimodal understanding of clinical processes in realistic teaching scenarios.

Recent multimodal LLMs have advanced visual understanding and zero-shot reasoning, yet clinical keyframe ordering remains challenging for three reasons. First, the input consists of only a few scrambled static images, requiring the model to infer temporal progression from subtle local differences. Second, multiple visual cues may evolve in parallel, such as chest exposure and hand positioning, which makes global ordering more difficult. Third, the task requires readable explanations grounded in clear evidence.

To address these challenges, we propose a structured zero-shot framework based on Qwen2.5-VL (Bai et al., 2025) that decomposes clinical keyframe ordering into three stages: global

overview, local evidence modeling, and ordering decision. The framework explores two ordering approaches: model-led reasoning and rule-guided correction with explicit evidence constraints. This design is motivated by a practical tension in the task: flexible reasoning may support better explanation generation, whereas stronger evidence constraints may improve ordering stability.

Our contributions are threefold:

1. We propose **EvidenceFlow**, a structured zero-shot framework for clinical keyframe ordering and rationale generation.
2. We design a three-stage pipeline with two ordering variants: model-led **EvidenceFlow-M** and rule-guided **EvidenceFlow-R**.
3. We show a trade-off between ordering stability and explanation quality, clarifying current MLLM limitations on this task.

2 Related Work

In recent years, multimodal research in medical and clinical scenarios has mainly focused on medical visual question answering, image dialogue, and multi-image understanding. Representative studies such as PMC-VQA (Zhang et al., 2024a), LLaVA-Med (Li et al., 2023), and Med-MIM (Yang et al., 2025) demonstrate the potential of multimodal models in medical image understanding, open-ended QA, and multi-image reasoning.

For structured multimodal reasoning, Multimodal-CoT (Zhang et al., 2024b) shows that making intermediate reasoning explicit can facilitate complex vision-language tasks. For temporal and multi-image understanding, MuirBench (Wang et al., 2025), Mementos (Wang et al., 2024), TempCompass (Liu et al., 2024), and TempVS (Song et al., 2025) indicate that robust cross-image or cross-frame temporal reasoning remains challenging for current MLLMs. In the

*Corresponding author.

zero-shot setting, Socratic Models (Zeng et al., 2023) further suggest that organizing intermediate evidence through language can improve interpretability without task-specific training.

Our work differs from these studies by focusing on structured zero-shot ordering in clinical keyframe sequences, where both temporal recovery and evidence-grounded explanation are required.

3 Method

3.1 Task Definition and Overall Framework

Given a set of clinical keyframes $X = \{x_1, x_2, \dots, x_n\}$ (typically $n \in [4, 6]$), the model must output the correct temporal order $Y = [y_1, y_2, \dots, y_n]$ and a natural language explanation. Unlike classification, the core task is to recover the full procedural order from subtle cross-image differences.

We propose EvidenceFlow, a structured zero-shot framework that decomposes clinical keyframe ordering into three stages—global overview, local evidence modeling, and ordering decision—with two variants: model-led EvidenceFlow-M and rule-guided EvidenceFlow-R, as shown in Figure 1.

3.2 Global Overview and Local Evidence Modeling

In the global overview stage, all keyframes in the same sample are first concatenated into a grid image according to their labels, and then input into the multimodal model for holistic analysis. The model needs to extract main change cues, early or late candidate anchors, and corresponding uncertainty information at the group level, denoted as $G = \{a, s^{early}, s^{late}, u\}$, where a denotes the dominant axis of change; s^{early} and s^{late} denote the confidence that an image serves as an early or late anchor, respectively; and u denotes global uncertainty information. Subsequently, the axis of change with the highest confidence is extracted as the main reference axis for subsequent sorting, and candidate anchors are extracted for subsequent ordering decisions.

In the local evidence stage, each image is independently analyzed along six dimensions (chest exposure, hand positioning, hand stability, airway, ventilation, AED), producing a vector $e_i = [chest_i, hand_i, stab_i, airway_i, vent_i, aed_i, c_i, z_i]$, where the first six encode local states, c_i denotes confidence, and z_i denotes a coarse stage label. This stage complements the global overview by

providing fine-grained cues for ordering.

3.3 EvidenceFlow Ordering Decision

Given the global overview and local evidence, the framework proceeds to the ordering decision stage. This stage requires further integrating the overall change trends at the image-group level with the fine-grained clinical cues at the single-image level to generate a consistent global order.

Under this framework, we instantiate two ordering variants: EvidenceFlow-M emphasizes modeled reasoning, whereas EvidenceFlow-R relies more on explicit evidence constraints and rule-based correction. By comparing these two variants, we analyze how different control strategies affect ordering stability and explanation generation.

3.3.1 EvidenceFlow-M: Model-Led Ordering Implementation

This stage first determines whether stable start/end anchors exist to select an ordering mode. To do so, it computes a pairwise relation score for each image from pairwise comparisons:

$$R_{\text{pair}}(x) = \frac{1}{|X| - 1} \sum_{y \neq x} c_{x,y} \delta_{x,y} \quad (1)$$

The resulting score characterizes the relative earliness or lateness tendency of each image within the set, where $c_{x,y}$ denotes the pairwise comparison confidence between images x and y , and $\delta_{x,y}$ is defined as follows:

$$\delta_{x,y} = \begin{cases} 1, & x \text{ is later than } y \\ -1, & x \text{ is earlier than } y \\ 0, & \text{UNCERTAIN} \end{cases} \quad (2)$$

Based on the candidate anchor information, main change cues, and coarse-grained stage labels from the global overview, we compute an anchor score for each image:

$$S_{\text{early}}(x) = A_{\text{early}}(x) - P_{\text{axis}}(x) + B_{\text{early}}(x) \quad (3)$$

$$S_{\text{late}}(x) = A_{\text{late}}(x) + P_{\text{axis}}(x) + B_{\text{late}}(x) \quad (4)$$

Here, $A_{\text{early}}(x)$ and $A_{\text{late}}(x)$ denote the early and late candidate confidences given in the global overview stage, respectively; $P_{\text{axis}}(x)$ denotes the relative progression value of the main change cues for the image; and $B_{\text{early}}(x)$ and $B_{\text{late}}(x)$ respectively denote the stage reward terms derived from the coarse-grained stage labels.

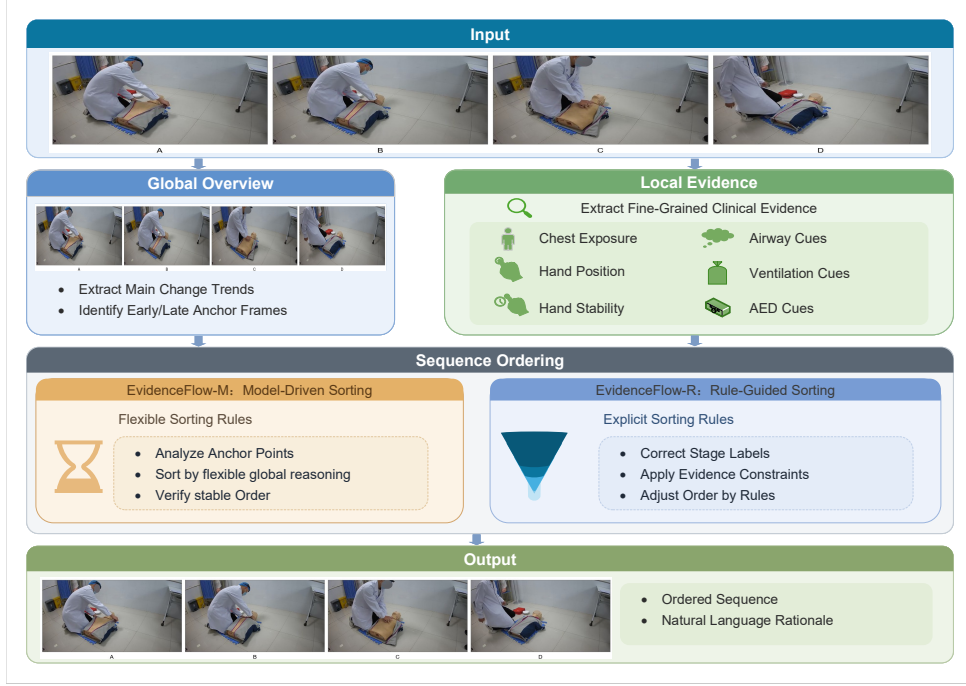


Figure 1: Overview of the EvidenceFlow framework for clinical keyframe ordering.

Based on anchor scores and marginal differences, the model adaptively selects a mode $m \in \{\text{dual, single, none}\}$. In dual mode, early and late anchors fix the start and end, with intermediate images sorted by axis progression. In single mode, one anchor fixes one end, and the remaining images are sorted according to their progression estimates. In none mode, the model estimates relative progression directly. After obtaining the initial sequence $Y^{(0)} = [y_1, y_2, \dots, y_n]$, local corrections using pairwise scores and cached comparisons, followed by lightweight global verification, yield the final sequence Y and its explanation.

3.3.2 EvidenceFlow-R: Rule-Guided Ordering Implementation

Unlike EvidenceFlow-M, EvidenceFlow-R generates ordering results through stage consistency correction, group-level position priors, continuous progression scoring, and directed pairwise comparisons. Specifically, we first refine the coarse-grained stage label z_i obtained from the local evidence stage by combining the local evidence e_i and the global overview information G , yielding a corrected stage value \tilde{z}_i . On this basis, the corrected stage information, the aggregated local evidence scores, and the group-level position priors are jointly mapped into a rule-flow ordering score:

$$S_R(x_i) = \alpha \tilde{z}_i + \beta E_i + \gamma P_i \quad (5)$$

Here, E_i denotes the aggregated local evidence score, and P_i denotes the group-level position prior. The coefficients α , β , and γ denote the corresponding weights. By sorting $S_R(x_i)$ in ascending order, an initial sequence $Y^{(0)} = [y_1, y_2, \dots, y_n]$ can be obtained. For image pairs with close ordering scores, the framework further performs directed pairwise comparisons for local refinement. If necessary, lightweight global verification is then applied to further correct adjacent errors and produce the final sequence Y together with its natural language explanation.

4 Experimental Setup

Experiments are conducted on the ClinSkill QA dataset (200 samples, each containing 4–6 keyframes). The base model Qwen2.5-VL-7B-Instruct is used in a zero-shot setting with 4-bit quantization and a temperature of 0 on a single RTX4090. All experiments are conducted without task-specific training. Both variants share the same backbone and evidence extraction setting, differing only in the final ordering strategy. Following the official evaluation protocol, we report the overall Score, Task Acc (exact sequence match), Pair Micro (pairwise order accuracy), and BERT F1 (Zhang et al., 2020), which measures semantic similarity between the generated explanation and the reference.

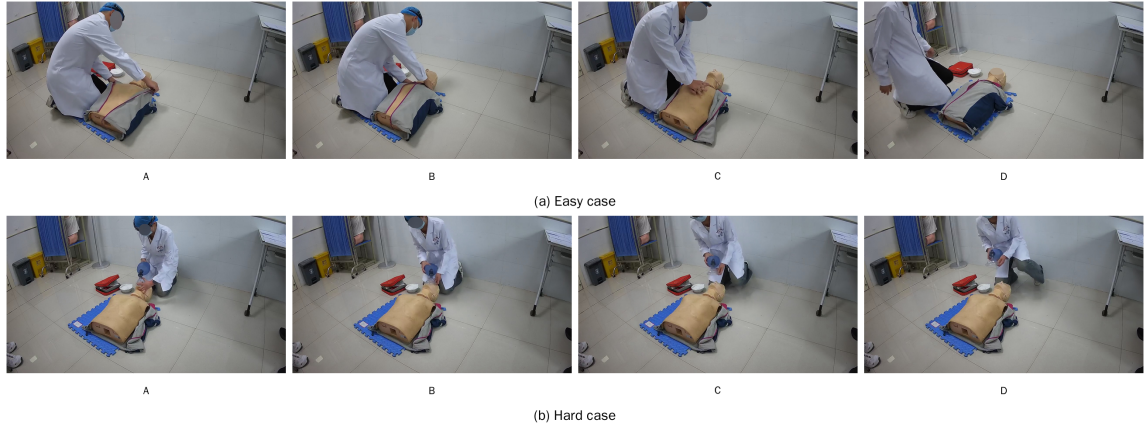


Figure 2: Case study of easy and hard scenarios in CPR emergency response.

5 Experimental Results

5.1 Main Experimental Results

We compare EvidenceFlow-R and EvidenceFlow-M under the same base model and official evaluation protocol. The results are shown in Table 1.

Variant	Score	Task Acc	Pair Micro	BERT F1
EvidenceFlow-R	30.44	0.10	0.55	0.57
EvidenceFlow-M	28.71	0.03	0.51	0.74

Table 1: Main results on the ClinSkill QA test set.

As shown in Table 1, EvidenceFlow-R outperforms EvidenceFlow-M on Score, Task Acc, and Pair Micro, indicating more stable sequence recovery, while EvidenceFlow-M achieves higher BERT F1, reflecting better semantic similarity in explanation generation.

5.2 Comparison with Public Submission Results

Table 2 compares our method with several public submissions from the same test phase. Compared

Method	Score	Task Acc	Pair Micro	BERT F1
baovy	71.43	0.63	0.86	0.79
qppprun	56.73	0.47	0.79	0.55
VerbaNexAI Lab	37.96	0.17	0.60	0.71
EvidenceFlow-R	30.44	0.10	0.55	0.57
EvidenceFlow-M	28.71	0.03	0.51	0.74

Table 2: Comparison of public submissions from the same test phase.

with public submissions, our framework appears more competitive in explanation quality than in ordering accuracy. This suggests that the model

can extract useful local cues, but still struggles to organize them into a consistent global temporal structure. Together with the internal comparison in Table 1, these results indicate that stable cross-image cue integration remains the main bottleneck of the task.

5.3 Error Analysis

Figure 2 shows that (a) is an easy case correctly predicted by Qwen2.5-VL-7B-Instruct; (b) is a hard case with ground truth $C \rightarrow A \rightarrow B \rightarrow D$, but the model predicts $C \rightarrow B \rightarrow A \rightarrow D$, swapping frames A and B. This error arises from high visual similarity during bag-valve mask ventilation, where subtle hand and mask changes are hard to capture, revealing insufficient fine-grained temporal discrimination as the main bottleneck.

6 Conclusion

We propose EvidenceFlow, a structured zero-shot framework for clinical keyframe ordering. The results show that the rule-guided variant yields better ordering performance, whereas the model-led variant produces better explanations, revealing a trade-off between ordering stability and rationale quality. This work provides an interpretable zero-shot baseline for multi-image temporal reasoning in clinical scenarios.

Limitations

The 7B model is limited in fine-grained cross-image temporal integration, especially on hard samples. Our analysis lacks ablations and large-scale verification, and the framework depends on prompt design and handcrafted constraints.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-VL technical report](#). *arXiv preprint arXiv:2502.13923*.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. [LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564. Curran Associates, Inc. Datasets and Benchmarks Track.
- Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. [TempCompass: Do video LLMs really understand videos?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772, Bangkok, Thailand. Association for Computational Linguistics.
- Yingjin Song, Yupei Du, Denis Paperno, and Albert Gatt. 2025. [Burn after reading: Do multimodal large language models truly capture order of events in image sequences?](#) In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24316–24342, Vienna, Austria. Association for Computational Linguistics.
- Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, and 2 others. 2025. [MuirBench: A comprehensive benchmark for robust multi-image understanding](#). In *The Thirteenth International Conference on Learning Representations*.
- Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. 2024. [Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 416–442, Bangkok, Thailand. Association for Computational Linguistics.
- Xikai Yang, Juzheng Miao, Yuchen Yuan, Jiaze Wang, Qi Dou, Jinpeng Li, and Pheng-Ann Heng. 2025. [Medical large vision language models with multi-image visual ability](#). In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15964 of *Lecture Notes in Computer Science*, pages 402–412. Springer Nature Switzerland.
- Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aavek Purohit, Michael S. Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2023. [Socratic models: Composing zero-shot multimodal reasoning with language](#). In *The Eleventh International Conference on Learning Representations*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations*.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024a. [Development of a large-scale medical visual question-answering dataset](#). *Communications Medicine*, 4:277.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alexander Smola. 2024b. [Multimodal chain-of-thought reasoning in language models](#). *Transactions on Machine Learning Research*.

UTS at PsyDefDetect: Multi-Agent Councils and Absence-Based Reasoning for Defense Mechanism Classification

Dima Galat  and Marian-Andrei Rizoiu 
University of Technology Sydney

Abstract

This paper describes our system for classifying psychological defense mechanisms in emotional support dialogues using the Defense Mechanism Rating Scales (DMRS), placing second (F1 0.406) among 64 teams.¹ A central insight is that defense mechanisms are defined by what is *absent*: missing affect, blocked cognition, denied reality. We encode this as an *affect-cognition integration spectrum* in prompt-level clinical rules, which account for the largest single gain (+11.4pp F1).

Our architecture is a multi-phase *deliberative* council of Gemini 2.5 agents where class-specific advocates rate evidence strength rather than voting, achieving F1 0.382 with no fine-tuning—a top-5 result on its own. We find, however, that the council is *confidently wrong* about minority classes: 59–80% of stable minority predictions are incorrect, driven by a systematic “L7 attractor” in which emotional content defaults to the majority class. A targeted override ensemble from three fine-tuned Qwen3.5 models applies 16 overrides (+2.4pp), selected by a structured multi-agent system (builder, critic, regression guard) that produced a larger F1 gain in one iteration than 8 prior attempts combined.

1 Introduction

The BioNLP 2026 shared task (Na et al., 2026a,b) requires classifying target utterances in emotional support dialogues into 9 levels of the Defense Mechanism Rating Scales (DMRS; Perry 1990), a hierarchical clinical instrument ranging from action-based defenses (Level 1) to highly adaptive coping (Level 7). The difficulty is that mechanisms are defined by psychological *function*, not linguistic form: the same surface expression can indicate denial (Level 3), intellectualization (Level 6), or adaptive coping (Level 7) depending on context (illustrated in §2).

¹Code available at <https://github.com/dimagalat/bionlp2026>

The task exhibits severe class imbalance (Table 1): Level 7 (Highly Adaptive) comprises 51.9% of training data, while Level 5 and Level 8 account for only 2.6% and 1.5%. We call this the *L7 attractor effect*: LLMs over-predict the majority class because emotional engagement in therapeutic dialogue looks like “adaptive coping.”

Our system uses a multi-phase deliberative council (§3.1), built on the Gemini 2.5 API (Gemini Team, Google, 2023): three specialist agents classify in parallel, class-specific advocates rate evidence strength, and a resolution stage adjudicates. We frame the DMRS hierarchy as an affect-cognition integration spectrum: many defense mechanisms are defined by what is absent (missing affect, blocked cognition, denied reality), which requires reasoning about what should be present but is not. A targeted override ensemble from three fine-tuned models applies 16 overrides to the council’s predictions, achieving macro-F1 0.406 (2nd out of 21 registered teams, or 64 CodaBench entries). We also flag a retrieval-leakage risk: same-dialogue exemplars in few-shot prompts inflate validation accuracy from 65% to 97.7% (§3.2).

2 Task and Data

The dataset contains 1,864 training samples from 200 dialogues and 472 test samples from 189 dialogues. All 189 test dialogue IDs overlap with training dialogues (different utterances from the same conversations), creating a retrieval leakage risk addressed in §3.2. The official metric is macro-averaged F1; we write “F1” throughout to mean macro-F1.

The core challenge is illustrated by this training example: a speaker responds to “How are you today?” with “I’m OK. Just dealing with a lot of unknowns.” This reads like Level 7 (*Suppression*, consciously managing distress). The ground

Level	Name	Train	%
0	No Defense / Neutral	296	15.9
1	Action Defense	108	5.8
2	Major Image-Distorting	61	3.3
3	Disavowal	99	5.3
4	Minor Image-Distorting	84	4.5
5	Neurotic Defense	48	2.6
6	Obsessional Defense	172	9.2
7	Highly Adaptive	968	51.9
8	Needs More Information	28	1.5

Table 1: DMRS class distribution in training data. Level 7 accounts for over half of all samples; the five lower-level defense classes (L1–L5) together comprise only 21.5%.

truth is Level 6 (*Isolation of Affect*): the speaker acknowledges difficulty cognitively (“a lot of unknowns”) but the expected emotional response is absent (“I’m OK”). The distinction turns on what is absent from the utterance, not what is present.

3 System Architecture

3.1 Multi-Phase Deliberative Council

Most LLM ensembles aggregate votes. A *deliberative* council instead evaluates *evidence strength per candidate* through structured advocacy. In our architecture, Phase 1 agents propose candidates with alternatives (not just top-1 predictions). Phase 2 spawns class-specific advocates that rate fit as STRONG, MODERATE, or WEAK; each argues *for* its assigned class using retrieved exemplars. Phase 3 resolves via evidence quality (unique STRONG wins immediately; ties require pairwise comparison), not vote count. Majority voting loses minority classes because L7 always outnumbers them; evidence-based resolution can select a minority class that receives STRONG even when the majority favors L7.

Figure 1 illustrates the pipeline, which uses the Gemini 2.5 API (primarily Flash, with Pro for resolution) and requires 3–10 LLM calls per sample depending on consensus.

Formally, given a dialogue d with target utterance u and label space $\mathcal{Y} = \{0, \dots, 8\}$, the council proceeds in three stages. Three agents a_1, a_2, a_3 each produce a candidate set $C_i = \{(y_i, y'_i, p_i)\}$ (primary label, alternative, confidence). Let $\mathcal{C} = \bigcup_i \{y_i, y'_i\}$ be the candidate pool. If all primaries agree ($y_1 = y_2 = y_3 = y^*$) with $\sum_i \mathbf{1}[p_i \geq \tau] \geq 2$, the council returns y^* immediately. Otherwise, for each unique candidate $c \in \mathcal{C}$, a class-specific advocate A_c produces a strength rating

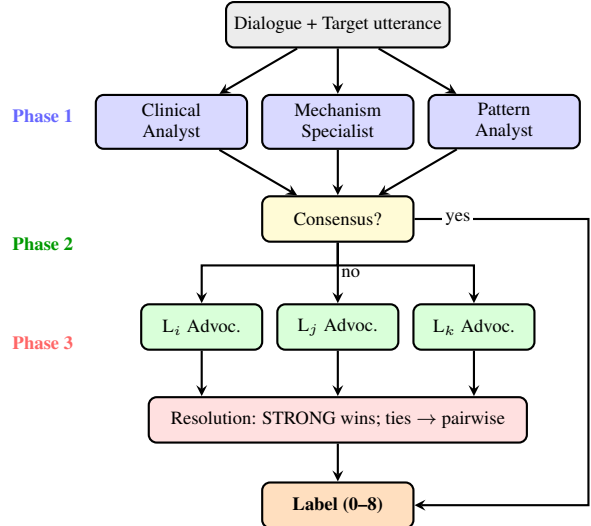


Figure 1: Council pipeline (3–10 LLM calls per sample). Phase 1: three specialist agents classify in parallel. If unanimous with high confidence, the pipeline exits early (3 calls). Otherwise, Phase 2 spawns class-specific advocates rating fit as STRONG/MODERATE/WEAK (2–5 calls). Phase 3 resolves via priority hierarchy (0–1 call).

$s(c) \in \{\text{STRONG, MODERATE, WEAK}\}$. A resolution function R selects the final label via priority ordering: unique STRONG wins; ties are resolved by pairwise head-to-head comparison. Phase details are in Appendix A.

Clinical Knowledge Encoding. We encode the DMRS hierarchy (Perry, 1990) as an *affect-cognition integration spectrum*. The discriminative question at each level is: what is the relationship between what the speaker knows and what they feel? Cognition present but affect drained \rightarrow L6 (Obsessional); affect present but cognition blocked \rightarrow L5 (Neurotic); cognition distorts reality to manage affect \rightarrow L2–4; affect and cognition integrated \rightarrow L7 (Adaptive). The most impactful single test is **Reporting vs. Processing** (L6/L7): describing painful facts without proportional emotion is Isolation of Affect (L6), not adaptive coping (L7). We complement this with five prompt-level rules: (1) 60+ DMRS-Q behavioral indicators per mechanism (Perry, 1990); (2) *Emotion \neq Defense*: “I feel sad” is not a defense without distortion or transformation; (3) prefer lower-level (less mature) defenses when ambiguous; (4) a watchlist of 8 high-confusion class pairs with discriminative tests; and (5) an L7 verification gate requiring a named adaptive mechanism before permitting L7. In ablation, the Gemini 2.5 Pro council without these clinical rules achieves F1 0.268 (Table 4); adding them

raises F1 to 0.382 (+11.4pp).

Phase-Level Bottleneck. Even with clinical rules, the council’s residual errors concentrate at *resolution* rather than detection: the correct label enters Phase 2 as a candidate in 96% of errors and receives a MODERATE+ rating in 76%, but in 94% of errors a wrong label (typically L7) also receives STRONG. The minority-class signal exists; what is missing is a way to prevent L7 from winning the head-to-head, which motivates the override ensemble (§3.3).

3.2 Retrieval and Fine-Tuned Models

Few-shot examples are retrieved via TF-IDF with MMR (Carbonell and Goldstein, 1998) for Phase 1 diversity and semantic re-ranking for Phase 2 within-class exemplars (Lewis et al., 2020). Dialogue-ID exclusion prevents same-conversation leakage; without it, council validation accuracy inflates from 65% to 97.7%.

We train three models via LoRA (Hu et al., 2022) in 4-bit quantization (Dettmers et al., 2023) using Unsloth (Unsloth AI, 2024) and TRL (von Werra et al., 2024): Qwen3.5-9B (65.1% val acc, strongest on L6/L1), Qwen3.5-35B-A3B MoE (61.7%, strongest on L2/L3), and Qwen3.5-9B f1_boost (62.5%, strongest on L1/L2) (Qwen Team, 2026). All use **completion-only loss** (`train_on_responses_only`): training on the full sequence wastes >99% of gradient updates on dialogue auto-completion, and this single change improves accuracy from 25–55% to 59–65%. Self-consistency inference (Wang et al., 2023) (multiple runs at temperature=0.3) provides per-sample confidence scores. A separate pairwise L6/L7 resolver (Qwen2.5-7B, 97.8% val accuracy) handles the dominant confusion pair. Per-model details are in Appendix B.

3.3 Ensemble Strategy

The ensemble applies minimal, high-confidence corrections to the council’s predictions (Figure 2).

Type A (L7→minority, high risk): we override when a fine-tuned model achieves $\geq 80\%$ self-consistency for a minority class and a council rerun or pairwise resolver corroborates. If wrong, we lose a true L7 and add a false minority, a double penalty. **Type B** (minority→minority, lower risk): requires ≥ 3 of 6 independent sources to agree, with FOR > AGAINST. A credibility gate discards models with <15% val recall on the target class. Our

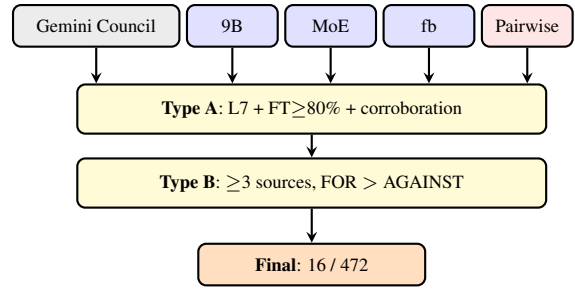


Figure 2: Override ensemble. The council’s predictions are checked against three fine-tuned models (20 self-consistency runs each) and a pairwise L6/L7 resolver. Type A corrects L7 over-predictions (7 overrides); Type B corrects minority confusions (9 overrides).

final submission applies 7 Type A and 9 Type B corrections (3.4% of predictions).

Agentic Override Selection. The override search space is combinatorial: $472 \times 8 = 3,776$ candidates. For the final submission, we decomposed override selection into three formal roles: (1) parallel *builder agents* scanning every sample against all sources with credibility gates; (2) an independent *critic agent* verifying every claim with separate data access; and (3) a programmatic *regression guard* hard-rejecting submissions below the evidence threshold. This structured approach found 5 corrections that ad-hoc exploration had missed, pushing F1 from 0.393 to 0.406, a larger gain than the preceding 8 iterations combined. The lesson: the value is in formal role decomposition, not automation; agents all the way down, but with structure at every level.

Propose–Verify–Decide. Both our council (§3.1) and the override system follow a three-stage *propose–verify–decide* pattern (Figure 3), differing in the verifier’s stance: council advocates argue FOR each candidate class while the override critic argues AGAINST every proposal, and only candidates surviving its scrutiny pass the programmatic regression guard. This mirrors the distinction between the generator–verifier paradigm (Cobbe et al., 2021) and adversarial debate (Irving et al., 2018). Separating proposer and verifier into distinct agents with independent data access prevents the confirmation bias of single-agent self-refine loops (Madaan et al., 2023).

4 Key Findings

L7 Attractor Effect. The council’s L7 advocate rates STRONG 73% of the time versus 32–43% for

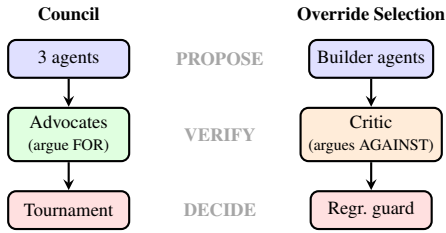


Figure 3: Both multi-agent systems follow a propose–verify–decide pattern. The council’s verifiers are advocates (each argues for one class); the override system’s verifier is an adversary (argues against all candidates).

True → Predicted	Count	% Errors
L6 → L7	66	16.8
L0 → L7	37	9.4
L4 → L7	34	8.7
L3 → L7	32	8.2
L7 → L0	27	6.9
L1 → L3	18	4.6
Any → L7	213	54.3

Table 2: Top 6 stable error confusions across 3 council runs (392 total errors). Over half are incorrect L7 predictions; L6→L7 alone accounts for 17% of all errors.

other classes, causing 54% of stable errors to be incorrect L7 predictions (Table 2). The effect intensifies with dialogue length (15.0 turns for errors vs. 12.6 for correct) and minority class accuracy decays sharply with position: L1 drops from 100% (early turns) to 16% (late turns).

Confidently Wrong. Three identical council runs show 22.2% prediction instability, driven by asynchronous execution order in the multi-agent architecture even at temperature=0. For minority classes L3, L4, L6, and L8, unstable predictions are paradoxically more accurate than stable ones (Table 7 in App. C); between 59% and 80% of stable minority predictions are wrong.

Prompt Overfitting. Every prompt modification that improved training-set F1 degraded test performance (Table 3). We hypothesize this reflects prompt overfitting: iteratively tuning prompts against training-set metrics acts as implicit gradient descent with no generalization check.

Additional findings. TF-IDF with MMR retrieves 38% L7 examples versus 42–46% for all semantic variants, because its lack of semantic understanding prevents emotional-content clustering, a useful property for diversity-dependent classification. Completion-only loss is the most impactful

Modification	Δ Train	Δ Test
L7 advocate rigor	+1.2pp	−3.6pp
Dialogue position metadata	+0.8pp	−2.1pp
Training exemplars in prompts	+1.5pp	−3.6pp
Few-shot $k=3 \rightarrow 5$	+1.5pp	−1.4pp
Length bias warning	+0.5pp	−2.0pp

Table 3: The overfitting paradox: modifications improving training F1 consistently degrade test F1.

System	Acc	P	R	F1
Council (no clin. rules)	.653	.395	.249	.268
Council baseline	.663	.473	.364	.382
+ 7 Type A overrides	.670	.477	.377	.391
+ 9 Type B overrides	.674	.461	.388	.406

Table 4: Test set results. Clinical knowledge rules account for +11.4pp F1 over the unconfigured council; the override ensemble adds +2.4pp through 16 overrides.

fine-tuning intervention: it improves accuracy from 25–55% to 59–65%.

Override Selection Lessons. Beyond architecture, our 9 submissions yielded replicable patterns for ensemble correction on imbalanced classification, four of which are not derivable from individual model accuracy. **Prediction volume signals per-sample reliability.** Under-predictors (9B predicts L2 for 11 samples vs. ~ 16 expected) carry higher precision; over-predictors (f1_boost: 31 vs. ~ 16) inflate false positives, so we weight votes by this volume discount. **Architectural independence is necessary but not sufficient.** A Gemma4-26B-A4B (Gemma Team, Google DeepMind, 2026) agreeing with the Qwen3.5 ensemble only 69–74% of the time (the most independent source we trained) never improved test F1, because per-class accuracy was too low for its disagreements to carry signal. **Override count vs. F1 is sharply non-monotonic.** 75 overrides \rightarrow F1 .367; 16 \rightarrow .406; 21 \rightarrow .399; the peak lives at a narrow intersection of evidence quality and quantity. **Regression guards must be programmatic.** Advisory warnings (“this override has weak evidence”) were ignored by human operators and LLM agents alike; only hard-rejecting submissions failing structural checks prevented regressions.

5 Results

Table 4 shows the ablated contribution of each component. Clinical rules account for the largest gain (+11.4pp). The 7 Type A overrides add +0.9pp F1; the 9 Type B overrides add +1.5pp (recall domi-

Team	Acc	P	R	F1
Nürnberg NLP	.701	.451	.404	.420
UTS (ours)	.674	.461	.388	.406
PerceptionLab	.674	.426	.409	.396
LinguUTics	.642	.400	.396	.392
LDI Lab	.636	.377	.389	.371

Table 5: Top 5 on the shared task leaderboard (21 registered teams). Our system has the highest precision.

nates despite -1.6 pp precision). On the shared task leaderboard (Table 5), our system achieves the highest precision among all teams (0.461), reflecting the conservative override strategy. The development progression (Table 8 in Appendix D) shows that the council’s minority predictions are $\sim 47\%$ correct, so only overrides backed by overwhelming evidence ($\geq 80\%$ FT confidence, ≥ 3 independent sources, zero opposition) reliably improve F1. Post-hoc analysis with the released test labels: 9 of 16 overrides individually corrected council errors (4 Type A, 5 Type B), 4 regressed correct predictions, and 3 were lateral, yielding the $+5$ net correct predictions behind the $+2.4$ pp F1 gain. The 56% override precision sits just above the 47% minority baseline, validating the conservative gating threshold.

6 Related Work

Our deliberative council builds on multi-agent debate (Du et al., 2023) and specialized medical prompting (Nori et al., 2023). The DMRS framework (Perry, 1990) provides the theoretical foundation; the shared task dataset (Na et al., 2026b) enables the first large-scale computational study of defense mechanisms in naturalistic dialogue, and a recent survey (Na et al., 2025) situates this within the broader LLM-psychotherapy landscape. Prior computational approaches to defense mechanisms have been limited to rule-based systems on structured clinical notes; our work is among the first to apply LLMs to this task. Our override framework relates to the generator–verifier paradigm (Cobbe et al., 2021) and adversarial debate (Irving et al., 2018); we adapt self-consistency (Wang et al., 2023) for classification confidence and find that TF-IDF diversity (Carbonell and Goldstein, 1998) outperforms semantic retrieval (Lewis et al., 2020) for few-shot selection under class imbalance.

7 Limitations

Our system depends on the Gemini API, limiting reproducibility to researchers with equivalent access. The validation set is a single GroupKFold split (373 samples); cross-validation was infeasible given API costs. All experiments use English emotional support dialogues from a single cultural context; generalization to other languages or therapeutic traditions is untested. The affect-cognition spectrum is our operationalization of clinical theory (Perry, 1990), not a validated clinical instrument. We address absence-as-signal heuristically; explicit counterfactual reasoning remains open.

8 Conclusion

Our council-ensemble system achieves 2nd place on DMRS defense mechanism classification. The core finding is that defense mechanisms are defined by what is absent, and encoding this insight in clinical rules produces the largest single gain ($+11.4$ pp), more than any model or architectural choice. Formalizing override selection into builder–critic–guard roles then added $+2.4$ pp in a single submission, more than 8 prior iterations combined. Future work could explore hierarchical classification, contrastive training from council error logs, and zero-shot classification via frontier embedding models.

Ethics Statement

This system classifies psychological defense mechanisms (constructs describing internal psychological states) and is a *research tool*, not a clinical diagnostic instrument. It should not be used to label individuals without clinical oversight, as misclassification could distort therapeutic understanding. The training data comes from the shared task organizers (Na et al., 2026a) who ensured appropriate consent and anonymization. Following Strubell et al. (2019), the system’s total compute footprint is ~ 22 kWh (council API at $\sim \$75$; fine-tuning 13 kWh; self-consistency inference 1.2 kWh; PUE 1.2), corresponding to ~ 8.5 kg CO₂eq on the US grid (4.3 kg renewable).

References

- Anthropic. 2026. Claude opus 4.6. <https://www.anthropic.com/news/claude-opus-4-6>.
- Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering doc-

- uments and producing summaries. In *Proceedings of SIGIR*, pages 335–336.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized language models. *arXiv preprint arXiv:2305.14314*.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multi-agent debate. *arXiv preprint arXiv:2305.14325*.
- Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Gemma Team, Google DeepMind. 2026. Gemma 4. <https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/>.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shanan Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899*.
- Chankyu Lee, Rajarshi Roy, Menber Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Han. 2024. NV-Embed: Improved techniques for training LLMs as generalist embedding and retrieval models. *arXiv preprint arXiv:2405.17428*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. In *Proceedings of NeurIPS*.
- Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of EACL*, pages 2014–2037.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the PsyDefDetect shared task at BioNLP 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.
- Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint arXiv:2311.16452*.
- OpenAI. 2024. The GPT model family. *OpenAI Technical Reports*.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- J. Christopher Perry. 1990. Defense Mechanism Rating Scales. Technical report, Cambridge Hospital, Harvard Medical School.
- Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. <https://qwen.ai/blog?id=qwen3.5>.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of ACL*, pages 3645–3650.
- Unsloth AI. 2024. Unsloth: Fast language model fine-tuning. <https://github.com/unslothai/unsloth>.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, and Nathan Lambert. 2024. TRL: Transformer reinforcement learning. <https://github.com/huggingface/trl>.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of ICLR*.

A Council Phase Details

Phase 1: Initial Assessment (3 parallel calls).

Three specialist agents independently classify each utterance: (a) a *Clinical Analyst* applying psychodynamic reasoning (stressor \rightarrow function \rightarrow mechanism \rightarrow level), (b) a *Mechanism Specialist* screening all 9 levels using DMRS-Q behavioral indicators (Perry, 1990), and (c) a *Pattern Analyst* performing analogical reasoning from TF-IDF-retrieved few-shot examples ($k=3$). Each agent outputs a primary label, alternative label, confidence score, and identified mechanism. If all three agents agree on the same label with ≥ 2 having high confidence, the pipeline exits early (3 calls total).

Phase 2: Differential Diagnosis (2–5 calls). For each unique candidate label from Phase 1 (primaries and alternatives), a class-specific advocate evaluates fit as STRONG, MODERATE, or WEAK using class-representative exemplars retrieved via semantic similarity. Advocate criteria are calibrated: STRONG requires “clear, specific evidence; functions similarly to examples”; MODERATE requires “partial or suggestive evidence”; WEAK indicates “little evidence; functions differently.” A minority class screening step injects at least one underrepresented class as a candidate. The L7 advocate rates STRONG 73% of the time (vs. 32–43% for other classes), creating the attractor effect.

Phase 3: Smart Resolution (0–1 call). Resolution follows a priority hierarchy: single STRONG advocate \rightarrow pick immediately (0 calls); multiple STRONG \rightarrow pairwise head-to-head comparison (1 call); multiple MODERATE \rightarrow pairwise comparison; single MODERATE \rightarrow pick; all WEAK \rightarrow deliberation moderator synthesis. The pairwise resolver compares two candidates by studying class-

representative examples for each, then determining which candidate the target utterance is more *functionally* similar to.

B Model Details

Qwen3.5-9B (Qwen Team, 2026) (LoRA $r=64$, attention + MLP targets): 65.1% val accuracy. Strongest on L6 (43% recall) and L1 (36%).

Qwen3.5-35B-A3B MoE (LoRA $r=32$, attention-only²): 61.7% val accuracy. Strongest on L2 (60% recall) and L3 (32%).

Qwen3.5-9B f1_boost (variant training recipe): 62.5% val accuracy. Strongest on L1 (46%) and L2 (60%).

All models use: (1) **completion-only loss** via `train_on_responses_only` (loss computed only on the label token, not the dialogue); (2) GroupKFold by `dialogue_id` for zero-leakage validation (Pedregosa et al., 2011); (3) balanced sampling (L7 capped at 300, minorities oversampled to 80).

Self-Consistency Inference (Wang et al., 2023). Each model runs multiple times at `temperature=0.3` per test sample; the majority vote serves as the prediction and the agreement fraction as a confidence score. Since our output is a single classification token, logit probabilities from a single forward pass could serve as an alternative confidence measure; we used sampling for implementation convenience.

Pairwise Differential Resolver. A separate Qwen2.5-7B model fine-tuned on 942 pairwise comparison examples serves as an L6-vs-L7 specialist. Given two candidate levels and a dialogue, it determines which better fits the target utterance. Per-pair val accuracy: L6/L7 97.8%, L3/L7 100%, L1/L7 100%, L0/L7 52.3%.

C Detailed Findings

C.1 The L7 Attractor Effect

Beyond the headline numbers in §4, the attractor effect intensifies with utterance verbosity: incorrectly classified samples average 23.0 words versus 17.1 for correct ones, and L5 accuracy decays from 100% (early turns) to 18% (late turns), mirroring the L1 pattern. The model uses utterance elaboration as a proxy for defense sophistication, inverting

²MoE expert layers are excluded from LoRA targets due to an Unsloth adapter reload bug. The model leverages pretrained expert routing.

Class	9B	MoE	fb
L1 (Action)	36	18	46
L2 (Image-Dist.)	40	60	60
L3 (Disavowal)	21	32	5
L4 (Image-Dist.)	7	29	14
L5 (Neurotic)	14	14	29
L6 (Obsessional)	43	23	23

Table 6: Per-class recall (%) on validation data for each fine-tuned model. Bold indicates the best model for each class. No model dominates; each is the best or only credible source for ≥ 1 class.

Class	Disagree Rate	Stable Acc	Stable Wrong	Unstable Acc
L0	17.9%	84%	16%	36%
L1	40.7%	41%	59%	32%
L2	37.7%	21%	79%	13%
L3	31.3%	32%	68%	48% [†]
L4	29.8%	24%	76%	40% [†]
L5	27.1%	23%	77%	31%
L6	35.5%	25%	75%	36% [†]
L7	15.5%	91%	9%	47%
L8	46.4%	20%	80%	38% [†]

Table 7: Prediction stability across 3 identical council runs. † marks classes where unstable predictions are more accurate than stable ones; the system is confidently wrong on these minority classes.

the clinical truth: long, detailed descriptions of hardship without emotional processing indicate L6, not L7.

C.2 Prediction Instability and the Confident-Wrong Problem

On the full training set, stable predictions achieve 73.0% accuracy versus 39.2% for unstable ones. However, for minority classes L3, L4, L6, and L8, unstable predictions are paradoxically more accurate than stable ones (Table 7). Between 59% and 80% of stable minority predictions are wrong. The system is not uncertain on hard examples; it is confidently incorrect.

C.3 The Overfitting Paradox in Prompt Engineering

Beyond the table in §4, the mechanism appears to be that iteratively tuning prompts against training-set metrics acts as implicit gradient descent with no generalization check, causing the prompt to memorize training distribution artifacts rather than capture the true classification signal.

#	Configuration	Ov.	F1
1	All FT overrides (aggressive)	75	.367
2	L6/L2 only	42	.375
3	Double corroborated	10	.385
4	Triple corroborated	4	.387
5	+ pairwise L6 resolver	7	.391
6	10 overrides variant	10	.391
7	8 overrides variant	8	.391
8	+ 4 minority \rightarrow minority (Type B)	11	.393
9	+ 5 more Type B	16	.406

Table 8: Development progression across 9 test submissions. All entries are official competition submissions; the final row (#9, F1 .406) was our selected leaderboard entry. The key insight: moving from 75 aggressive overrides (F1 .367) to 16 surgical ones (F1 .406). Fewer, higher-confidence corrections consistently outperform larger override sets.

C.4 Retrieval: TF-IDF Beats Semantic for Diversity

We compared five retrieval strategies for Phase 1 few-shot selection. All four semantic variants (Gemini embedding-001, enriched, task-framed, and focused TF-IDF on last 3 turns) produced 42–46% L7 in retrieved examples. TF-IDF with MMR achieved 38%, well below the 52% base rate, because its lack of semantic understanding prevents emotional-content clustering. For Phase 2 within-class retrieval, however, semantic re-ranking finds better functional matches: embeddings capture functional similarity when the class label constrains the search space. This phase-dependent pattern is worth noting for future work: frontier embedding models trained for zero-shot classification (Lee et al., 2024; Wang et al., 2022; Muennighoff et al., 2023) perform a similar constraint implicitly.

D Development Progression and Failed Approaches

Failed Approaches.

GPT-5.4 (OpenAI, 2024). Standalone F1 of 0.265 with heavy L0 bias (30.5% of predictions). Near-zero L1 detection.

Claude Opus 4.6 agent council (Anthropic, 2026). F1 0.261 without retrieval augmentation, with 69.3% L7 over-prediction.

Aggressive class balancing. Oversampling minorities to 150 (vs. 80) and capping L7 at 200 (vs. 300) caused minority over-prediction, reducing val accuracy from 59.0% to 51.5%.

Chain-of-thought fine-tuning. Template reasoning (“The speaker is managing internal state...”) taught the model to parrot templates rather than classify. Removing CoT and training on label-only output was strictly better.

Gemma4 MoE (Gemma Team, Google DeepMind, 2026). Attention-only LoRA on Gemma4-26B-A4B achieved 63.5% val accuracy with the best L4 recall of any model (36%). However, its test-set predictions failed to improve F1: the model’s L4 signal was contradicted by all Qwen models, and its L1 predictions were heavily over-predicted (47 vs ~ 27 expected). Despite genuine architectural independence from Qwen (69–74% agreement), this independence did not translate to useful override evidence, a cautionary result for cross-architecture ensembling. An earlier attempt applying LoRA to MoE expert layers (not just attention) failed due to an Unsloth adapter serialization bug: the model trained correctly but collapsed to all-L7 when reloaded from checkpoint.

Otter at MedExAct2026: Diverse Encoder Ensemble for Medical Decision Span Detection

Lalita Lowphansirikul
VISTEC
lalita.l_s22@vistec.ac.th

Piyalitt Ittichaiwong
Siriraj Hospital, Mahidol University
piyalitt.itt@mahidol.ac.th

Abstract

We build an ensemble of 10 transformer encoders for the MedExACT 2026 shared task on medical decision span detection. The ensemble is diversified along three training directions: encoder initialization (including domain-adaptive pre-training on clinical text), loss function, and data augmentation with LLM-generated synthetic notes and silver-labeled clinical documents. Greedy forward search selects the combination with the highest validation score. The system achieves a final score of 58.09 (Span F1 51.81, Token F1 66.66) on the test set. Analysis shows that each added model differs from existing members along at least one training direction, producing span predictions that allow majority voting to screen out individual errors.

1 Introduction

Clinical discharge summaries document medical decisions such as diagnoses, prescriptions, and treatment plans, yet these decisions are embedded in unstructured text, making systematic analysis difficult. The MedExACT 2026 shared task (Elgaar et al., 2026) formalizes this as a span detection problem over a 9-category subset of the DICTUM taxonomy (Ofstad et al., 2016), adapting the MedDec dataset (Elgaar et al., 2024). The evaluation uses a subgroup-robust fairness metric that averages overall and worst-group performance across patient demographics (sex, race, and language).

This task poses two key challenges: (1) clinical discharge summaries often exceed the 512-token context window of standard encoders, requiring documents to be split into overlapping segments whose predictions must be reconciled; and (2) only a few hundred annotated documents are available for training across 9 fine-grained categories, three of which (Gathering Information, Treatment Goal, and Deferment) together account for a small fraction of annotations.

Rather than optimizing a single model, we train multiple encoders with different backbones, loss functions, and training data, then aggregate their predictions through majority voting, retaining only spans on which multiple models agree.¹

2 System Description

We train a candidate pool of models that vary along three directions and apply greedy forward selection to identify the best-performing subset.

2.1 Task Formulation

We formulate decision detection as token-level BIO (Begin, Inside, Outside) sequence labeling. Each token is classified by a pretrained encoder followed by a dropout layer and a linear projection with softmax.

2.2 Candidate Pool

We vary three directions of training to build a diverse candidate pool: encoder backbone, loss function, and training data augmentation.

Encoder initialization. We train from two base encoders: BiomedBERT-base (109M parameters) (Gu et al., 2022) and GatorTron-base (345M parameters) (Yang et al., 2022). We additionally introduce a clinical variant of each via domain-adaptive pre-training (DAPT; Gururangan et al., 2020) with entity-centric masking (Lin et al., 2021), continuing the masked language modeling (MLM) objective on 7,814 MIMIC-III discharge summaries (Johnson et al., 2016), excluding MedExACT training and validation documents. We first identify *entities* in the MIMIC-III text: token spans labelled as medical decisions by a majority of five preliminary BIO taggers fine-tuned on the MedExACT training data. At each training step, 15% of tokens are then masked, with 80% sampled from positions inside

¹Code available at <https://github.com/lalital/diverse-encoder-ensemble-medexact>

these entities and the remaining 20% drawn at random. DAPT runs for 5,000 optimizer steps with learning rate 5×10^{-5} , batch size 16, and cosine decay with 300-step warmup, yielding the DAPT-BiomedBERT and DAPT-GatorTron variants in the candidate pool.

Loss function. Each encoder is trained with both cross-entropy and focal loss ($\gamma=2$; Lin et al., 2017), which down-weights well-classified tokens and focuses training on ambiguous examples.

Training. Each model is trained independently as a single-task BIO tagger. Non-DAPT variants undergo a single supervised NER fine-tuning stage. DAPT variants undergo two sequential stages: self-supervised MLM (DAPT) followed by supervised NER fine-tuning, rather than a joint objective. We use AdamW (weight decay 0.01, batch size 16) for 8,000 steps with 6% warmup, with learning rates 3×10^{-5} for BiomedBERT variants and 2×10^{-5} for GatorTron variants. The schedule is cosine decay, except for the DAPT-GatorTron variant which uses linear decay. At inference, predictions from all models are combined through whole-span majority voting, retaining only spans that multiple models agree on (§2.4).

Data augmentation. The training set is augmented with two sources, each with its own labelling pipeline.

- **Synthetic notes.** MedGemma-27B (Sellergren et al., 2025) generates clinical notes with inline span markers of the form `<d cat="N">text span</d>`; we parse these markers into BIO labels (104 and 152 documents from two prompt iterations; see Appendix A and Figure 2).
- **Silver-labeled MIMIC-III.** Discharge summaries are labelled by a five-model ensemble of BiomedBERT and GatorTron models fine-tuned on the gold training data (Span F1 50.3 on the validation set).

During training, weighted random sampling draws 30% of each batch from the augmented pool on average. We train each backbone in three settings: train split, train split with synthetic notes, and train split with silver-labeled data, making data augmentation one of the dimensions varied across the candidate pool.

2.3 Input Segmentation

Documents are segmented into overlapping windows of 512 tokens with a stride of 384 during

training, yielding 128 tokens of overlap between consecutive windows. All windows are used as training examples, ensuring full document coverage regardless of length.

At test time, the stride is reduced to 256 tokens so that each token receives predictions from more windows. Overlapping regions are merged via center-priority weighting: for a window of length $L=512$, each token at position t receives weight $w(t) = \exp(-(\frac{t-L/2}{L/2+1})^2)$, assigning higher weight to tokens closer to the window center.

2.4 Ensemble Strategy

We select the final ensemble by greedy forward selection (Caruana et al., 2004) from the candidate pool. The search begins with the highest-scoring individual model and iteratively adds the candidate that yields the largest improvement in final score, with the global voting threshold k re-optimized at every step. Selection terminates when no remaining candidate improves the score. Predictions are aggregated via span-level majority voting: each model’s predicted spans are normalized (expanded to full words, punctuation-stripped, and lowercased) to form (category, text) pairs per document, and a span is included if at least k models predict the same pair. The final configuration uses 10 models with a default threshold of $k=3$. This composition is the deterministic outcome of greedy selection, not a designed coverage constraint: 6 of 10 models use augmented data, all three encoder initializations are represented, and both loss functions are present. A separate exhaustive search over per-category thresholds raises k to 4 for Defining Problem and to 5 for Advice and Precaution, reducing false positives in these high-recall categories.

3 Experiments

3.1 Setup

The MedExACT dataset (Elgaar et al., 2026, 2024) consists of clinical discharge summaries annotated with medical decision spans across 9 DICTUM categories. We use the provided training and validation splits. For the test submission, each model’s best checkpoint (selected on the validation set) is used directly to generate test predictions.

The official evaluation metric, final score, averages the overall base score (mean of Span F1 and Token F1) and the worst-group base score (lowest base score among demographic subgroups defined by race, sex, and language); see Elgaar et al. (2026)

Encoder	Loss	Data	Span F1	Token F1	Final Score [†]
BiomedBERT	CE	+ 104 synthetic	42.7	61.6	50.5
BiomedBERT	Focal	Train split only	41.4	59.8	49.2
BiomedBERT	Focal	+ 104 synthetic	41.8	62.2	51.7
GatorTron	CE	+ 152 synthetic	41.2	62.6	49.1
GatorTron	CE	+ Silver annotated	45.0	63.8	52.7
GatorTron	CE	Train split only	43.9	63.4	51.2
BiomedBERT	CE	+ Silver annotated	44.8	61.1	50.5
GatorTron	CE	+ 104 synthetic	42.4	63.8	52.5
BiomedBERT	CE	Train split only	45.3	62.7	51.8
DAPT-GatorTron	CE	Train split only	45.0	64.1	48.6
10-model ensemble			50.4	62.7	56.9

Table 1: Ensemble models in greedy selection order, with majority vote $k=3$. Selection halted when no remaining candidate improved the validation final score. [†]Subgroup-robust final score on corrected validation set. Data: 104/152 synthetic = MedGemma-27B documents from two generation rounds at 0.3 mixing ratio, silver annotated = silver-labeled MIMIC-III documents, train split only = no augmentation.

for the full specification.

During development, we select checkpoints by final score on the validation set using corrected annotations, where 28 mislabeled Treatment Goal spans in a single Hispanic-subgroup document are reclassified as Defining Problem (§4); test submissions are scored by the organizers against the held-out test set with the original annotations. All models use 512-token sequences and are trained on NVIDIA A100 GPUs.

3.2 Results

Our system achieves a final score of 58.09 on the test set (Span F1 51.81, Token F1 66.66). Greedy forward selection yields a 10-model ensemble (Table 1) covering all three training directions. Notably, DAPT-GatorTron has the lowest score (48.6) yet is selected, suggesting that the ensemble benefits from diverse span predictions, allowing majority voting to screen out individual errors.

Per-category breakdown. Token F1 exceeds Span F1 across all categories (Table 2): the ensemble locates decision-relevant tokens but misaligns span boundaries. Drug Related achieves 91.0 Token F1 but only 57.8 Span F1, reflecting consistent drug vocabulary with variable boundary conventions. Contact Related shows the widest gap (76.1 vs 36.8), likely because referral and follow-up spans vary in how much surrounding context is included. Rare categories (Gathering Information at 12.2/26.2, Deferment at 15.4/23.0) remain low on both metrics, consistent with their low frequency in the training data.

Category	Span F1	Token F1
Contact Related (CR)	36.8	76.1
Gathering Information (GI)	12.2	26.2
Defining Problem (DP)	55.7	77.4
Treatment Goal (TG)	40.0	57.4
Drug Related (Dr)	57.8	91.0
Therapeutic Procedure (TP)	39.3	63.7
Evaluating Test Result (ETR)	35.5	68.9
Deferment (De)	15.4	23.0
Advice and Precaution (A&P)	54.9	70.4

Table 2: Per-category Span F1 and Token F1 of the 10-model ensemble on the corrected validation set.

4 Analysis

Ensemble composition. Greedy selection produces an ensemble covering all three encoder initializations, both loss functions, and all data augmentation variants (Table 1). This coverage emerges from the selection criterion, not by design: each added model differs from existing members along at least one direction.

False negative patterns. Two-thirds of false negatives (1,936 of 2,905) are missed by every individual model in the ensemble (oracle-missed). Over half of these oracle-missed spans (54%) cross line boundaries, and long spans (11+ words) account for 34% of misses compared to 21% for short spans (1–3 words). Multi-line decision spans and long spans appear particularly difficult, even with overlapping sliding windows.

Category confusion. The dominant error pattern is bidirectional confusion between Defining Problem (DP) and Evaluating Test Result (ETR). For example (target spans underlined), “Head CT re-

Error type	Count	%
<i>Which boundary is wrong</i>		
Start only	1,513	62.0
End only	633	25.9
Both	294	12.0
<i>Direction</i>		
Under-extension (pred shorter)	1,487	60.9
Over-extension (pred longer)	953	39.1

Table 3: Boundary mismatch breakdown for the 10-model ensemble on the validation set (2,440 spans with $\geq 30\%$ overlap and matching category).

vealed a SAH with IVH. Neurosurgery was consulted for further management.” is labeled ETR in the annotations but predicted as DP. Conversely, “She was started on tamiflu...and has completed a 5 day course. BPs have been stable to slightly elevated...Pulmonary nodule stable.” is labeled DP but predicted as ETR; the model likely confused a clinical observation for a test result. Both directions reflect the difficulty of distinguishing problem definitions from test findings in clinical text.

Span boundary mismatch. Of 2,440 predicted spans that overlap a gold span of the same category (character-level intersection-over-union ≥ 0.3) but differ in offsets, 62% have only the start boundary wrong, 26% only the end, and 12% both (Table 3). The ensemble exhibits a conservative bias: 61% of boundary errors are under-extensions where the predicted span is shorter than the gold span, versus 39% over-extensions. The missed tokens are often clinically meaningful. For under-extension, the model drops clinical indications, e.g., predicting “Alprazolam. . . as needed for” while the gold span extends to include “Alprazolam. . . as needed for anxiety”; the indication word anchors the prescribing rationale but falls outside the model’s predicted boundary. For over-extension, the model absorbs related findings, e.g., predicting “Mildly thickened mitral valve leaflets. Mild mitral annular calcification. No MS.” while the gold span ends before the negative finding “No MS.”; the model treats the negative exclusion as part of the same evaluative statement.

Boundary error recoverability. Error analysis of the ensemble’s false positives suggests that most are located adjacent to gold spans with incorrect offsets rather than in unrelated text. As an exploratory diagnostic, we trained a BiomedBERT-base boundary refiner on validation data. For each ensemble-predicted span, the refiner embeds it in a context window with special marker tokens indicating the

predicted boundaries. For example, if the ensemble predicts “hydrocortisone 100mg IV” as a Drug span but the correct span is “She was given hydrocortisone 100mg IV”, the refiner input is:

. . .was given [START] hydrocortisone 100mg IV [END] and started. . .

The model predicts where the span should actually start and end, similar to extractive question answering; here it would move the start marker left to include “She was given”. Training data is generated from the validation set by independently shifting the start and end offsets of each reference span by $\pm n$ words ($n \leq 5$); the refiner is trained to recover the original boundary (AdamW, lr 2×10^{-5} , batch size 32, 5 epochs).

On validation data, the refiner raises the final score from 56.90 to 73.79. Because training and evaluation draw from the same validation gold spans, this value is an in-distribution ceiling rather than a held-out generalization estimate. The refiner was applied to ensemble-predicted spans in our test submission.

Annotation quality. Worst-group analysis revealed that a single validation document accounts for the Hispanic subgroup’s low score: 28 spans annotated as Treatment Goal appear to be Defining Problem (e.g., “Denies chest pain...” in Review of Systems, “Regular rate and rhythm” in Physical Exam). Because this subgroup contains only one validation document, the subgroup-robust metric is sensitive to individual annotation inconsistencies.

5 Conclusion

We presented a 10-model diverse encoder ensemble for the MedExACT 2026 shared task, scoring 58.09 on the test set. Greedy selection consistently prefers models that differ from existing members, and error analysis identifies two residual challenges: start-boundary errors (62% of boundary mismatches) and Defining Problem/Evaluating Test Result confusion, both of which may benefit from incorporating clinical domain knowledge, for example through expert-guided annotation refinement or models that explicitly represent clinical reasoning structure.

Limitations

The ensemble, synthetic data generation, and silver labeling are designed for this task. With only 53

validation documents, tuning decisions carry overfitting risk; the test score (58.09) exceeds the validation ensemble score (56.9), but worst-group estimates are noisy as some subgroups contain fewer than ten documents. Because test annotations are held out, we cannot ablate the boundary refiner; the alignment between test and pre-refiner validation scores is suggestive but not controlled.

The synthetic and silver-labeled documents were not validated by clinical experts. Their value is primarily distributional: they expose the encoder to clinical vocabulary patterns and decision-span boundary conventions, not to clinically accurate reasoning. Despite this, six of ten selected members use augmented data, indicating a net benefit for token-level span detection even when the underlying clinical content may be implausible.

Acknowledgements

We used computing resources from the LANTA HPC system at the National Science and Technology Development Agency (NSTDA), Thailand.

Ethics Statement

This work uses de-identified MIMIC-III data (Johnson et al., 2016) under the PhysioNet Data Use Agreement. We release correction scripts for the annotation inconsistencies identified in §4, noting these reflect our judgment and have not been verified by the shared task organizer. Synthetic clinical notes generated by MedGemma-27B may contain clinically implausible content and should not be used for medical decision-making.

References

- Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. 2004. [Ensemble selection from libraries of models](#). In *Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04*. ACM.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024. [MedDec: A dataset for extracting medical decisions from discharge summaries](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16442–16455, Bangkok, Thailand. Association for Computational Linguistics.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Mehrnaz Sadrolashrafi, Mitra Mohtarami, Adrian Wong, Hadi Amiri, and Leo A. Celi. 2026. [Overview of medical decision extraction, analysis, and classification task \(MedExACT\) 2026](#). In *The 25th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, San Diego, California, USA. Association for Computational Linguistics.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. [Domain-specific language model pretraining for biomedical natural language processing](#). *ACM Transactions on Computing for Healthcare*, 3(1):1–23.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don't stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. [MIMIC-III, a freely accessible critical care database](#). *Scientific Data*, 3:160035.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2021. [EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain](#). In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201, Online. Association for Computational Linguistics.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. [Focal loss for dense object detection](#). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.
- Eirik H. Ofstad, Jan C. Frich, Edvin Schei, Richard M. Frankel, and Pål Gulbrandsen. 2016. [What is a medical decision? A taxonomy based on physician statements in hospital encounters: A qualitative study](#). *BMJ Open*, 6(2):e010098.
- Andrew Sellergren and 1 others. 2025. [MedGemma technical report](#). *arXiv preprint arXiv:2507.05201*.
- Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. 2022. [A large language model for electronic health records](#). *npj Digital Medicine*, 5(1):194.

Iteration 1

Brief Hospital Course:

[Multiple units of packed red blood cells (pRBCs)]^{TP} and fresh frozen plasma were administered. [Norepinephrine infusion]^{Dr}, titrated to achieve [target mean arterial pressure (MAP) > 65 mmHg]^{TG}. [A central venous catheter was placed]^{TP} for hemodynamic monitoring. Initial labs demonstrated [hemoglobin of 6.5 g/dL]^{ETR}, [platelets of 80,000/uL]^{ETR}, and [lactate of 4.2 mmol/L]^{ETR}. [An ABG was drawn]^{GI} revealing acute respiratory acidosis. [Bilateral pulmonary contusions and pneumothoraces]^{ETR} on repeat CXR. [Bilateral chest tubes were placed]^{TP} under image guidance. [Fentanyl 100 mcg/hr]^{Dr} and [Propofol 4 mg/kg/hr]^{Dr} were initiated for sedation. [Nephrotoxic antibiotics were deferred pending culture results]^{De}. [Blood cultures were obtained]^{GI} on admission. [Patient remained hemodynamically labile]^{DP} requiring ongoing vasopressor support. [Consultations were made with Trauma Surgery, Pulmonology, and Nephrology]^{CR}.

Discharge Medications:

- [Oxycodone 5mg PO Q4-6H PRN pain]^{Dr}
- [Omeprazole 20mg PO QD]^{Dr}

Discharge Instructions:

[Report any dark or black stools, vomiting blood, or abdominal pain immediately]^{AP}.

Iteration 2

Brief Hospital Course:

Upon arrival, the patient was hypotensive, tachycardic, and tachypneic. Initial labs revealed [significant hyperkalemia (7.5 mEq/L)]^{ETR}. The patient was started on [bicarbonate and insulin with glucose drip]^{Dr}. [Hemodialysis was initiated emergently]^{TP} due to [refractory hyperkalemia]^{DP}. [CXR demonstrated pulmonary edema]^{ETR} consistent with fluid overload. [Maintain adequate hydration to prevent dehydration]^{TG}. [Control blood sugar levels]^{TG}.

[Hyperkalemia]^{DP}. [Volume overload]^{DP}. [End-stage renal disease]^{DP}. [Anemia]^{DP}.

[Renal function labs to monitor kidney health]^{GI}. [Complete blood count (CBC) to monitor hemoglobin level]^{GI}. [Avoid strenuous activity until cleared by your physician]^{De}.

Discharge Medications:

- [Lisinopril 20mg PO daily]^{Dr}
- [Furosemide 40mg PO daily]^{Dr}

Discharge Instructions:

[Monitor blood pressure and heart rate closely]^{AP}. [Strictly adhere to dietary sodium restriction (<2 grams per day)]^{AP}.

Followup Instructions:

[Follow up with Nephrologist within 2 weeks]^{CR}.

Legend: **CR** Contact Related **GI** Gathering Info **DP** Defining Problem **TG** Treatment Goal **Dr** Drug Related **TP** Therapeutic Proc. **ETR** Eval. Test Result **De** Deferment **AP** Advice & Precaution

Figure 1: Annotated excerpts from synthetic discharge summaries generated by MedGemma-27B. Colored brackets denote decision spans; superscripts indicate DICTUM categories. Iteration 1 (left) produces longer narrative text with annotations embedded in clinical sentences; Iteration 2 (right) combines narrative with structured lists. Both documents cover all 9 categories.

A Synthetic Data Generation Details

A.1 Generation Pipeline

Synthetic notes are generated by MedGemma-27B (Sellergren et al., 2025) via vLLM (Kwon et al., 2023) on 4×A100 40GB GPUs (temperature 0.8, top-*p* 0.95, max tokens 4,096, repetition penalty 1.1); the prompt was refined using Claude Opus 4.6. Iteration 1 (104 docs) uses DICTUM definitions, a few-shot example, and per-document demographics; Iteration 2 (152 docs) revises category targets to reduce DP/Dr imbalance (Figure 1).

The prompt (Figure 2) conditions each document on a sampled age group, sex, ICU care unit (weighted toward MICU/CCU), and one of 25 diagnostic profiles (e.g., STEMI, ARDS, sepsis, TBI); demographics are oversampled for underrepresented subgroups. It enforces per-category minimums, boundary rules with correct/incorrect examples, contrastive category disambiguation, negative examples, span length targets, and a few-shot fragment covering all 9 categories.

System: You are a clinical documentation specialist generating synthetic discharge summaries for medical NLP research. . .

User:
 Generate a complete ICU discharge summary with inline annotations marking EVERY medical decision.
 Format: <d cat="N">text span</d>
 Categories: 1=Contact, 2=Test ordered, 3=Diagnosis, 4=Treatment goal, 5=Drug, 6=Procedure, 7=Test result, 8=Deferment, 9=Advice

Patient: {age} {sex}, admitted to {care_unit} for {principal_dx}.
 {secondary_diagnoses and comorbidities}

Requirements:

- Include sections: Chief Complaint, HPI, Brief Hospital Course, Pertinent Results, Discharge Medications, Instructions, Followup
- Mark ALL decisions. Aim for 60-120 annotations.
- Category minimums: ≥8 Cat3, ≥5 Cat5, ≥5 Cat6, ≥5 Cat7, ≥3 Cat2, ≥3 Cat4, ≥2 Cat8

{10 boundary rules, e.g.:
 CORRECT: <d cat="5">She was given hydrocortisone 100mg IV</d>
 WRONG: <d cat="5">hydrocortisone 100mg IV</d> (missing verb)}

{Cat 3 vs Cat 7 disambiguation, e.g.:
 Cat 3: <d cat="3">bilateral pneumonia</d> (diagnosis)
 Cat 7: <d cat="7">CXR showed bilateral infiltrates</d> (test result)}

{Negative examples, e.g.:
 NOT tagged: "BP was 85/35 and drifted as low as 57/35" (context, not a decision)

{Span length guidance, e.g.:
 Short: <d cat="3">COPD</d> Medium: <d cat="1">transferred to [**Hospital**]</d>
 Long: <d cat="8">Bronchoscopy was deferred given clinical improvement</d>
 Target: 30% short, 40% medium, 20% long, 10% very long}

{Few-shot example fragment covering all 9 categories}

Figure 2: Prompt template for synthetic note generation. Italicized sections are condensed; the full prompt includes 10 boundary rules, category disambiguation pairs, and a few-shot example covering all 9 categories.

Eraserhead at PsyDefDetect: Prompt Design and Class Rebalancing for Psychological Defense Mechanism Detection

Muhammad Abu Horaira, Mehreen Rahman, Nahian Chowdhury

Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004029, u2004033, u2004026}@student.cuet.ac.bd

Abstract

We describe the Eraserhead system submitted to the PsyDefDetect shared task at BioNLP 2026, which frames psychological defense level detection as a nine-class utterance classification problem over supportive dialogue. Our system is based on Qwen3-14B and combines clinically informed prompt design, per-label oversampling, and careful inference settings for stable prediction. A central challenge of the task is strong class imbalance, with High-Adaptive responses appearing far more often than several minority classes. This makes it easy for models to favor the majority class and achieve reasonable accuracy while performing poorly on rarer categories. To address this, we iteratively adjusted oversampling targets based on error analysis and predicted label distributions across submission rounds. Our final system achieved an official macro F1 of 0.3418 on Leaderboard 1 and 0.3947 on Leaderboard 2, ranking 7th among the 21 registered teams on both leaderboards. We further analyze the main failure modes of the system, especially the difficulty of distinguishing Minor Image Distorting defenses from High-Adaptive responses and the persistent tendency to over-predict the majority class. These findings highlight the broader difficulty of modeling psychological function from text alone.

1 Introduction

Psychological defense mechanisms are unconscious processes through which individuals manage anxiety and emotional distress in interpersonal interaction (Vaillant, 1992). The Defense Mechanism Rating Scales (Perry, 1990) provide a clinically grounded hierarchy of defense levels. Recent work has also shown growing interest in the use of large language models in psychotherapy related settings (Na et al., 2025).

The PsyDefDetect shared task (Na et al., 2026a) formulates this problem as a nine class utterance

level classification task over supportive conversations. The task is challenging because label distinctions depend on psychological function rather than surface wording, while the class distribution is also highly imbalanced.

Our system, Eraserhead (CodaBench: nahian_abu), combines clinically informed prompt design, per-label oversampling, and careful inference design. The main contributions of this work are as follows:

- We develop a theoretically grounded prompt design based on DMRS, incorporating explicit label definitions and targeted disambiguation rules for fine-grained psychological classification.
- We introduce a per-label oversampling strategy with empirically tuned target distributions to reduce majority class bias and improve representation of minority defense categories.
- We provide a focused analysis of prediction patterns and failure modes, highlighting the difficulty of distinguishing subtle defense mechanisms from text alone.

2 Related Work

Recent advances in NLP have substantially expanded the scope of computational mental health and psychological text analysis. Early work in this area often relied on domain-adapted transformer models trained on mental health-related social media data. A representative example is MentalBERT (Ji et al., 2022), which was pre-trained on social media data and showed improved performance across several mental health detection benchmarks. Despite these results, its reliance on social media corpora introduces domain bias and may limit generalization to other settings, especially supportive or therapeutic conversations. Such models are also

typically optimized for symptom or condition detection rather than for identifying deeper psychological processes.

Subsequent studies explored improved architectures and broader applications. DEENT (Narvaez Burbano et al., 2025) was proposed for depression detection on social media and showed gains over traditional machine learning baselines. Likewise, Ajayi et al. (2025) examined the detection of mental health conditions and cyberbullying from social media using machine learning methods. Another line of work has emphasized domain-adaptive pre-training. Chinese MentalBERT (Zhai et al., 2024), for example, leveraged large-scale social media corpora to improve Chinese mental health text analysis and outperformed several baseline models. Related work on expressive narrative stories (Tang et al., 2024) further suggests that language models can capture subtle linguistic patterns in emotionally rich text, while still showing limitations in deeper semantic reasoning.

Overall, prior work demonstrates that transformer-based models are effective for mental health text classification, especially in social media settings. However, most studies focus on surface-level categorization rather than fine-grained psychological mechanisms. In contrast, our work addresses defense mechanism classification grounded in the Defense Mechanism Rating Scales, where labels may appear similar on the surface but differ in underlying psychological function. This makes conversational context, clinically informed prompt design, and class imbalance handling especially important.

3 Task and Data

PsyDefDetect is a nine-class utterance-level classification task for detecting levels of psychological defense mechanisms in supportive conversations (Na et al., 2026b). Given a short dialogue context and a target utterance from the help-seeker, the model must predict one DMRS label.

The task is challenging for two reasons. First, the distinctions between labels are psychologically subtle and often depend more on underlying function than on surface wording. Second, the label distribution is highly imbalanced, with High-Adaptive responses appearing much more often than several minority classes. This imbalance makes prediction of the rarer categories especially difficult and motivates the rebalancing approach used in our system.

4 System Overview

Our classification system is based on Qwen3-14B¹, which we fine-tuned with 4-bit QLoRA on Kaggle T4 GPUs.

4.1 Model and Fine-Tuning

We fine-tuned Qwen3-14B using Low-Rank Adaptation through the Unsloth library², which enables memory-efficient 4-bit quantized training. LoRA adapters were applied to all attention and feed-forward projection matrices (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj) with LoRA rank 16 and scaling factor 16. Training used AdamW 8-bit with a learning rate of 2×10^{-4} , cosine decay, an effective batch size of 4 (per-device batch size 1 with gradient accumulation of 4), 3 epochs, and a 2,048-token context window.

4.2 Class-Rebalanced Oversampling

The training set is highly imbalanced, with Label 7 (High-Adaptive) dominating several other classes. When trained on the original distribution, the model tends to favor the majority class, producing high accuracy but poor macro F1. This pattern is also visible on the leaderboard, where many systems achieve relatively strong accuracy while still struggling on macro F1.

To address this, we used per-label oversampling with empirically tuned target counts. Minority classes were oversampled to improve representation, while the majority class (Label 7) was capped below its natural frequency to reduce its dominance without removing it entirely. This helped preserve calibration while limiting majority-class bias. These target counts were adjusted through iterative empirical tuning across multiple submission rounds. After each run, we compared the predicted label distribution on the test set with the training prior. Over-predicted labels had their targets reduced, while under-predicted labels had their targets increased. We did not use a separate held-out validation split for these adjustments; instead, oversampling targets were revised across official submission rounds by comparing predicted label distributions with the training prior. Table 1 summarizes this process across two rounds.

¹<https://huggingface.co/unsloth/Qwen3-14B-unsloth-bnb-4bit>

²<https://github.com/unslothai/unsloth>

L	Name	v1	v3 Error	v4
0	No Defenses	200	+5.5%	200
1	Action Defenses	180	-0.5%	180
2	Major Image-Distorting	140	-1.0%	140
3	Disavowal Defenses	130	+6.6%	130
4	Minor Image-Distorting	160	+3.3%	80
5	Neurotic Defenses	120	-0.4%	120
6	Obsessional Defenses	220	+0.8%	220
7	High-Adaptive Defenses	600	-8.3%	900
8	Needs More Information	70	-0.5%	70
Total		1820		2040

Table 1: Oversampling targets across versions. “v3 Error” denotes predicted frequency minus training frequency after the v3 submission, where positive values indicate over-prediction and negative values indicate under-prediction. Boldface highlights the targets changed for v4: Label 4 was reduced after over-prediction, while Label 7 was increased after under-prediction.

One important finding is that Label 7 was difficult to calibrate reliably. In v3, it remained under-predicted relative to its training prior, which motivated increasing its oversampling target in v4. This likely reflects an asymmetry in the DMRS framework: other labels require *positive evidence* for a specific defense mechanism, whereas Label 7 is often assigned when such evidence is absent. As a result, an LLM biased toward cooperative and interpretive readings may still default too easily to a more adaptive classification.

4.3 Prompt Design

Prompt design was the most influential part of the system. The classification prompt evolved through four versions shaped by error analysis from prior submissions: v1 used basic label descriptions, v2 added full label definitions with sub-mechanism names, v3 introduced targeted Critical Distinctions rules, and v4 refined those rules while adding the explicit rule-out step described below.

The final prompt (v4) is structured as a clinical rubric with three parts. Key Principles instructs the model to focus on *psychological function* rather than surface wording and clarifies that direct emotional expression does not by itself constitute a defense. Full Label Definitions list all nine DMRS levels together with their sub-mechanisms, which reduced uncertainty more effectively than high-level descriptions alone. Finally, Critical Distinctions provides eight targeted rules for the label pairs most often confused in earlier submissions:

- *Label 0 vs. 7*: A specific Label 7 sub-mechanism must be identifiable; otherwise, assign 0.
- *Label 1 vs. 7*: Label 1 rejects all offered solutions while complaining; Label 7 is genuinely open to help.
- *Label 3 vs. 4*: Label 3 involves evasion of a *fact or reality*, whereas Label 4 distorts the *image of self or others*.
- *Projection vs. Displacement (3 vs. 5)*: Projection means the speaker is unaware of their own feeling and attributes it to others; Displacement means the feeling is recognized but redirected toward the wrong target.
- *Label 4 vs. 2*: Label 4 preserves some degree of reality testing, whereas Label 2 is more absolute and leaves little room for ambivalence.
- *Label 6 vs. 7 (Isolation of Affect)*: Label 6 implies that emotion is genuinely absent; Label 7 still shows engagement, even when understated.
- *Label 6 (Undoing) vs. 7*: Undoing tends to appear hesitant and self-contradictory rather than analytical.
- *Plain self-disclosure*: Purely descriptive reporting of a past feeling with no distortion or mature coping mechanism defaults to Label 0; genuinely unclassifiable utterances should be assigned Label 8.

The prompt instructs the model to identify the psychological function, rule out the closest alternative, and output a single digit from 0 to 8. This rule-out step was added to discourage the model from selecting the first plausible label.

4.4 Inference

At inference time, the prompt is wrapped in the model’s native chat template with `add_generation_prompt=True` and `enable_thinking=False`. We disabled Qwen3’s thinking mode because, in early experiments, chain-of-thought generation often shifted predictions toward overly charitable High-Adaptive interpretations. Non-thinking mode reduced this tendency and produced more stable, more easily parseable single-digit outputs.

System	LB1-Acc	LB1-F1	LB2-F1
Nürnberg NLP (1/21)	0.7013	0.4200	0.4732
UTS (2/21)	0.6737	0.4055	0.4450
PerceptionLab (3/21)	0.6737	0.3956	0.4402
zzucs (6/21)	0.6441	0.3585	0.4135
Eraserhead (7/21)	0.6462	0.3418	0.3947
zzunlp (8/21)	0.6758	0.3300	0.3909

Table 2: Official leaderboard results. LB1 = Leaderboard 1 for positive classes (Labels 1–8), and LB2 = Leaderboard 2 for all classes. Rankings are reported with respect to the 21 registered teams in the official evaluation.

We decoded with temperature 0.1 and a maximum of 10 new tokens. The predicted label was extracted using a two-stage regular expression: first matching label: $[\theta-8]$, then falling back to the first standalone digit in the valid range.

5 Results

5.1 Main Results

Table 2 presents our official leaderboard results together with selected comparison systems.

Our system achieved an official LB1 macro F1 of 0.3418 and an official LB2 macro F1 of 0.3947, placing it 7th among the 21 registered teams in the official evaluation. Its LB1 accuracy of 0.6462 is also reasonably close to that of several higher-ranked systems, such as zzunlp at 0.6758. This suggests that the model is fairly competitive in overall agreement with the gold labels, but still falls behind the strongest systems in handling minority classes, particularly in terms of precision and recall.

6 Analysis

6.1 The Label 4 Failure: Why Minor Image-Distorting Is the Hardest Class

The clearest weakness in our system is Label 4 (Minor Image-Distorting), which was predicted less often than expected. This class requires the model to detect an *unrealistically* inflated or deflated view of a person while preserving some reality testing, a distinction that often depends on broader emotional context unavailable in a short conversation window.

These cases are difficult because their surface language can still appear reasonable or adaptive. Our model often resolved this ambiguity in favor of Label 7, the safer prediction under the imbalanced training distribution. Reducing the Label 4

oversampling target in v4 then pushed the model toward under-prediction, suggesting that the boundary between Labels 4 and 7 is not well captured by frequency calibration alone.

6.2 Label 7 as a Fallback Class

Even after increasing the oversampling target for Label 7, this class remained difficult to control consistently. This suggests that the problem is not purely one of class frequency, but also of decision boundary definition. In the DMRS framework, other labels require *positive evidence* for a specific defense mechanism, whereas Label 7 is selected when such evidence is weak or absent. As a result, Label 7 becomes a natural fallback under uncertainty. Because LLMs tend toward cooperative and charitable interpretations, they may be especially likely to resolve ambiguous cases in favor of a more adaptive reading.

6.3 Confusion Among Labels 3, 6, and 7

We observe persistent confusion among Labels 3 (Disavowal), 6 (Obsessional), and 7 (High-Adaptive). Although these labels often look similar on the surface, they differ in psychological function: Label 3 avoids an uncomfortable truth, Label 6 is analytical but affectively detached, and Label 7 reflects genuine emotional engagement.

This distinction is difficult to recover from text alone, since affect is often conveyed through prosody and other nonverbal cues that are absent from transcripts. As a result, models fall back on surface heuristics, such as mapping analytical language to Label 6 and positive framing to Label 7. This reflects a broader difficulty of text-only DMRS assessment rather than a limitation of our system alone.

7 Conclusion

We presented Eraserhead, a Qwen3-14B based system for the PsyDefDetect shared task that combines clinically informed prompt design, class rebalanced oversampling, and careful inference settings. Our results show that this approach can achieve competitive performance, while our analysis highlights persistent difficulty around Label 4 and the tendency to over-predict Label 7. More broadly, the task underscores the challenge of inferring psychological function from text alone, especially under strong class imbalance.

Limitations

This work has several limitations. The task is based on text-only dialogue, which excludes prosodic and nonverbal cues that may be important for distinguishing psychological defense mechanisms. This especially affects subtle distinctions among Labels 3, 6, and 7, as well as the boundary between Label 4 and Label 7.

Our class-rebalancing strategy was tuned across official submission rounds by comparing predicted label distributions with the training prior, without using a separate held-out validation split. Although effective in practice, this limits how strongly we can interpret the results in terms of generalization.

Finally, our results are based on a single model family and a specific prompt design, so they may not transfer directly to other architectures or prompting setups. More broadly, reliable text-only classification remains challenging because labels can appear similar on the surface while differing in psychological function.

Ethical Considerations

This work examines automatic detection of psychological defense mechanisms in supportive conversations. Since the task relates to psychological functioning, model predictions should not be treated as clinical judgments. The system is intended only for shared-task research, not for diagnosis, treatment, or mental health assessment.

Misclassification and bias are also important concerns. The model may be affected by class imbalance and may over-predict majority or fallback labels such as High-Adaptive in ambiguous cases. In real-world use, such errors could lead to misleading interpretations of a speaker’s psychological state. Language models may also reflect biases from pretraining data, especially when interpreting emotionally expressive or culturally diverse language.

To reduce these risks, we present the system strictly as a research prototype, report results transparently, and discuss its limitations. Any future use in mental health settings would require human oversight, stronger validation, and proper ethical review.

References

Edward Ajayi, Martha Kachweka, Mawuli Deku, and Emily Aiken. 2025. [A machine learning approach](#)

[for detection of mental health conditions and cyberbullying from social media](#). *arXiv preprint arXiv:2511.20001*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. [Mentalbert: Publicly available pretrained language models for mental healthcare](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. [Overview of the PsyDefDetect shared task at BioNLP 2026: Detecting levels of psychological defense mechanisms in supportive conversations](#). In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. [You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.

Robinson Narvaez Burbano, Oscar Mauricio Caicedo Rendon, and Carlos A. Astudillo. 2025. [An encoder-only transformer model for depression detection from social network data: The deent approach](#). *Applied Sciences*, 15(6):3358.

J. Christopher Perry. 1990. *Defense Mechanism Rating Scales (DMRS), 5th Edition*. McGill University / DMRS Research Group.

Jinwen Tang, Qiming Guo, Yunxin Zhao, and Yi Shang. 2024. [Decoding linguistic nuances in mental health text classification using expressive narrative stories](#). In *2024 IEEE 6th International Conference on Cognitive Machine Intelligence (CogMI)*, pages 207–216.

George E. Vaillant. 1992. *Ego Mechanisms of Defense: A Guide for Clinicians and Researchers*. American Psychiatric Press, Washington, DC.

Wei Zhai, Hongzhi Qi, Qing Zhao, Jianqiang Li, Ziqi Wang, Han Wang, Bing Xiang Yang, and Guanghui Fu. 2024. [Chinese mentalbert: Domain-adaptive pre-training on social media for chinese mental health text analysis](#). *arXiv preprint arXiv:2402.09151*.

Nürnberg NLP at PsyDefDetect: Multi-Axis Voter Ensembles for Psychological Defence Mechanism Classification

Philipp Steigerwald, Eric Rudolph and Jens Albrecht

Technische Hochschule Nürnberg Georg Simon Ohm

{philipp.steigerwald,eric.rudolph,jens.albrecht}@th-nuernberg.de

Abstract

Detecting levels of psychological defence mechanisms in supportive conversations is inherently ambiguous. In the PsyDefDetect shared task at BioNLP 2026 the eight positive defence categories share surface language and differ only in pragmatic function and trained raters reach only moderate inter-annotator agreement. On such a task the decisive lever is not a stronger single model but error independence, since any single representation will waver on the overlapping defence boundaries. We translate this insight into a 9-voter ensemble spanning three orthogonal axes: class granularity (all nine classes for the gatekeeper, only the eight defence classes for the specialists), training method (generative and discriminative) and base model. The system reaches $F1_{test}=.420$ on the hidden test set, placing first among 21 registered teams.

1 Introduction

The PsyDefDetect shared task (Na et al., 2026a) asks a model to classify each seeker utterance in an emotional-support conversation by its level of psychological defence. The PSYDEFCONV corpus (Na et al., 2026b) pairs ESConv (Liu et al., 2021), a corpus of crowdsourced support dialogues, with the Defense Mechanism Rating Scale (DMRS; Perry, 1990)—a clinical taxonomy of eight hierarchical defence levels plus a “No Defence” category—and is evaluated on macro-F1 over classes 1–8. The task is clinically motivated (Perry, 2014) but difficult: trained raters reach only a moderate Cohen’s $\kappa=.639$ and the corpus is heavily imbalanced (C7 covers 52%, the three rarest classes together 7.4%). Several defence categories share surface-level language and differ only in pragmatic function, so the semantic boundaries are inherently fuzzy.

Given these fuzzy boundaries, voting across diverse voters was our starting point. Rather than chasing a stronger single model, we sought voters

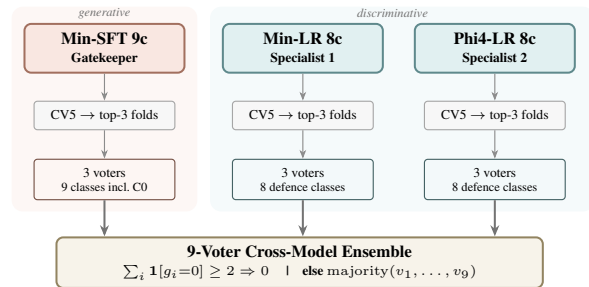


Figure 1: Architecture of our 9-voter cross-model ensemble.

with uncorrelated errors that arbitrate the ambiguity. We tested different training methods (generative and discriminative), several base LLMs and different class granularities. A geometric analysis of the QLoRA-adapted hidden states indicates that only the no-defence class is reliably separable, motivating a *generalist-specialist* split (9-class generalist + 8-class specialists). To counter the heavy imbalance, we additionally augmented the minority defence classes with GPT-5.2 synthetic dialogues.

All training uses 5-fold cross-validation (CV5) as both a voter pool and an internal performance estimate ($F1_{cv}$, mean macro-F1 over classes 1–8) since test labels were hidden. Comparing voters across folds, methods and base models, each axis produces systematically different errors: where some voters get confused on the fuzzy defence boundaries, others succeed and majority voting sharpens those boundaries—the error independence (Dietterich, 2000) we sought, realised in a 9-voter cross-model cross-method ensemble (Figure 1) that reaches $F1_{test}=.420$ (+33.4% over the baseline).

Our contributions: (i) the winning 9-voter cross-model cross-method ensemble; and (ii) an embedding-level analysis quantifying the defence-class semantic overlap that drives task difficulty. Additionally, we release our class-imbalance synthetic dialogues for replication.

2 System

We build our 9-voter ensemble (Figure 1) step by step, adding one voter voice at a time. Each step posed a design choice—which method, which base model, which folds to trust—which we settled by the CV5 signal, guided by the principle that a diversity of voices sharpens fuzzy class boundaries better than any single strong voice.

2.1 Data Augmentation

The PSYDEFCONV training set is heavily imbalanced (C7 covers 52%, the three rarest classes together 7.4%). Our first step was to replicate the organisers’ baseline on a dialog-stratified 80/20 split, and the resulting 1,520-sample training split was augmented with up to $\min(200 - n_c, 3 \cdot n_c)$ GPT-5.2 synthetic dialogues per class—the first term targets 200, the second caps synthetic at 75%. Classes 0 and 7 are excluded as already well-represented; this yields 738 synthetic dialogues. When we later moved to CV5 for the voter pool, we reused the same 738 synthetic dialogues unchanged across all five folds (per-class counts in Appendix Table 4). Validation and test sets (472 samples) remain original human-annotated data only.

2.2 Voting

On this augmented data, voting was our first step toward the voter diversity we kept extending throughout the system. Given the dataset’s fuzzy class boundaries, the decisive lever is error independence between voters rather than a stronger single model; where one voter wavers, another may be more confident on the same sample and majority voting sharpens the joint decision. CV5 provides both a voter pool (five trained models per configuration) and an internal performance estimate ($F1_{cv}$, mean macro-F1 over classes 1–8) since test labels were hidden, with the majority across the five fold-models giving the ensemble prediction. Our first ensemble ran the organisers’ baseline approach (Ministral-8B with generative supervised fine-tuning on all 9 classes of the augmented data, Min-SFT 9c) as 5-fold majority voting, lifting $F1_{test}$ from .315 to .373 (Table 1)—already a substantial gain with no architectural diversity yet present.

2.3 Training Axis

We tested three adaptation methods for fine-tuning a base LLM. **SFT** (supervised fine-tuning) fine-tuned the LLM end-to-end with QLoRA on a gen-

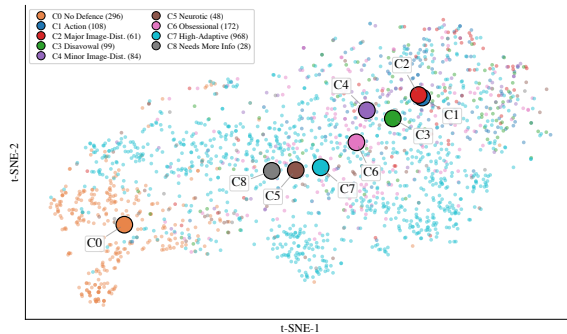


Figure 2: Out-of-fold per-class t-SNE of SFT QLoRA-adapted 9-class Ministral-8B hidden states on the 1,864 original training utterances, with C0 No Defence forming the only well-separated cluster.

erative objective, learning to emit the class digit as text. **ClsHead** (classification head) attached a randomly initialised head to the base LLM and jointly fine-tuned both with QLoRA and focal loss. **LR** (logistic regression) froze the ClsHead-adapted LLM, discarded the trained classification head and fitted a new linear head on the frozen last-token hidden states—architecturally identical to the discarded one but retrained as an L2-regularised logistic regression on frozen features rather than jointly with the backbone. Because LR reused the ClsHead-adapted backbone (extraction in minutes, fit in seconds), it added essentially no compute and let us screen many base-model and class-mode combinations cheaply. All three shared the same input prompt (see appendix).

2.4 Class Granularity Axis

To strengthen the 5-voter baseline we faced two design questions, which classes are reliably distinguishable in the hidden-state space and which configurations to combine for uncorrelated errors.

A t-SNE of the SFT QLoRA-adapted 9-class Ministral-8B hidden states (Figure 2) indicates that C0 No Defence forms the most separable cluster, while the eight defence classes overlap substantially. This motivates a class granularity split. The *gatekeeper* keeps all nine classes, using C0 predictions for the no-defence override and C1–C8 predictions for the defence vote. The 8-class *specialists* focus entirely on the overlapping defences.

For the 8-class specialist we tested seven base models with ClsHead and LR (Appendix Table 3). LR matched or outperformed ClsHead in most cells and achieved the highest $F1_{cv}$ on the majority of base models, becoming the specialist default.

For the gatekeeper, although LR 9c slightly outperforms SFT 9c on $F1_{cv}$, we deliberately chose generative SFT, expecting two different methods to disagree on a different subset of samples than two LR branches would and trading a small per-voter $F1_{cv}$ loss for a larger gain in ensemble error independence. A post-hoc ablation supports this. Pairing the Min-SFT 9c gatekeeper with another SFT specialist (Min-SFT 8c, top-3 folds, $t=2$) yields $F1_{test}=.373$, no improvement over the 5V baseline, while pairing with the discriminative Min-LR 8c lifts it to .391 (Table 1).

We paired the Min-SFT 9c gatekeeper with the Min-LR 8c specialist, keeping only the top-3 folds by $F1_{cv}$ per branch to drop each branch’s most uncertain folds and save 40% inference cost. The resulting 6-voter ensemble fortuitously covers all five folds (Min-SFT 9c $\{f0, f1, f4\}$, Min-LR 8c $\{f0, f2, f3\}$).

With a dedicated gatekeeper, the simple majority vote extends to a two-stage rule. The gatekeeper voters first decide whether the sample is C0 and the remaining defence classification is settled by majority across all voters. Letting g_1, \dots, g_G denote the G gatekeeper predictions and v_1, \dots, v_V all V voter predictions, the ensemble decision is

$$\hat{y} = \begin{cases} 0 & \text{if } \sum_i^G \mathbf{1}[g_i=0] \geq (G+1)/2 \\ \operatorname{argmax}_c \sum_j^V \mathbf{1}[v_j=c] & \text{else} \end{cases} \quad (1)$$

with ties broken in favour of class 7 (the majority class). The gatekeeper voters participate in both branches of Equation 1—triggering the C0-override when a majority of gatekeepers predicts C0, otherwise voting on defence classes alongside the LR specialists and adding method diversity since SFT and LR fail on different subsets of ambiguous samples.

2.5 Model Axis

To extend the diversity principle to the third (model) axis, we tested three additional 8-class LR variants—Phi-4-14B (Phi4-LR 8c), Llama-3.1-8B (Llama-LR 8c) and PsychoCounsel-Llama3-8B (PCounsel-LR 8c, a counselling-domain Llama3-8B finetune)—and ranked them by per-fold Pearson correlation with Min-LR 8c’s $F1_{cv}$ profile (negative values indicate anti-aligned per-fold strengths, contributing independent voter signal). We selected Phi4-LR 8c on the most anti-aligned per-fold profile ($r=-.544$; Llama-LR 8c $+.06$, PCounsel-LR 8c $-.09$), completing the three-axis 9-voter ensemble. With only $n=5$ folds, the gap between

System	$F1_{test}$
Baseline (Na et al., 2026b) (Min-SFT 9c, no-aug)	.315
Min-SFT 9c full-train, augmented (single model)	.307
<i>Voting baseline (no axes)</i>	
5V Min-SFT 9c (5 folds)	.373
<i>Class + training axis</i>	
6V Min-SFT 9c + Min-LR 8c	.391
<i>Class + training + model axis</i>	
6V Min-SFT 9c + Phi4-LR 8c	.391
6V Min-SFT 9c + Llama-LR 8c	.392
9V Min-SFT 9c + Min-LR 8c + PCounsel-LR 8c	.414
9V Min-SFT 9c + Min-LR 8c + Llama-LR 8c	.417
9V Min-SFT 9c + Min-LR 8c + Phi4-LR 8c	.420

Table 1: Hidden-test-set scores ($F1_{test}$, classes 1–8) for our submitted systems (all trained on augmented data), grouped by which diversity axes are active.

Phi4-LR ($r=-.544$) and Llama-LR ($r=+.06$) is indicative rather than statistically decisive; consistent with this, all three candidates land within .006 of each other on the test set (Table 1).

3 Results and Analysis

The exploratory design above, guided by the $F1_{cv}$ signal across methods, base models and folds, produced the configurations in Table 1. The rest of this section asks three questions of the winning 9V—where do its per-class errors concentrate, what does the third specialist arbitrate and how much of the gain depends on augmentation.

3.1 Per-Class Analysis

On the hidden test set ($n=472$), the winning 9V performs strongly on surface-identifiable defences (C0 No Defence $F1_{test}=.899$, C7 High-Adaptive .833) but struggles where categories overlap semantically (Figure 3, Appendix Table 5). The C0-override (Equation 1) fires on 17.6% of test samples, close to the 15.9% training prevalence.

Two error patterns dominate. C6 and C7 are swapped on 28 samples (16 C6→C7, 12 C7→C6) and 7 of 13 C5 Neurotic samples (54%) are labelled C7 High-Adaptive—the highest relative confusion rate. All three classes produce measured, reflective language and distinguishing them needs intent or longitudinal context rather than a single utterance (Perry, 2014). The model therefore defaults to C7—the clinically costly direction, where a neurotic defence read as mature coping misses the signal for intervention.

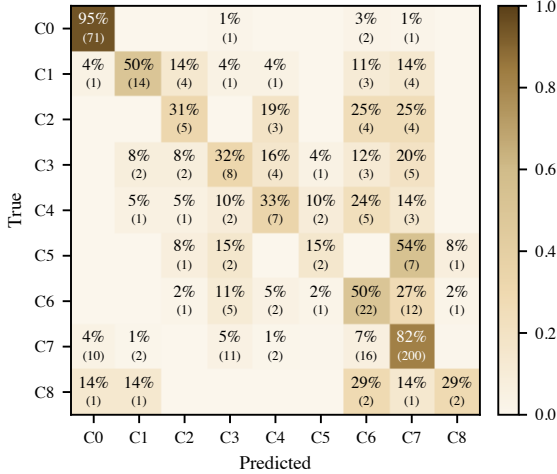


Figure 3: Row-normalised confusion matrix of the winning 9V system on the hidden test set. C7 High-Adaptive absorbs most misclassifications from mid-hierarchy classes.

3.2 Voter Diversity: Flipping and Arbitration

In the 9V the third specialist cannot overrule a confident Ministeral majority since 5/6 and 6/6 are mathematically locked against 3 specialist votes, so it can only intervene on the $n=142$ samples where Ministeral is itself split (Figure 4). The question is whether the specialist flips Ministeral’s wrong calls (helpful arbitration) or its correct ones (harmful noise). Adding Phi4-LR 8c lowers the system Krippendorff’s α from .451 (6V Min-SFT 9c + Min-LR 8c) to .397 (9V) and the drop sits across branches rather than within them (within-branch Min-SFT .617, Min-LR .630, Phi4-LR .464; lowest cross-pair Min-SFT \times Phi4-LR .382). Such cross-branch disagreement among accurate base models is the well-known prerequisite for ensemble gains beyond the strongest member (Dietterich, 2000). But low Krippendorff’s α only proves the voters disagree, not whether they disagree where it matters, so we trace the actual flips.

Phi4-LR flips Ministeral on 39 of those 142 samples and 33 (85%) touch the C6/C7 boundary that dominates the per-class confusion matrix (Figure 3). C7 is the most common flip source (12 of 39, 31%), reflecting Min’s tendency to over-call the majority class, but the redistribution is heterogeneous with no single direction dominating (top flip C7 \rightarrow C3 is 6 of 39). Swapping Phi4-LR for Llama-LR or PCounsel-LR yields .417 and .414, so the gain reproduces across all three candidates.

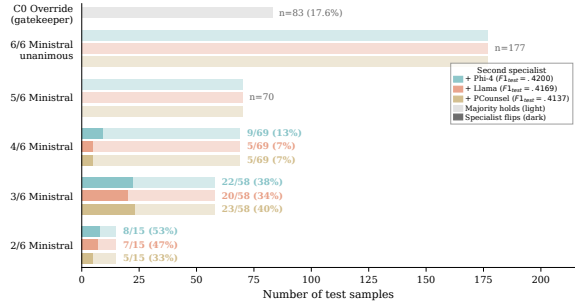


Figure 4: Each bar is one 9V system (6 Ministeral + 3 specialist voters); y-axis groups samples by how many Ministeral voters agreed—the specialist can flip the Ministeral majority only from 4/6 downwards—and dark portions mark the actual flips.

3.3 Augmentation Ablation

Without GPT-5.2 augmentation (*no-aug*; Table 2), the 9V loses only .042. Augmentation alone hurts a single model ($-.008$), but voting averages over the synthetic noise and turns augmentation into a $+.042$ lift on top of voting’s $+.063$ —they are interlocked, not additive.

System	aug	no-aug
Min-SFT 9c (single, no voting)	.307	.315
5V Min-SFT 9c	.373	.319
+ Min-LR 8c (6V)	.391	.369
+ Phi4-LR 8c (9V)	.420	.378

Table 2: Augmentation ablation ($F1_{test}$, classes 1–8); augmentation hurts a single model but helps once voting averages the noise.

4 Conclusion

Our 9-voter ensemble reaches $F1_{test}=.420$ (+33.4% over baseline), driven by voter error independence. C0 forms a well-separated cluster while the defence-class clusters are less distinct—exactly the fuzzy boundaries our approach disentangles. We got there by stacking diversity axes (class granularity, training method, base model) and following the $F1_{cv}$ signal. GPT-5.2 augmentation hurts a single model alone but lifts the 9V by $+.042$. Post-hoc analysis reveals $aug \times no-aug$ as another diversity axis—a 6V combining aug Min-SFT 9c and $no-aug$ PCounsel-LR 8c clears .40 (Appendix Table 6), supporting our hypothesis that independent errors are the lever. Voter diversity tends to help but does not scale arbitrarily. Disentangling these clusters with richer signal and matching the 9V more cheaply remain open.

Limitations

PSYDEFCONV provides only 1,864 training samples, so the +.029 gain from the best non-cross-model 6V (.391) to the 9V (.420) is a single hidden-test observation and what generalises is the complementary-model selection principle, not the exact 9V ranking. Several design choices—specialist selection by an $n=5$ Pearson correlation, top-3 fold selection without cross-validation and the C0-override threshold—rest on limited statistical support, so the submitted configuration is one of several plausible winners. Moderate annotator agreement ($\kappa=.639$) bounds the macro-F1 target and places the rare clinical classes C2, C5 and C8 inside the annotator-disagreement band. All evaluation is on PSYDEFCONV/ESConv (English) with 738 GPT-5.2 synthetic dialogues that carry generator-specific artefacts. The 9V needs distillation for real-time deployment, though the 5V single-branch already captures more than half of the gain at one third of the inference cost.

Ethics Statement

At $F1_{test}=.420$ the system misclassifies the majority of defence-bearing utterances with an adaptive-skew bias that under-flags exactly the patients who most warrant clinical attention. Assigning defence labels to a person’s utterances is itself a psychological intervention and should not occur outside supervised clinical workflows with informed consent (Steigerwald et al., 2026; Na et al., 2025). Like other LLM-based tools in mental health, such systems should augment, not replace, the human practitioner (Steigerwald et al., 2025); outputs are categorical labels only and could be misused to pathologise individuals in adversarial contexts, mirroring known privacy, bias and accountability risks of mental-health LLMs (Steigerwald and Albrecht, 2026). The DMRS taxonomy reflects Western, English-language therapeutic traditions and PSYDEFCONV (Liu et al., 2021) is simulated rather than clinical data whose crowdworkers consented to support-dialogue collection, not to psychodynamic re-annotation.

Data Availability

The 738 GPT-5.2 synthetic dialogues, generation prompt and parameters are released under CC BY-NC 4.0 at <https://github.com/th-nuernberg/nuernberg-nlp-psydefdetect>.

References

- Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *MCS 2000, Lecture Notes in Computer Science*, pages 1–15. Springer.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 3469–3483.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the PsyDefDetect shared task at BioNLP 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.
- J Christopher Perry. 1990. *Defense Mechanism Rating Scales (DMRS)*, 5th edition. Cambridge, MA.
- J Christopher Perry. 2014. Anomalies and specific functions in the clinical identification of defense mechanisms. *Journal of Clinical Psychology*, 70(5):406–418.
- Philipp Steigerwald and Jens Albrecht. 2026. From “Help” to helpful: A hierarchical assessment of LLMs in mental e-health applications.
- Philipp Steigerwald, Nico Bienlein, Jennifer Burghardt, Mara Stieler, Robert Lehmann, and Jens Albrecht. 2025. CAIA in practice: Field evaluation of an AI-assisted support system for text-based online counselling. In *2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE.

Philipp Steigerwald, Jennifer Burghardt, Eric Rudolph, and Jens Albrecht. 2026. AI systems in text-based on-line counselling: Ethical considerations across three implementation approaches.

A Hyperparameter Details

All fine-tuning uses 4-bit NF4 QLoRA on all linear projections (dropout 0.05, cosine schedule with 10% warm-up, 10 epochs, effective batch size 8, max sequence length 4,096 tokens). SFT uses LoRA rank 32 ($\alpha=64$), learning rate 10^{-4} , cross-entropy on the label digit. ClsHead uses rank 16 ($\alpha=32$), learning rate 2×10^{-5} , focal loss (Lin et al., 2017) ($\gamma=2$) with inverse-frequency class weights $w_c = N/(K n_c)$. LR is L2-regularised multinomial logistic regression on frozen last-token hidden states ($C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ swept per fold; class_weight="balanced"; scikit-learn defaults). Augmented samples enter only training.

SFT Prompt Template

System: You are an expert psychologist specialising in the Defense Mechanism Rating Scale (DMRS). You analyse emotional support conversations and classify the psychological defence mechanisms used by the help-seeker in their utterances. You always respond with exactly one line in the format 'label: <number>' where <number> is 1–8.

User: Below is an emotional support conversation between a SEEKER and a SUPPORTER. Your task is to classify the TARGET utterance according to the Defense Mechanism Rating Scale (DMRS).

Conversation

[Full dialogue history with SEEKER and SUPPORTER turns]

Target Utterance

[The seeker utterance to be classified]

DMRS Defence Mechanism Categories

[Definitions for all 8 (or 9) DMRS levels, e.g. "1: Action: The speaker uses action-oriented defences such as acting out, passive aggression, or help-rejecting complaining..."; full descriptions for every level]

Examine the dialogue carefully and select the single most appropriate defence tier. When multiple defences seem plausible, choose the tier with the strongest supporting evidence. Every utterance contains a defence mechanism. Return exactly one line: label: <1–8>.

Assistant: label: {class_id}

B Multi-Model CV5 Comparison

Table 3 reports $F1_{cv}$ for all candidate base models, methods and class modes. The LR column dominates across models and Ministral-8B and Phi-4-14B top the 8-class LR column—the pairing adopted in the winning system.

C Data Augmentation and Class Balancing

GPT-5.2 (temperature 0.9) generated 738 synthetic dialogues against the 80/20 dialog-stratified train

Model	SFT		ClsHead		LR	
	8c	9c	8c	9c	8c	9c
Ministral-8B	.321	.306	.333	.311	.342	.315
Phi-4-14B	–	.293	.337	–	.337	–
Llama-3.1-8B	.251	.279	.246	.284	.312	.284
Qwen2.5-7B	.266	.256	.302	.268	.307	.283
PsychoCounsel-8B	–	–	.316	–	.301	–
PsyLLM-8B	–	–	.295	–	.289	–
GPT-OSS-20B	.212	.183	.278	–	.292	–

Table 3: $F1_{cv}$ for all candidate base models (mean over $n=5$ folds, classes 1–8; $\sigma \in [.019, .045]$).

split (1,520 originals) using a few-shot prompt with the full DMRS taxonomy and five randomly sampled originals of the target class; each dialogue is a 2–6 turn emotional support exchange ending in a target seeker utterance that demonstrates the specified defence level. Table 4 reports the per-class budget against this 80/20 split. The same 738 synthetic dialogues underpin both the “Min-SFT 9c full-train, augmented (single model)” baseline in Table 1 (trained on all 1,864 originals plus the 738 synthetic) and every CV5 voter (each trained on its fold-train split of $\sim 1,493$ originals plus the same 738 synthetic). Synthetic data enters only training splits, while validation and test sets remain exclusively original human-annotated data.

Data Augmentation Prompt (Abridged)

System: You are a psychology expert generating training data. Output only valid JSON.

User: You are an expert psychologist specialising in psychological defence mechanisms.

[Full DMRS taxonomy with definitions and example markers for all 9 levels]

Your Task: Generate $\{n\}$ NEW and DIVERSE examples of Level $\{\ell\}$: **{name}** in emotional support conversations.

Requirements:

1. Create REALISTIC dialogues between a help-seeker and emotional supporter
2. The TARGET UTTERANCE must clearly demonstrate Level $\{\ell\}$
3. VARY the topics: work stress, relationships, health anxiety, family conflict, finances, grief, ...
4. Use natural, conversational English
5. The SEEKER uses the defence mechanism (not the supporter)
6. Dialogue context: 2–6 turns before the target

Few-Shot Examples:

[5 randomly sampled original training examples of the target class, each showing dialogue context + target utterance + label]

Output Format: Generate exactly $\{n\}$ examples as JSON objects, each with a dialogue (list of speaker/text turns) and a target utterance.

D Embedding Geometry after 8-Class Specialist Training

Figure 5 shows per-class t-SNEs for the two LR 8c specialists. 8-class training does not separate the

ID	Defence Level	Orig.	+Aug	Total
0	No Defence	244	0	244
1	Action	88	112	200
2	Major Image-Dist.	54	146	200
3	Disavowal	83	117	200
4	Minor Image-Dist.	67	133	200
5	Neurotic	34	102	136
6	Obsessional	135	65	200
7	High-Adaptive	794	0	794
8	Needs More Info	21	63	84
<i>Total</i>		<i>1,520</i>	<i>738</i>	<i>2,258</i>

Table 4: Per-class composition of the 80/20 train split against which the augmentation budget was computed; C0 and C7 are excluded from augmentation but remain in training.

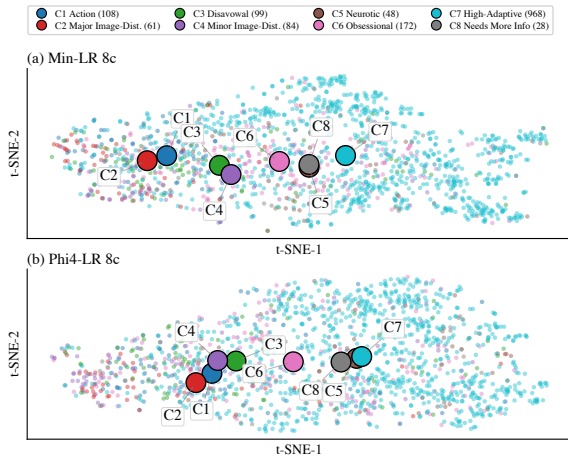


Figure 5: Per-class t-SNE of the LR 8c specialist hidden states ($n=1,568$, C0 excluded).

defence clusters in either model—C6 and C7 remain the dominant overlap (cf. Figure 3)—yet the two models produce visibly different local geometries, consistent with our finding that error independence drives the ensemble.

E Per-Class Results

Table 5 reports per-class scores of the 9V ensemble, with C3 and C5 below .300 and confirming the C7 absorption seen in Figure 3.

F Post-Hoc Re-Voting Search

The configurations in this appendix are not part of our shared-task submission and were never uploaded to the leaderboard. After the test labels were released we re-voted over our 18 cached per-fold prediction sets (12 aug + 6 no-aug) without retraining; an 18-branch search takes seconds on cached predictions. Fold selection is post-hoc, so these

ID	Defence Level	F1	P	R	n
0	No Defence	.899	.855	.947	75
1	Action	.583	.700	.500	28
2	Major Image-Dist.	.333	.357	.312	16
3	Disavowal	.291	.267	.320	25
4	Minor Image-Dist.	.350	.368	.333	21
5	Neurotic	.200	.286	.154	13
6	Obsessional	.436	.386	.500	44
7	High-Adaptive	.833	.844	.823	243
8	Needs More Info	.333	.400	.286	7

Table 5: Per-class $F1_{test}$ of the winning 9V ensemble on the hidden test set ($n=472$).

$F1_{test}$	t	Configuration (gatekeeper + specialists)
6V		
.402	2	Min-SFT 9c + PCounsel-LR 8c (n)
.396	2	Min-SFT 9c + Phi4-LR 8c (n)
.395	2	Min-SFT 9c + Min-LR 8c
9V		
.452	1	Min-SFT 9c + Min-SFTinit-LR 8c + Phi4-LR 8c (n)
.449	2	Min-SFT 9c + Min-SFTinit-LR 8c + Phi4-LR 8c (n)
.445	3	Min-SFT 9c + Min-SFTinit-LR 8c + Phi4-LR 8c (n)
12V		
.471	3	Phi4-SFT 9c + Min-Cls 9c + Min-LR 8c (n) + Phi4-LR 8c (n)
.464	3	Phi4-SFT 9c + Llama-LR 8c + Min-LR 8c (n) + Phi4-LR 8c (n)
.456	3	Phi4-SFT 9c + Min-SFTinit-LR 8c + Min-LR 8c (n) + Phi4-LR 8c (n)

Table 6: Top-3 post-hoc voter combinations on the test set per ensemble size; (n) marks no-aug branches (unmarked = augmented), t is the C0-override threshold; Min-SFTinit-LR is a Min-LR 8c specialist initialised from the SFT 9c adapter rather than ClsHead.

scores are oracle upper bounds rather than blind submissions.

Three patterns emerge (Table 6). First, mixing aug with no-aug branches is itself a diversity axis: 9V ensembles combining aug and no-aug specialists average $F1_{test}=.391$ vs. .372 for aug-only pairs and every top-3 entry from 9V upward draws on at least one no-aug branch. Second, for 6V the best gatekeeper + specialist pairing is the no-aug PCounsel-LR 8c (.402), narrowly ahead of no-aug Phi4-LR 8c (.396); the cross-architecture LR specialists dominate the top of the 6V leaderboard. Third, ensemble gain grows from 9V to 12V (best .452 \rightarrow .471, +.019) and begins to plateau beyond 12V, suggesting that adding more branches from a saturated voter pool re-introduces correlated errors faster than independent signal.

Neural Nexus at PsyDefDetect: Fine-Tuning RoBERTa with Focal Loss and Role-Tagged Dialogue History for Defense Level Detection

Subhrajyoti Basu

Heritage Institute of Technology
subhrajyoti479@gmail.com

Abstract

We describe our system for the PsyDefDetect shared task at BioNLP 2026, which focuses on classifying help-seeker utterances in multi-turn supportive conversations into nine psychological defense mechanism levels defined by the Defense Mechanism Rating Scales (DMRS). Our approach fine-tunes `roberta-base` using a composite training objective that combines focal loss, label smoothing, and square-root dampened class weights to address the severe label imbalance present in the PSYDEFCONV corpus, where the dominant class constitutes 52% of the training data. The input representation is constructed by concatenating up to eight dialogue turns with role-specific tags, separated using RoBERTa’s native `</s>` tokens, followed by the target utterance marked using a `[TARGET]` token. Model selection is performed using macro-F1 based early stopping on a stratified 15% validation split, along with cosine learning rate decay for stable optimization. Our best submission achieves an official Leaderboard 1 (positive classes) macro-F1 score of **0.2556**, ranking 11th among 21 registered teams.

1 Introduction

Psychological defense mechanisms are unconscious strategies through which individuals manage anxiety and regulate emotional distress (Na et al., 2026b). In supportive conversations, a help-seeker’s choice of defense—whether mature coping, intellectualization, or denial—shapes the trajectory of the interaction and the kind of support that is likely to be effective. Automatically detecting these defense levels is therefore relevant for clinical decision support, psychotherapy research, and emotionally intelligent dialogue systems. Despite this relevance, the problem remains largely unexplored in NLP.

The PsyDefDetect shared task at BioNLP 2026 (Na et al., 2026a) introduces this problem through

the PSYDEFCONV dataset, the first conversational corpus annotated with DMRS-based defense levels. The task requires classifying help-seeker utterances into nine hierarchical categories using their dialogue context. This setting connects to ongoing work in psychotherapy-focused language modeling and mental health NLP (Na et al., 2025).

The task presents three main challenges. **(1) Severe class imbalance:** Level 7 (High-Adaptive Defenses) accounts for 52% of the training data, whereas Level 8 (Needs More Information) contains only 28 instances. **(2) Fine-grained clinical distinctions:** Adjacent DMRS levels often show strong surface overlap. For example, Disavowal (Level 3) and Obsessional defenses (Level 6) both involve verbal rationalization, but differ in whether the speaker avoids or intellectualizes affect. **(3) Context dependency:** The same utterance may correspond to different defense levels depending on the preceding dialogue, including both supporter and help-seeker turns.

To address these challenges, we adopt a focal loss objective with square-root dampened class weights, label smoothing, and a context-rich input representation that encodes up to eight dialogue turns with explicit role tags and RoBERTa separator tokens. Our best system achieves a private test macro-F1 score of 0.2556. We further provide per-class analysis and discuss the observed failure modes to inform future work.

2 Related Work

Liu et al. (2021) introduced the ESConv dataset and emotional support strategy classification, which also serves as the source corpus for PSYDEFCONV, while Na et al. (2025) highlight the gap between current NLP capabilities and real clinical requirements in psychotherapy-related applications. In imbalanced text classification, focal loss (Lin et al., 2017) has been widely used to focus

learning on harder and underrepresented examples, and Müller et al. (2019) show that it can work well with label smoothing by reducing overconfidence while preserving learning signal on difficult cases. More broadly, encoder-based dialogue models have shown strong performance in emotion recognition in conversation (Poria et al., 2019), where context dependency and class imbalance are also important challenges. Our input formulation follows this line of work by using role-tagged dialogue history with separator tokens to better capture conversational structure and dependencies.

3 Task and Data

The shared task is based on the PSYDEFCONV dataset (Na et al., 2026b), which is constructed from a stratified subset of the ESConv corpus of emotional support conversations (Liu et al., 2021). The dataset consists of 200 multi-turn dialogues between help-seekers and supporters, with a total of 2,336 help-seeker utterances annotated using the DMRS clinical framework by expert annotators.

Label schema. The task requires classifying each utterance into one of nine categories spanning the full DMRS hierarchy. This includes seven hierarchical defense levels along with two auxiliary categories: “No Defenses”, which corresponds to functional utterances without engagement in internal conflict, and “Needs More Information”, which captures ambiguous cases where the context is insufficient for reliable classification. Table 1 presents the full label distribution and corresponding training frequencies.

Class imbalance. The dataset exhibits significant class imbalance, which directly influences the modeling approach. Level 7 (High-Adaptive Defenses) is the dominant class constituting 52% of the data. In contrast, Levels 1, 2, 3, 4, 5 and 8 appear far less frequently which can be seen in Table 1. This level of imbalance makes standard fine-tuning difficult.

Evaluation. The shared task reports two official leaderboards based on macro-averaged F1 scores. Leaderboard 1 evaluates performance over the positive defense classes (Levels 1–8), while Leaderboard 2 evaluates performance across all classes (Levels 0–8). In both settings, macro-F1 is computed uniformly across classes, making performance on minority classes particularly important for overall system ranking.

Label	Defense Level	Count	%
0	No Defenses	296	15.9
1	Action	108	5.8
2	Major Image-Distorting	61	3.3
3	Disavowal	99	5.3
4	Minor Image-Distorting	84	4.5
5	Neurotic	48	2.6
6	Obsessional	172	9.2
7	High-Adaptive	968	52.0
8	Needs More Info	28	1.5
Total		1,864	100

Table 1: Label distribution in the training set.

4 System Description

4.1 Input Representation

Each instance consists of a dialogue history and a target help-seeker utterance to classify. We concatenate the last $k=8$ dialogue turns, prefixing each turn with an explicit speaker role tag (`Seeker:` or `Supporter:`), and join turns with RoBERTa’s native sentence-separator token `</s>` as a soft semantic boundary. The target utterance is appended after a final `</s>` under a `[TARGET]` `Seeker:` prefix that explicitly marks the utterance to classify:

```
Seeker: t1 </s> Supporter: t2
</s> ... </s>
[TARGET] Seeker: target
utterance
```

All sequences are truncated and padded to a maximum of 256 tokens. Using $k=8$ turns—rather than the shorter windows ($k=5$) trialled in preliminary experiments—is important because defensive functioning often manifests as a *pattern* across multiple exchange rounds. For instance, escalating help-rejecting complaining (Level 1) is more detectable when the model observes a sequence of prior supporter suggestions being consistently deflected. The explicit `[TARGET]` marker prevents the model from attending ambiguously to any seeker turn in the history when classifying.

4.2 Pre-trained Model

We fine-tune `roberta-base` (Liu et al., 2019), a 12-layer Transformer encoder with 125M parameters, appending a two-layer classification head (dense + dropout + output projection) over the `<s>` (CLS) representation for 9-way classification. Hidden dropout and attention dropout are both set to 0.1 to regularize the encoder during fine-tuning on the small corpus.

In preliminary experiments, we also trained `distilbert-base-uncased` (Sanh et al., 2019) as a pilot baseline under a simpler setup with a shorter dialogue history window ($k=5$), standard inverse-frequency class weights, focal loss with $\gamma=2.0$, and no early stopping. The final `roberta-base` system improved over this pilot by approximately +0.04 macro-F1 on the official LB1 leaderboard. However, this improvement reflects not only the stronger encoder backbone but also the revised training and input configuration. RoBERTa’s more aggressive pretraining strategy, including dynamic masking, larger batch training, and the removal of the next-sentence-prediction objective, likely contributes to richer contextual representations for the subtle pragmatic distinctions required by DMRS-level detection.

4.3 Training Objective

Class weights. To counteract majority-class dominance, we compute per-class inverse-frequency weights via scikit-learn’s `compute_class_weight` ("balanced"):

$$w_c = \frac{N}{K \cdot n_c} \quad (1)$$

where N is total training samples, $K=9$, and n_c is the count of class c . Raw balanced weights are extreme for very rare classes (Level 8 yields $w \approx 7.4$), which destabilizes training by amplifying gradient noise from the few available samples. We apply square-root dampening to moderate the range while preserving the relative ordering:

$$\tilde{w}_c = \sqrt{w_c} \quad (2)$$

This keeps weights in the range ≈ 0.6 – 2.7 , which we found empirically more stable than either undampened or cube-root dampening.

Focal loss with label smoothing. We combine focal loss (Lin et al., 2017) with label smoothing (Müller et al., 2019) to address both class imbalance and overconfidence simultaneously.

Label smoothing with $\varepsilon=0.05$ produces a soft target distribution:

$$\tilde{y}_c = \begin{cases} 1 - \varepsilon & \text{if } c = y \\ \frac{\varepsilon}{K - 1} & \text{otherwise} \end{cases} \quad (3)$$

The label-smoothed cross-entropy loss per sample i is:

$$\ell_i = - \sum_{c=1}^K \tilde{y}_c \log p_c^{(i)} \quad (4)$$

The focal weight down-weights confidently classified (high p_t) examples:

$$p_t^{(i)} = p_{y_i}^{(i)}, \quad f_i = (1 - p_t^{(i)})^\gamma \quad (5)$$

with $\gamma=1.5$. We use a lower γ than the original $\gamma=2$ (Lin et al., 2017) to avoid over-suppressing gradient signal from the dominant Level 7 class, which anchors shared encoder representations.

The final loss for a batch of B samples is:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^B \tilde{w}_{y_i} \cdot f_i \cdot \ell_i \quad (6)$$

This objective directs the model’s representational capacity toward hard, often minority-class examples while preserving a meaningful learning signal from easier majority-class instances.

4.4 Training Configuration

We implement the custom objective in a subclassed `HuggingFaceTrainer` (Wolf et al., 2020). The full hyperparameter configuration is given in Table 4 (Appendix A). Key settings:

- AdamW optimizer, learning rate 2×10^{-5}
- Cosine decay schedule, 6% linear warmup
- Weight decay 0.01, gradient clipping 1.0
- Batch size 16 / 32 (train / eval), fp16 precision
- Up to 12 epochs, early stopping with patience 3 on val macro-F1
- Stratified 85/15 train/validation split

At test time, logits are divided by temperature $T=1.3$ before the argmax to reduce overconfidence under the domain shift between training and test data (Guo et al., 2017).

5 Results

5.1 Validation Performance

Early stopping selected the best checkpoint at epoch 6 (of a maximum of 12), with validation macro-F1 of 0.288. Table 2 shows the per-class breakdown on the validation split (280 utterances). The model performs strongly on Level 0 (F1=0.832) and Level 7 (F1=0.677), which together constitute the majority of validation labels. Performance degrades sharply for low-frequency classes, with Levels 1, 3, and 5 showing very low or near-zero F1 scores, highlighting the difficulty of learning from extremely limited data.

L	Defense	P	R	F1
0	No Defenses	0.841	0.822	0.832
1	Action	0.059	0.063	0.061
2	Maj. Img.-Dist.	0.333	0.333	0.333
3	Disavowal	0.000	0.000	0.000
4	Min. Img.-Dist.	0.200	0.154	0.174
5	Neurotic	0.000	0.000	0.000
6	Obsessional	0.263	0.385	0.313
7	High-Adaptive	0.754	0.614	0.677
8	Needs More Info	0.167	0.250	0.200
Macro avg.		0.291	0.291	0.288

Table 2: Per-class results on the validation split (280 utterances, 15% stratified split).

5.2 Private Test Performance

Table 3 compares our final submitted system against a preliminary DistilBERT pilot baseline. The final system achieves an official LB1 macro-F1 of 0.2556. Compared with the DistilBERT pilot run, this corresponds to an improvement of approximately +0.041 macro-F1. However, this comparison should be interpreted as a full-system comparison, since the pilot baseline used a different context window, loss configuration, and optimization setup as mentioned before.

System	Val F1	Official LB1 F1
DistilBERT pilot baseline	0.327	0.215
RoBERTa + focal (ours)	0.288	0.256

Table 3: Macro-F1 results on the validation split and the official LB1 test set. The DistilBERT pilot used a different training configuration, including a shorter context window ($k=5$), different loss hyperparameters, and no early stopping.

6 Analysis

Validation-private gap. Our system achieves a validation macro-F1 of 0.288 but only 0.256 on the private test set, leaving a gap of 0.032. We think this comes mainly from two factors. First, the validation split is small, with only 280 utterances, and some classes have fewer than ten samples, so even a few mistakes can move macro-F1 a lot. Second, repeated hyperparameter tuning across runs likely caused some implicit overfitting to this split, even though each run used fresh initializations. A more reliable setup would be stratified cross-validation or a separate validation set.

Failure modes on rare classes. The model struggles most on Levels 1, 3, and 5, which together have only 255 training samples. Data scarcity is

one reason, but the bigger issue is that these classes often overlap with each other and with more frequent classes. In the validation set, Levels 3 and 5 collapse to $F1 = 0.0$, which shows how hard these labels are to learn from limited data. For example, a statement like “*nobody ever really understands me*” may fit Projection (Level 3, Disavowal) or Passive Aggression (Level 1, Action), depending on the surrounding context. These cases require a deeper understanding of the DMRS framework, not just surface cues. Level 8 is even harder, since it is defined by missing context rather than any clear positive signal, and 28 examples are simply not enough.

Effect of dialogue history window. Using a history window of $k=8$ turns worked better, especially for Action Defenses (Level 1) and Disavowal (Level 3). These defenses often appear as patterns across multiple turns. For example, repeated deflection of suggestions can signal Level 1, while more consistent engagement can align with Level 7. Shorter windows like $k=5$ miss these interaction patterns.

Temperature scaling. Temperature scaling with $T=1.3$ gave a small improvement on the private leaderboard. This is not surprising, since fine-tuned models on small domain-specific datasets often become overconfident. Temperature scaling is a simple fix, but it helps calibration (Guo et al., 2017).

7 Conclusion

We greatly appreciate the dataset as the strong class imbalance makes this task quite challenging. Although we attempted to address this issue, the performance still remains limited. In future work, we plan to explore more reliable ways of handling minority classes including controlled data augmentation methods such as paraphrasing and to incorporate DMRS-specific modeling techniques.

Limitations

Our system has several notable limitations. First, the small training set (1,864 utterances) and severe class imbalance fundamentally constrain fine-tuning approaches; Level 8 with 28 instances is practically unlearnable. Second, our models are encoder-only and operate without any explicit representation of the DMRS clinical framework; they learn surface correlates of defense levels rather than the underlying clinical constructs. Third, the

val/private gap (0.032) indicates unresolved overfitting to the specific validation split used throughout development. Finally, all experiments were conducted on a single GPU (NVIDIA T4, Google Colab), which precluded exploration of larger models or extensive ensemble methods due to time constraints.

Acknowledgments

We thank the PsyDefDetect organizing committee for preparing the PSYDEFCONV dataset and running a well-organized shared task at BioNLP 2026.

References

- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pages 3469–3483.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

A Hyperparameter Configuration

Hyperparameter	Value
Base model	roberta-base
Parameters	125M
Max sequence length	256 tokens
Dialogue history	$k = 8$ turns
Batch size (train/eval)	16 / 32
Learning rate	2×10^{-5}
LR schedule	Cosine decay
Warmup ratio	0.06
Weight decay	0.01
Gradient clipping	1.0
Max epochs	12
Early stop patience	3 epochs
Early stop metric	Val macro-F1
Focal γ	1.5
Label smooth ϵ	0.05
Weight dampening	Square root
Hidden dropout	0.1
Attention dropout	0.1
Temperature T	1.3
Precision	fp16
Hardware	NVIDIA T4 (Colab)

Table 4: Full hyperparameter configuration.

ELiRF-UPV@MedExACT 2026: Dynamic Section Conditioning for Medical Decision Span Detection in Discharge Summaries

Vicent Ahuir[†], Lluís-F. Hurtado^{†,§}, María-José Castro-Bleda^{†,§}

[†]VRAIN: Valencian Research Institute for Artificial Intelligence
Universitat Politècnica de València, Spain

[§]ValgrAI: Valencian Graduate School and Research Network of Artificial Intelligence

[†]vahuir@upv.es, lhurtado@upv.es, mcastro@dsic.upv.es

Abstract

Extracting medical decisions from discharge summaries is essential for downstream clinical analytics, yet the task remains challenging due to the heterogeneous structure of electronic health records. For the MedExACT track at ACL 2026, we proposed a system that achieved the 4th position. Our approach first applies dynamic section conditioning to capture the contextual dependencies inherent in each document. A transformer backbone is then augmented with category- and section-aware layer mixing, enabling us to fuse global document structure with fine-grained semantic cues. To further improve robustness, we employ an ensemble of instruction-tuned large language models for automatic section extraction, while a fairness-oriented model selection criterion ensures that performance does not degrade on minority demographic subgroups. The resulting system attains a final score of 0.5806 on the held-out test set and demonstrates significant gains over the baseline across all evaluated subpopulations.

1 Introduction

The extraction of medical decisions from clinical text is a critical task that holds significant potential for enhancing patient care, advancing medical research, and developing predictive models (Oliveira et al., 2025; Klug et al., 2024). However, this process is fraught with challenges due to the inherent complexity of clinical text, which is characterized by specialized jargon, abbreviations, and diverse formats such as notes and reports (Mortadi et al., 2025; Ando et al., 2022). Discharge summaries, in particular, are rich repositories of medical decisions, yet their extraction remains non-trivial due to the inherent complexity and contextuality of clinical language (Navarro et al., 2023; Hossain et al., 2023).

The MedExACT task at ACL 2026 (Elgaar et al., 2026) builds on this challenge by requiring systems

to identify and classify medical decisions within ICU discharge summaries. The task is grounded in the MedDec dataset (Elgaar et al., 2024), which comprises over 56 000 expert-annotated spans derived from de-identified MIMIC-III clinical notes (Johnson et al., 2016) across nine decision categories. This dataset reflects real-world clinical practice, with significant variations in demographic and linguistic distributions, making it a robust benchmark for evaluating both accuracy and equity in medical decision extraction systems.

This task faces significant challenges: contextual complexity, demographic variability, and overlapping semantics. Contextual complexity demands models that capture fine-grained semantic dependencies across document sections. Demographic variability requires consistent performance across diverse groups to ensure fairness. Overlapping semantics necessitate robust methods to accurately disambiguate decisions.

Our paper addresses these issues through an innovative approach involving dynamic section conditioning and the integration of semantic segmentation with context-aware modeling. This method not only enhances the accuracy of medical decision extraction but also ensures fairness across diverse populations.

The remainder of the paper is structured as follows: Section 2 provides an overview of the task, Section 3 details our methodology, Section 4 presents our experimental results, and Section A describes our validation findings. Finally, Section 6 discusses the implications of our findings and future directions.

2 The Shared Task

The MedExACT shared task was guided by the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM). Participating systems were required to detect contiguous

Table 1: Gold span counts and percentage distribution per demographic subgroup across dataset splits in the MedDec dataset (Elgaar et al., 2024). Test counts (n) are derived from the shared-task evaluation output.

Dimension	Group	Train		Eval		Test	
		n	%	n	%	n	%
Sex	Female	17342	42.1	3262	49.7	2298	37.8
	Male	23831	57.9	3307	50.3	3780	62.2
Race	White	28342	68.8	5159	78.5	4673	76.9
	African American	4078	9.9	902	13.7	313	5.1
	Hispanic	2670	6.5	175	2.7	89	1.5
	Asian	1154	2.8	0	0.0	148	2.4
	Other	4929	12.0	333	5.1	855	14.1
Language	English	25762	62.6	4744	72.2	4134	68.0
	Non-English	15411	37.4	1825	27.8	1944	32.0

text spans and classify them into nine categories: *Contact related, Gathering information, Defining problem, Treatment goal, Drug, Therapeutic procedure, Evaluating test result, Deferment, and Advice/precaution.*

Table 1 shows how gold span counts vary by sex, race, and language across the three splits in the MedDec dataset. In the training set, males dominate (57.9%), but the evaluation split is almost balanced (50.3%), and the test set again skews male (62.2%). Racial composition shifts as well: White subjects rise from 68.8% in training to 76% in both evaluation and test, while African-American spans peak at 13.7% during evaluation and fall to 5.1% in the test set. Language distribution mirrors this trend, with English spans peaking at 72.2% in evaluation and non-English spans being highest in training (37.4%).

In the shared task, the systems were evaluated with a composite metric that balances accuracy and fairness. Accuracy is captured by the Base Score S_{base} , the arithmetic mean of Span- F_1 (exact span-and-label match) and Token- F_1 . Fairness is measured via the Worst-group Score $S_{\text{worst}} = \min_{g \in G} S_{\text{base},g}$, where G denotes all demographic or linguistic subgroups. The Final Score, encouraging uniform performance across groups, is

$$S_{\text{final}} = \frac{S_{\text{base}} + \min_{g \in G}(S_{\text{base},g})}{2}.$$

The baseline was a RoBERTa-based BIO sequence tagger. Because clinical notes average 1600 words, the model processed them in 512-token chunks to respect input limits (Elgaar et al., 2024).

3 Developed System

The developed system comprises a modular pipeline that integrates semantic segmentation of

clinical reports with a context-aware extraction phase. The architecture employs a pretrained transformer encoder to generate a sequence of hidden states across all layers; subsequently, a category- and section-aware layer-mixing module aggregates these states to produce the final embeddings for the classification heads. By utilizing an attention mechanism, the layer-mixing module computes a weighted sum of the encoder’s hidden states, conditioned on the section’s canonical identity and the targeted entity type. This configuration is designed to facilitate the integration of global document structure with local semantic features, providing a framework to manage the contextual dependencies inherent in segmented clinical text.

3.1 Section Extraction

Clinical reports often lack formal partitioning, leading to segmentation strategies that ignore natural semantic boundaries. However, clinical discourse is fundamentally structured around distinct headers –such as *History, Findings, and Impression*– which provide critical inductive biases for entity contextualization. For instance, an entity labeled as a condition in the *History* section implies a past state, whereas the same entity in *Findings* denotes a current observation.

To overcome the brittleness of rule-based parsing, this system delegates section identification to an ensemble of instruction-tuned Large Language Models (LLMs), including *Llama 3.2* (Grattafiori et al., 2024), *Gemma 3* (Team et al., 2025), *Qwen3* (Yang et al., 2025), and *Mistral* (Jiang et al., 2023). These models are leveraged to resolve the stylistic variations and semantic synonymy common in free-text reports. We implement a sequential inference strategy to maintain computational efficiency,

terminating the ensemble query once a valid segmentation is achieved. This process transforms unstructured text into structured records characterized by canonical names and precise character offsets.

To ensure the reliability of these extractions, we utilize a multi-dimensional scoring framework. This metric evaluates the completeness of text coverage, penalizes overlapping segments, and validates that section lengths remain within clinically plausible bounds. By applying a statistical penalty to high variance in segment distribution, the system prioritizes structural consistency, ensuring that internally balanced segmentations are selected over irregular partitions.

3.2 Architecture

To handle the presence of tokens belonging to multiple semantic categories, the architecture employs 9 independent BIO classification heads operating in parallel, one per category. This parallelized structure enables the representation of overlapping spans without specialized decoding, as each head independently predicts the presence of a specific entity type. The system’s training objective is formulated as the mean of the 9 per-head Conditional Random Field (CRF) negative log-likelihood losses.

To provide structured conditioning, section names are clustered with the Hierarchical Density-Based Spatial Clustering of Applications with Noise, HBSCAN, (Campello et al., 2013) into semantically coherent groups using the all-MiniLM-L6-v2 model of SentenceTransformer (Reimers and Gurevych, 2019). During training, headers are grouped into clusters of at least 3; headers in the validation and test sets are subsequently mapped to their nearest training cluster using cosine similarity. This establishes a fixed vocabulary of canonical section identities, allowing the model to incorporate document-level context despite variations in header nomenclature.

Documents are processed at the section level, with each section treated as an independent training instance. Sections exceeding the transformer’s token capacity are partitioned into overlapping windows of 768 characters with a 50-character stride. Annotations are aligned to sub-tokens via offset mapping, with overlapping boundaries resolved greedily to maximize character coverage. Those parts of the text that are not covered by any section are assigned to a dedicated “undefined” section.

The architecture utilizes a pretrained transformer

backbone to extract a sequence of hidden states H . A subsequent layer-mixing module computes a weighted combination of all $L + 1$ hidden states. Using an attention mechanism, the model selects layers based on both the category type and the specific section identity. For a given head, we define a 2-token query sequence $Q = [e_{category}, e_{section}]$, where $e_{section}$ is derived from the section cluster embedding and $e_{category}$ from one of the nine category embeddings. The attention weights α for each layer $l \in \{0, \dots, L\}$ are computed by first calculating the alignment scores:

$$a_{b,s,l} = \frac{1}{2}((W_k H_{b,s,l}) \cdot (W_q e_{category}) + (W_k H_{b,s,l}) \cdot (W_q e_{section}))$$

$$\alpha_{b,s,l} = \frac{\exp(a_{b,s,l})}{\sum_{j=0}^L \exp(a_{b,s,j})}$$

The resulting conditioned representation $h'_{b,s}$ is the weighted sum of the previously extracted hidden states:

$$h'_{b,s} = \sum_{l=0}^L \alpha_{b,s,l} \cdot H_{b,s,l}$$

Final sequence labeling is performed by the 9 labeling heads, each containing a specialized Bidirectional LSTM (BiLSTM) and a CRF layer. This architecture ensures that each head models category-specific sequential patterns.

4 Submitted Runs

We submitted three runs, differing only in encoder, initialization, and model-selection strategy; all were fine-tuned on the shared-task data (LR = 5×10^{-5} , 10% linear warm-up over 5 epochs unless noted).

- **Run 1:** Bio-ClinicalBERT (Alsentzer et al., 2019) encoder; checkpoint chosen to maximize the overall competition score on validation.
- **Run 2:** RoBERTa-large (Zhuang et al., 2021) encoder; same fine-tuning and selection criterion as Run 1, allowing a direct size and domain-vs-general comparison.
- **Run 3:** The fine-tuning process of Run 2, but changing the checkpoint selection to the maximization of the harmonic mean across demographic subgroups to prioritize fairness over aggregate performance.

Validation results are presented in Appendix A.

5 Shared Task Results

Table 2 reports the overall performance of our three runs on the MedExACT @ ACL2026 test set, while Table 3 gives a fine-grained analysis by demographic subgroup. The results show a clear improvement in both accuracy and stability over validation, indicating strong generalization to unseen clinical data.

Table 2: Results for the three submitted runs in the MedExACT @ ACL 2026 Test set.

Metric	Run 1	Run 2	Run 3
Span F1	0.4952	0.5174	0.5237
Token F1	0.6181	0.6610	0.6541
Worst-Group Score	0.5397	0.5650	0.5723
Final Score	0.5482	0.5771	0.5806

Table 2 shows that Run 3 is our best submission, with a *Final Score* of 0.5806 and a *Worst-Group Score* of 0.5723. In contrast to the validation phase (Appendix A), where *RoBERTa-large* lagged behind domain-specific *Bio-ClinicalBERT*, the test results favor the larger general-domain model. Thus, while Run 1 (*Bio-ClinicalBERT*) establishes a strong baseline, Run 3’s scaling and fairness-oriented selection deliver superior overall performance and demographic equity on the test set.

Table 3: Subgroup base scores for the three submitted runs in the MedExACT @ ACL 2026 Test set.

Group	Subgroup	Run 1	Run 2	Run 3
Sex	Female	0.5816	0.6118	0.6149
	Male	0.5413	0.5757	0.5723
Race	White	0.5568	0.5876	0.5778
	African American	0.5419	0.5843	0.5966
	Hispanic	0.5454	0.5650	0.6134
	Asian	0.5460	0.5764	0.5860
	Other	0.5601	0.5988	0.6376
Language	English	0.5670	0.5980	0.5912
	Non-English	0.5397	0.5739	0.5837
Harmonic Mean		0.5530	0.5854	0.5964

Table 3 shows that the bottleneck for the *Hispanic* cohort seen in validation is largely eliminated on the test set: Run 3 scores 0.6134, surpassing many other groups. This confirms that optimizing the *Harmonic Mean* during checkpoint selection balances sensitivity across demographics, and that

Run 3 also attains the highest overall *Harmonic Mean* (0.5964).

Table 4: Top 5 participants on the MedExACT @ ACL 2026 test leaderboard. Subscripts indicate the rank position for each metric among the five participants. The baseline (BL) is shown for reference. The columns refer to: Final Score (F-S), Baseline Score (B-S), Worst-Group Score (WG-S), Span F1 (S-F1), and Token F1 (T-F1).

#	F-S	B-S	WG-S	S-F1	T-F1
1	0.5965 ₁	0.6043 ₁	0.5886 ₁	0.5419 ₁	0.6667 ₃
2	0.5942 ₂	0.6003 ₂	0.5881 ₂	0.5257 ₂	0.6750 ₂
3	0.5809 ₃	0.5924 ₃	0.5695 ₄	0.5181 ₄	0.6666 ₄
*4	0.5806 ₄	0.5889 ₄	0.5723 ₃	0.5237 ₃	0.6541 ₅
5	0.5724 ₅	0.5848 ₅	0.5601 ₅	0.4900 ₅	0.6796 ₁
BL	0.5111	0.5301	0.4922	0.4363	0.6238

Table 4 places our best submission (Run 3) 4th overall, surpassing the official baseline by +0.0695 in *Final Score* and +0.0801 in *Worst-Group Score*. Although we rank 5th on Token-F1, our 3rd-place *Worst-Group Score* underscores the model’s effectiveness at achieving robust, equitable clinical entity extraction.

6 Conclusions

We demonstrated that dynamic section conditioning—semantic segmentation of clinical documents followed by a context-aware encoder—substantially improves medical decision extraction while maintaining demographic equity. Across three experimental runs, the *RoBERTa-large* model with a fairness-focused selection strategy (Run 3) achieved the highest Final Score (0.5806) and Worst-Group Score (0.5723), outperforming both the baseline and other configurations. The improvement from validation to test data confirms strong generalization to unseen clinical records.

Maximizing the harmonic mean across demographic subgroups in Run 3 not only boosted overall performance but also ensured equitable outcomes, illustrating that fairness-aware model selection can enhance both accuracy and inclusivity.

Future work will focus on reducing computational complexity (e.g., via distillation or quantization) and extending robustness evaluations to diverse clinical settings, thereby increasing the real-world applicability of our approach. In addition, we plan to conduct a thorough ablation study to better understand the contribution of key components.

This will include analyzing the impact of varying the number of retrieved similar sections on final performance, as well as systematically removing or varying components such as section conditioning, ensemble models, and layer mixing.

Limitations

While our approach shows strong performance overall, several limitations should be noted.

First, using multiple large language models (LLMs) for section extraction creates significant computational overhead—each model must run separately, increasing latency and GPU memory demands, which hampers scalability in resource-constrained clinical settings.

Second, the method assumes reliable section identification. In practice, documents vary widely in formatting and terminology; misidentified sections can propagate errors to downstream tasks such as entity or relation extraction.

Third, although promising on MedDec, performance on other clinical domains (radiology, pathology, and outpatient notes) has not yet been evaluated. Different specialties use distinct vocabularies and structures that may challenge both the heuristics and the LLMs' adaptability without fine-tuning.

Finally, reliance on large pre-trained models raises privacy and regulatory concerns. Even with local inference, model parameters can contain sensitive patterns from training data, requiring careful audit and mitigation before deployment in regulated environments.

Acknowledgments

This research is supported by Grant PID2024-155948OB-C55 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

References

Emily Alsentzer, John Murphy, William Boag, Weihung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. [Publicly available clinical BERT embeddings](#). In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Kenichiro Ando, T. Okumura, Mamoru Komachi, H. Horiguchi, and Yuji Matsumoto. 2022. [Is artificial intelligence capable of generating hospital discharge](#)

[summaries from inpatient records?](#) *PLOS Digital Health*, 1.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024. [MedDec: A dataset for extracting medical decisions from discharge summaries](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16442–16455, Bangkok, Thailand. Association for Computational Linguistics.

Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Mehrnaz Sadrolashrafi, Mitra Mohtarami, Adrian Wong, Hadi Amiri, and Leo A. Celi. 2026. [Overview of medical decision extraction, analysis, and classification task \(medexact\) 2026](#). In *The 25th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, San Diego, California, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.

Elias Hossain, R. Rana, N. Higgins, J. Soar, P. Barua, Anthony R. Pisani, and K. Turner. 2023. [Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review](#). *Computers in biology and medicine*, 155:106649.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.

Alistair Johnson, Tom Pollard, and Roger Mark. 2016. [MIMIC-III Clinical Database](#). *PhysioNet*. Version 1.4.

Katrin Klug, Katharina Beckh, Dario Antweiler, Nilesh Chakraborty, Giulia Baldini, Katharina Laue, R. Hosch, F. Nensa, Martin Schuler, and Sven Gieselbach. 2024. [From admission to discharge: a systematic review of clinical natural language processing along the patient journey](#). *BMC Medical Informatics and Decision Making*, 24.

Ahmad Mortadi, Waleed Nazih, Mohamed I. Eldesouki, and Yasser Hifny. 2025. [Intelligent de-identification of medical discharge summaries using hybrid nlp](#)

techniques. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24:1 – 17.

David Fraile Navarro, K. Ijaz, Dana Rezasadegan, H. Rahimi-Ardabili, M. Dras, E. Coiera, and S. Berkovsky. 2023. [Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review](#). *International journal of medical informatics*, 177:105122.

Juliana Damasio Oliveira, H. D. P. Santos, Ana Helena D. P. S. Ulbrich, Julia Colleoni Couto, Marcelo Arocha, Joaquim Santos, Manuela Martins Costa, Daniela Faccio, F. Tabalipa, and Rodrigo F. Nogueira. 2025. [Development and evaluation of a clinical note summarization system using large language models](#). *Communications Medicine*, 5.

Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. [A robustly optimized BERT pre-training approach with post-training](#). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Validation Results

The validation results for the three submitted runs are presented in Table 5, showing the overall performance of the models. The *RoBERTa-large* configurations (Run 2 and Run 3) yielded higher Span F1 and Token F1 scores than the domain-specific *Bio-ClinicalBERT* (Run 1), yet Run 1 reached the highest Final Score of 0.4732. The *Worst-Group Score* was highest in Run 1 (0.4089), suggesting that the domain-specific representations in this run

Table 5: Validation set results for the three submitted runs.

Metric	Run 1	Run 2	Run 3
Span F1	0.4802	0.5013	0.5013
Token F1	0.5947	0.6153	0.6294
Worst-Group Score	0.4089	0.3772	0.3538
Final Score	0.4732	0.4677	0.4596

maintain a higher performance floor on the validation set compared to the larger general-domain models.

Table 6 provides the breakdown of scores across sex, race, and language subgroups. It shows that selecting for the Harmonic Mean in *Run 3* results in the highest consistency across most categories, notably improving the *Other* race subgroup to 0.5129. However, this optimization for parity across the majority of groups coincided with the lowest observed score for the *Hispanic* subgroup (0.3538), which remained the performance bottleneck across all three runs. While Run 3 achieved the highest overall Harmonic Mean (0.5221), the validation data suggests a trade-off: increasing broad demographic consistency could lower the score of the lowest-performing subgroup.

Table 6: Validation set subgroup base scores for the three submitted runs.

Group	Subgroup	Run 1	Run 2	Run 3
Sex	Female	0.5617	0.5819	0.5861
	Male	0.5138	0.5353	0.5452
Race	White	0.5469	0.5673	0.5726
	African American	0.5429	0.5712	0.5739
	Hispanic	0.4089	0.3772	0.3538
	Other	0.4109	0.4419	0.5129
Language	English	0.5250	0.5512	0.5577
	Non-English	0.5639	0.5736	0.5803
Harmonic Mean		0.5014	0.5138	0.5221

VISHC AT PSYDEFDETECT: Mitigating Data Scarcity in Psychological Defense Classification with Context-Aware Synthetic Augmentation

Hoang-Thuy-Duong Vu^{1,2}, Quoc-Cuong Pham^{1,2}, Huy-Hieu Pham^{1,2,3}*

¹College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam

²VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam

³Center for Innovations in Health Sciences, VinUniversity, Hanoi, Vietnam
{26duong.vht, 24cuong.pq, hieu.ph}@vinuni.edu.vn

Abstract

Psychological defense mechanisms (PDMs) are unconscious cognitive processes that modulate how individuals perceive and respond to emotional distress. Automatically classifying PDMs from text is clinically valuable but severely hindered by data scarcity and class imbalance, challenges which generative augmentation alone cannot resolve without psychological grounding. In this work, we address these challenges in the PsyDefDetect shared task (BioNLP@ACL 2026) by proposing a context-aware synthetic augmentation framework combined with a hybrid classification model. Our hybrid model integrates contextual language representations with basic clinical features, along with 150 annotated defense items. Experiments demonstrate that definition quality in prompting directly governs generation fidelity and downstream performance. Our method surpasses DMRS CO-PILOT, reaching an accuracy of 58.26% (+40.25%) and a macro-F1 of 24.62% (+15.99%), thereby establishing a strong baseline for psychologically grounded defense mechanism classification in low-resource settings. Source code is available at: <https://github.com/htdgv/CASA-PDC>.

1 Introduction

Psychological Defense Mechanisms (PDMs) present a unique challenge for Natural Language Processing (NLP) field, particularly, they are unconscious, context dependent processes that appear through subtle cues such as narrative inconsistency, shifts in emotional framing, and distorted attribution, rather than clear lexical markers (Vaillant, 1994; Cramer, 1987; Bond et al., 1983). This implicit nature creates semantic ambiguity in which identical surface text may reflect distinct defensive processes, depending on underlying intent and psychological context, leading standard token or

sentence level models to conflate adaptive coping with maladaptive defenses.

Data scarcity and class imbalance further compound these difficulties. Synthetic augmentation via Large Language Models (LLMs) offers a natural remedy, yet without psychologically grounded constraints, generative models produce fluent but theoretically invalid text, creating hallucinating defenses that introduce noise and erode model reliability (Ji et al., 2023; Na et al., 2025; Anaby-Tavor et al., 2020; Kumar et al., 2020). A key point is that the PsyDefDetect shared task (Na et al., 2026a) on the PSYDEFCONV dataset (Na et al., 2026b), based on the ESCONV dataset (Liu et al., 2021), introduces two auxiliary labels, *No Defense* (Level 0) and *Need More Information* (Level 8), that carry no corresponding clinical defense items (Di Giuseppe and Perry, 2021). These labels violate standard multi-class assumptions and produce skew distributions, making defense-item-based feature extraction underspecified, demanding a principled reformulation of the task.

We address these challenges in the PsyDefDetect shared-task through context-aware synthetic augmentation paired with a Hybrid Feature Fusion architecture. Specifically, our contributions are:

- **Psychologically grounded augmentation.** Stressor-anchored, theory-driven prompts with class-specific definitions from the Defense Mechanisms Rating Scales (DMRS) for synthetic augmentation to ensure generating high-fidelity examples, demonstrating that definition quality in prompting governs downstream performance.
- **Clinical feature engineering.** Structured features from all 150 defense items, along with basic clinical features, are fused with contextual language representations, bridging clinical theory and neural classification.

*Corresponding author: hieu.ph@vinuni.edu.vn

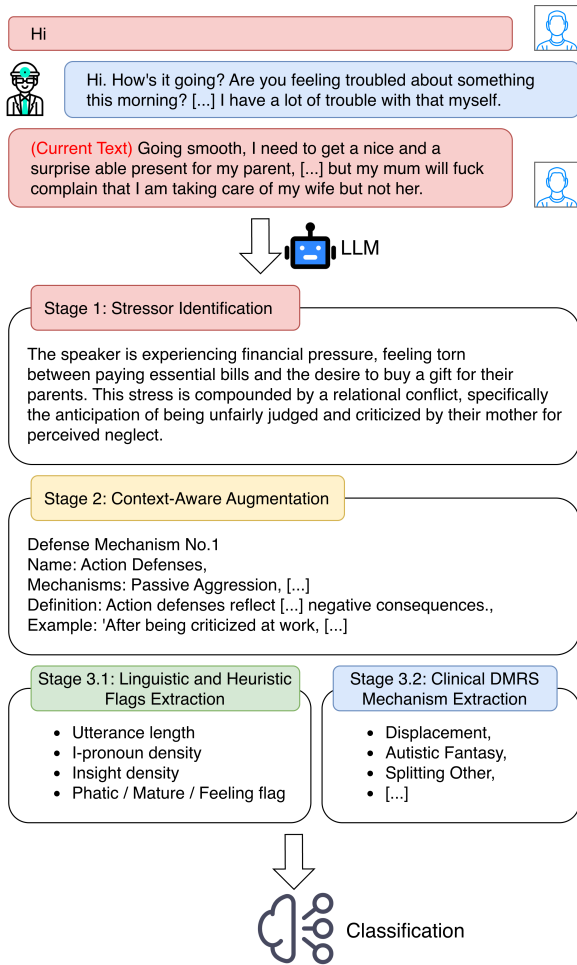


Figure 1: Overview of the multi-stage research pipeline. The process begins with (1) LLM-based stressor identification to establish contextual grounding; (2) context-aware synthetic data augmentation to address class imbalance; and (3) a dual-domain feature extraction stage targeting linguistic heuristics and clinical DMRS mechanism indicators; followed by the final classification.

- **Strong low-resource baseline.** Using Llama-3-8B-Instruct as data generator, our system improves accuracy (18.01% to 58.26%) and macro-F1 (8.63% to 24.62%) on the PSYDEFCONV blind-test set, establishing a competitive foundation for PDM classification.

2 Methodology

We address PDM classification under data scarcity through three stages: (1) a context aware synthetic data augmentation pipeline grounded in clinical theory; (2) a dual-domain feature extraction stage, combining linguistic heuristics with DMRS derived defense profiles; and (3) a hybrid fusion architecture that integrates contextual language representations with structured clinical features (Figure 1).

Class	Label	N	SB \downarrow	SA \uparrow
No Defense	0	500	0.399	0.481
Action	1	500	0.496	0.585
Major Image	2	488	0.440	0.559
Disavowal	3	500	0.453	0.597
Minor Image	4	500	0.416	0.592
Neurotic	5	384	0.392	0.525
Obsessional	6	500	0.429	0.619
Needs Info	8	224	0.413	0.601
Avg.			0.430	0.570

Table 1: Synthetic data quality per class. **SB**: Self-BLEU (Zhu et al., 2018) (lower = greater lexical diversity); **SA**: Semantic Adherence via Natural Language Inference (NLI) entailment (higher = stronger label alignment). Class 7 (High-Adaptive) is excluded from augmentation due to sufficient original samples; all other classes are capped at $N=500$ synthetic instances.

2.1 Context-Aware Data Augmentation

Standard augmentation methods such as paraphrasing and back translation (Wei and Zou, 2019) introduce diversity in phrasing but do not retain the functional role of defense mechanisms. We present a Synthetic Data Augmentation (SDA) pipeline based on Llama-3-8B-Instruct that captures the psychological conditions that give rise to defenses, with emphasis on function instead of form.

Stressor-Anchored Generation. Defense mechanisms emerge in response to perceived stressors rather than in isolation. Each prompt is anchored in a key stressor identified from the dialogue, such as interpersonal conflict, job loss, or social rejection. This approach promotes responses that reflect realistic defensive behavior instead of generic expressions of emotion.

Theory-Driven Prompting. To control semantic drift and reduce label inconsistency, each prompt defines the target defense level using structured clinical details from the DMRS framework (Di Giuseppe and Perry, 2021). These details include the defense name, its formal definition, and common linguistic and behavioral patterns. Paired with few-shot examples (Appendix A.3), this design steers generation toward samples that express the theoretical role of each defense rather than relying on surface plausibility (Brown et al., 2020).

2.2 Data Quality Control

Uncontrolled generation may introduce label noise and artifacts. To mitigate this, we use two quality control steps. First, a soft balancing scheme restricts each minority class to 500 total samples

(real plus synthetic), which helps reduce overfitting to generation specific patterns. We examine five augmentation settings: $\times 1$, $\times 2$, $\times 5$, $\times 8$, and $\times 10$, together with the 500 cap variant. Second, a machine as annotator filter applies a secondary classifier to assign labels to generated batches; Only batches achieving a Cohen’s Kappa of $\kappa \geq 0.60$ (Cohen, 1960), reflecting substantial agreement, are retained (Table 1).

2.3 Feature Extraction

Each seeker utterance is represented by two complementary feature sets: (i) lightweight linguistic heuristics capturing surface-level cues, and (ii) clinically grounded DMRS-derived features encoding latent defensive functioning.

Linguistic and Heuristic Features. We define six lightweight features to distinguish non-defensive (Label 0) from defensive responses, which are often conflated: *Utterance Length* (narrative elaboration proxy), *I-Pronoun Density* (self-focus), *Insight Density* (reflective reasoning), *Phatic Flag* (short filler utterances), *Mature Coping Flag* (triggered by length > 12 , high insight, and elevated I-pronouns), and *Emotion Intensity* (model confidence in non-neutral predictions).

DMRS Defense Profile. We approximate latent defensive functioning using a four-step indicator inference procedure:

1. *Indicator Scoring:* An NLI model estimates entailment probability $P(T \Rightarrow I_j)$ for each of 150 DMRS indicators given utterance T .
2. *Mechanism Aggregation:* Indicator scores are grouped into 30 defense mechanisms and normalized to form mechanism scores $S(M_k)$.
3. *Profile Construction:* The resulting 30-dimensional vector defines the Defense Profile of the utterance.
4. *Level Mapping:* Mechanism scores are aggregated by DMRS level to obtain the predicted defense level: $\hat{y} = \arg \max_{\ell} \sum_{M_k \in \ell} S(M_k)$.

2.4 Hybrid Feature Fusion Architecture

Our system integrates contextual language representations with structured clinical features using a late fusion approach, as described below:

1. **Textual Encoder:** MentalRoBERTa (Ji et al., 2022) encodes each instance formatted as

[Stressor:S|Turn:T], conditioning the representation on both the triggering context and the response, yielding a 768-dim embedding.

2. **Feature Encoders:** The heuristic (7-dim) and DMRS-derived (30-dim) features are each passed through a dedicated Multilayer Perceptron (MLP) with the following structure: 64 \rightarrow Batch Normalization \rightarrow ReLU \rightarrow Dropout ($p = 0.3$) \rightarrow 32, producing two 32-dim vectors.
3. **Fusion and Classification:** The three representations are concatenated into an 832-dim vector (768 + 32 + 32) and passed through two fully connected layers (256 \rightarrow 128, ReLU + Dropout ($p=0.4$)) and a final linear layer producing a probability distribution over 9 labels (Kiela et al., 2020).

3 Experiment & Results

3.1 Experimental Setup

Data Pre-processing. The training corpus combines human-annotated dialogues with synthetic samples generated by Llama-3-8B-Instruct under theory-driven prompting. We evaluate six augmentation scales, $\times 1$, $\times 2$, $\times 5$, $\times 8$, $\times 10$, and a hard cap of $N=500$ per class, yielding corpora ranging from approximately 1,800 to 5,100 instances. Class 7 is excluded from augmentation given its already substantial representation. The baseline corresponds to the organizer-provided code rerun with Llama-3-8B-Instruct and no augmentation.¹

Implementation Details. The model is implemented in PyTorch using the Hugging Face Transformers library. MentalRoBERTa (mental-roberta-base) (Ji et al., 2022) serves as the textual encoder; its parameters are fine-tuned end-to-end with a learning rate of 1×10^{-6} . Task-specific layers (MLPs, fusion head, and classifier) use a higher learning rate of 1×10^{-4} , optimized with AdamW (Loshchilov and Hutter, 2019). Training runs for up to 20 epochs with batch size 16, early stopping on validation macro-F1, weight decay of 1×10^{-2} , and label smoothing ($\epsilon = 0.1$) to mitigate noise introduced by synthetic samples.

Evaluation Metrics. We follow the same protocol employed in Na et al. (2026a) and report all results on both the PSYDEFCONV development and blind test sets using macro-averaged Precision,

¹Per-class metrics and DMRS activation patterns are detailed in Appendix A.

Setting	Acc \uparrow	P \uparrow	R \uparrow	F1 \uparrow
<i>Baseline</i>				
DMRS CO-PILOT*	0.1801	0.1904	0.1715	0.0863
<i>Our System</i>				
$\times 1$	0.5508	0.2555	0.2601	0.2543
$\times 2$	0.5508	0.2789	0.2882	0.2799
$\times 5$	0.5487	0.2764	0.2821	0.2783
$\times 8^\dagger$	0.5826	0.2588	0.2503	0.2462
$\times 10$	0.5254	0.2237	0.2289	0.2238
$N=500$	0.5275	0.2659	0.2654	0.2628

Table 2: Classification performance on the PSYDEFCONV official test set. *Baseline rerun with Llama-3-8B-Instruct on the released test set; original results in Na et al. 2026b use a different backbone. $\times k$: each minority class expanded to k times its original size. $N=500$: hard cap of 500 instances per class. Metrics are macro-averaged. \dagger Official leaderboard submission; all other rows are post-hoc evaluations on the released test set. Best results per column in **bold**.

Recall, and F1, as well as overall Accuracy, to account for class imbalance.

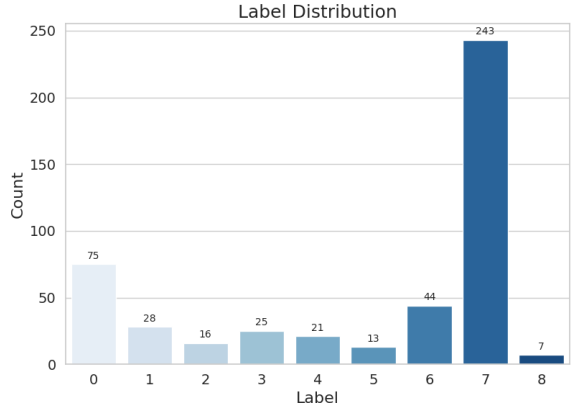
3.2 Results Analysis

Classification performance across settings.

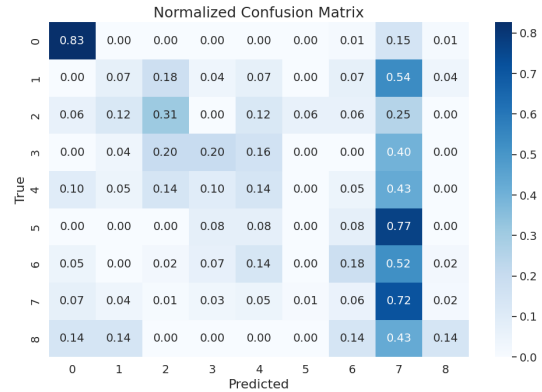
Our submission ranked 13 out of 21 registered teams in the official evaluation. Table 2 reports results across all six augmentation configurations on the official blind-test set. Every augmented variant substantially outperforms DMRS CO-PILOT in accuracy (+40.25 pp) and macro-F1 (+15.99 pp), confirming that theory-driven augmentation delivers robust gains over a prompt-only LLM baseline in this low-resource setting. Performance improves at lower augmentation scales but deteriorates as augmentation becomes more aggressive. The $\times 2$ configuration yields the highest macro-F1 (27.99%), indicating an effective balance between expanded class coverage and synthetic generation noise. Further scaling leads to a steady decline in macro-F1, which falls to 22.38% at $\times 10$, consistent with noise accumulation in heavily augmented corpora (Kumar et al., 2020). While $\times 8$ records the highest accuracy (58.26%), its macro-F1 remains 3.37 pp below $\times 2$, revealing that overall accuracy is disproportionately influenced by dominant Label 7 predictions at the expense of minority-class recall.

The Label 7 sink effect and class imbalance.

The confusion matrix (Figure 2b) confirms Label 7 as a universal prediction sink. The most severe case is Label 5 (Neurotic), where 77% of instances are misclassified as Label 7, rendering this class unlearnable. Per-class F1 scores shows that La-



(a) Label distribution of the PSYDEFCONV official test set.



(b) Row-normalized confusion matrix of our official leaderboard submission (PSYDEFCONV test set, $\times 8$).

Figure 2: (a) The PSYDEFCONV official test set label distribution and (b) row-normalized confusion matrix of our official leaderboard submission ($\times 8$). Label 7 dominates both the distribution (243/472 instances) and predictions, absorbing errors from all other classes.

Labels 0 and 7 exceed $F1 > 0.70$, while all remaining classes fall below 0.30, with four classes below 0.15. This implies that the accuracy (0.55-0.58) substantially overstates practical utility. The Label 5/7 confusion is semantically meaningful: both involve reflective discourse, but differ in whether anxiety is intellectualized or channeled.

The primacy of definition quality. A key finding is the sensitivity of model performance to prompt design. Compared to the baseline of Na et al. (2026b), which uses shallow class descriptions and achieves a Macro-F1 of 8.63%, our best setting ($\times 2$) improves by 19.36 pp. We attribute this to the richer definitional context provided by our DMRS-based definitions derived from Di Giuseppe and Perry (2021), which better disambiguate overlapping classes and reduce label ambiguity during generation.

4 Conclusion

Clinical specificity of class definitions is the primary driver of synthetic augmentation effectiveness in PDM classification. Our hybrid system, combining MentalRoBERTa with DMRS-derived features and stressor-anchored generation, achieves substantial gains over DMRS CO-PILOT (Accuracy: 18.01%→58.26%; macro-F1: 8.63%→24.62%). However, the Label 7 sink effect and resulting bimodality indicate that augmentation alone cannot overcome majority-class bias and clinical proximity. Future works should consider including constraint-based decoding, human-in-the-loop validation, and dialogue-level modeling to address temporal volatility identified in our analysis.

Limitations

Majority-class dominance and augmentation ceiling. The most critical limitation is the Label 7 sink effect identified in Section 3.2. Despite augmenting minority classes to $N = 500$, the model’s decision boundary remains heavily biased toward Label 7 (243/472 development instances), and macro-F1 performance on six of eight classes remains below 0.30. This suggests that naive count-balancing is insufficient when the majority class also exhibits high linguistic surface overlap with adjacent classes. Addressing this will require loss re-weighting strategies (e.g., focal loss (Lin et al., 2017)), hard-negative mining during augmentation, or explicit contrastive learning objectives that sharpen inter-class boundaries rather than simply expanding minority class size.

Turn-level modeling and temporal blindness. Our proposed architecture operates on isolated seeker utterances, each formatted with only the preceding stressor context. However, our dataset analysis (Appendix A.1) demonstrates that defense levels are unstable, frequently undergoing abrupt transitions across turns, and that larger defensive shifts tend to occur rapidly (Figure 5). A turn-level classifier observes only the outcome of a defensive transition, not the trajectory that produced it. This architectural limitation is especially problematic for clinically adjacent classes (e.g., Labels 6 and 7), whose distinction may reside in discourse-level patterns spanning multiple turns rather than in any single utterance.

Synthetic data validity and clinical reliability. Our quality control pipeline filters for inter-

annotator agreement ($\kappa \geq 0.60$) using a secondary classifier, providing a practical proxy for label consistency. However, this process does not guarantee clinical validity. A generated utterance may receive consistent classification by both the primary LLM and the secondary classifier while still failing to instantiate the functional psychological role of the target defense. Without human expert validation of a representative sample of synthetic instances, we cannot bound the rate of theoretically invalid but classifier-plausible samples in our training data. This is a fundamental limitation of machine-as-annotator pipelines in psychologically grounded domains, and future work should incorporate systematic clinician review.

Label underspecification and task formulation ambiguity.

Labels 0 (No Defense) and 8 (Needs More Information) do not correspond to clinical defense mechanisms and thus lack the DMRS indicator structure used to construct our Defense Profile features. For Label 8, the classification signal must be derived from the textual encoder and heuristic features, with the DMRS branch contributing noise rather than discriminative signal. Meanwhile, Label 0 is defined by the absence of a positive mechanism, making it harder to synthesize and harder for the NLI-based indicator scoring to characterize. A principled resolution, treating Labels 0 and 8 as a prior detection stage (defensive and non-defensive and ambiguous) before running the eight-class classifier, is deferred to future work.

Ethical considerations

This work relies on publicly released datasets (PSYDEFCONV and ESCONV) collected under informed consent and Institutional Review Board (IRB) oversight, with no new human data collection. Synthetic utterances simulating psychological distress are used exclusively for model training; clinical validity is not guaranteed, and expert review is required before any real-world deployment. The system is a research prototype and must not be used as a diagnostic tool.

Acknowledgments

We would like to thank the organizers of the BioNLP 2026 PsyDefDetect shared task and acknowledge the PSYDEFCONV dataset as the foundation of this work.

References

- Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. [Do not have enough data? Deep learning to the rescue!](#) *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7383–7390.
- Michael Bond, Susan Gardner, J. Christian, and J. Sigal. 1983. An empirical study of self-rated defense style. *Archives of General Psychiatry*, 40:333–338.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS ’20, Red Hook, NY, USA. Curran Associates Inc.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20:37 – 46.
- Phebe Cramer. 1987. [The development of defense mechanisms](#). *Journal of Personality*, 55(4):597–614.
- Mariagrazia Di Giuseppe and J. Christopher Perry. 2021. [The hierarchy of defense mechanisms: Assessing defensive functioning with the defense mechanisms rating scales Q-sort](#). *Frontiers in Psychology*, 12:718440.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. [MentalBERT: Publicly available pretrained language models for mental healthcare](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12):1–38.
- Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2020. [Supervised multimodal bitransformers for classifying images and text](#). *Preprint*, arXiv:1909.02950.
- Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. [Data augmentation using pre-trained transformer models](#). In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. [Focal loss for dense object detection](#). In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). In *Proceedings of ACL-IJCNLP 2021*, pages 3469–3483. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.
- George E. Vaillant. 1994. Ego mechanisms of defense and personality psychopathology. *Journal of Abnormal Psychology*, 103(1):44–50.
- Jason Wei and Kai Zou. 2019. [EDA: Easy data augmentation techniques for boosting performance on text classification tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. [Txygen: A benchmarking platform for text generation models](#). In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR ’18*, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Dataset Analysis

We conduct an exploratory analysis of PSYDEF-CONV to characterize its structural and temporal properties. Class imbalance is also reported in (Na et al., 2026b).

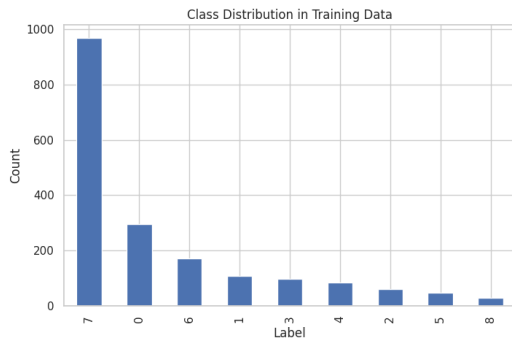


Figure 3: Class distribution across defense levels in the development set. Level 7 (High-Adaptive) dominates with 968 instances; Level 8 (Needs Info) contains only 28, motivating soft-balancing augmentation.

Temporal Volatility of Defense States. Defense levels are not stable within a dialogue (Fig. 4). Frequent transitions across levels, including abrupt shifts between adaptive and disavowal patterns within a single session, indicate that classification cannot rely on static turn-level features alone and must account for broader discourse context.

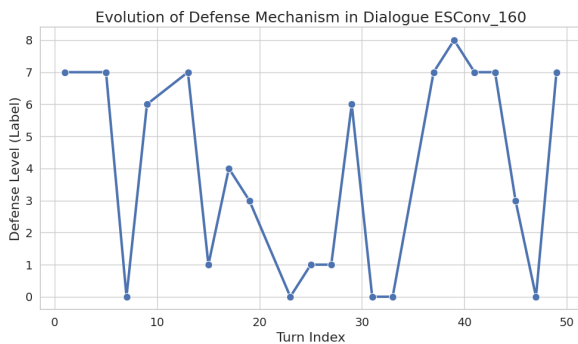


Figure 4: Defense level trajectory across turns in dialogue ESConv_016. Frequent transitions, including abrupt shifts between Level 7 and Level 0, indicate that defense states are temporally unstable and cannot be modeled from isolated turns.

Defense Volatility: Magnitude and Speed of Change. The scatter plot (Fig. 5) shows a positive correlation between the magnitude and speed of defense-level transitions: larger shifts in defensive functioning tend to occur over fewer turns.

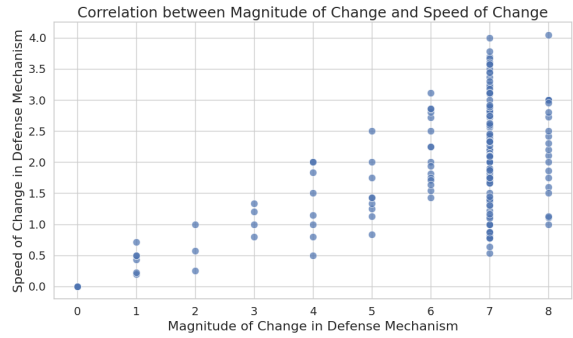


Figure 5: Correlation between magnitude and speed of defense level change across dialogue turns. Larger shifts in defense level tend to occur more rapidly, suggesting that defensive transitions are abrupt rather than gradual, a property that static turn-level classifiers are structurally unable to capture.

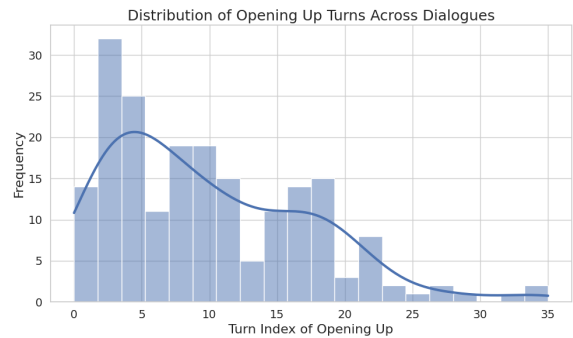


Figure 6: Distribution of turns at which seekers exhibit increased openness. The modal opening-up turn is 3–4, indicating early disclosure before defensive consolidation.

This is a key empirical finding, it implies that when a seeker’s defense changes, it changes quickly and dramatically, rather than gradually. This property motivates dialogue-level or sequential modeling as a future direction, as turn-level classifiers observe only the outcome of a transition, not its dynamics.

Disclosure Dynamics. Analysis of the CDI (Fig. 7) reveals a consistent disclosure peak around the 10-20% mark, followed by gradual stabilization. The opening-up distribution (Fig. 6) confirms that seekers tend to disclose early (modal turn \approx 3–4), suggesting that defensive activation intensifies *after* initial vulnerability rather than preceding it.

Response Latency as a Defensive Signal. Seeker response time varies across defense levels (Fig. 8). Label 0 (No Defense) shows higher latency variance, while defense-related responses cluster more tightly. This supports the use of temporal features as auxiliary signals.

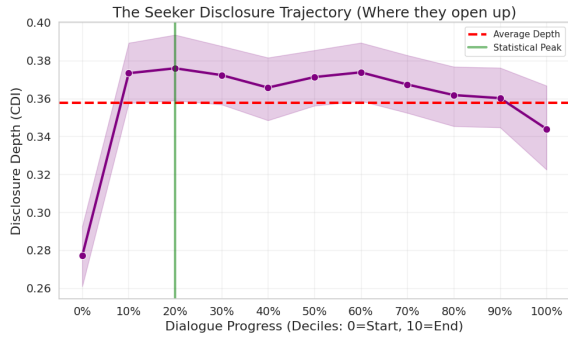


Figure 7: Composite Disclosure Index (CDI) across normalized dialogue progression. Disclosure peaks at the 10–20% mark then stabilizes, suggesting defensive activation intensifies after initial vulnerability.

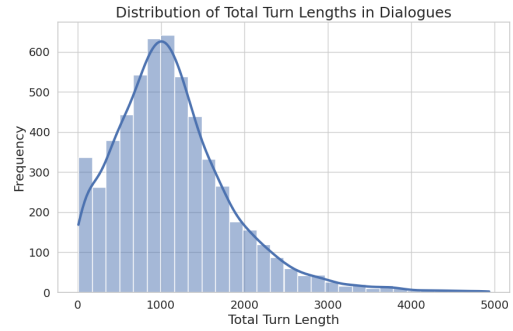


Figure 10: Distribution of total turn lengths per dialogue. The right-skewed distribution peaks around 1,000 tokens, with a long tail of extended sessions up to 5,000 tokens.

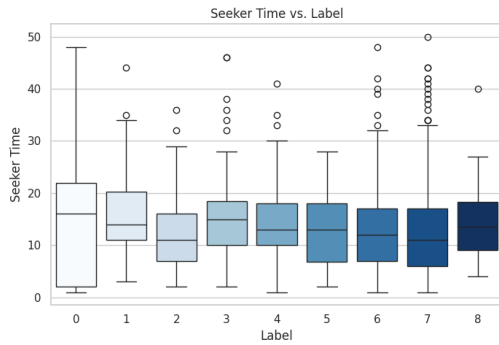


Figure 8: Seeker response time per defense label. Label 0 (No Defense) shows notably higher latency variance; defense-related labels cluster in tighter distributions, supporting temporal features as auxiliary classification signals.

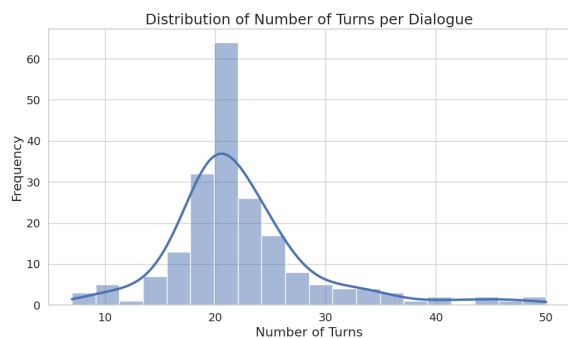


Figure 9: Distribution of number of turns per dialogue. Dialogues average approximately 20 turns, confirming the multi-turn nature of the classification task.

Corpus Structure. Dialogues average 20 turns and 1,000 tokens in total turn length (Figs. 9 and 10), confirming the multi-turn nature of the task and the need for context-aware modeling beyond single utterances.

A.2 Extended Result Analysis

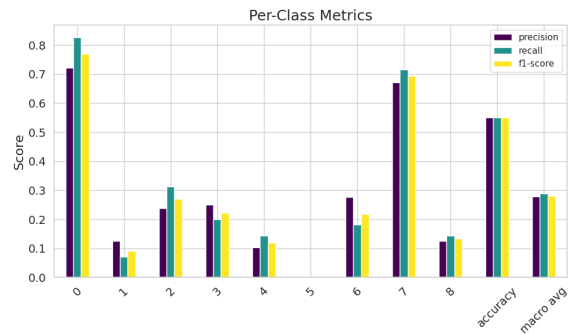


Figure 11: Per-class Precision, Recall, and F1 under the best setting ($\times 2$). Labels 1, 4, 5, and 8 remain below $F1 = 0.15$; Label 5 (Neurotic) achieves zero precision and recall, consistent with its severe underrepresentation (13 dev instances).

Per-Class Metrics. Labels 1, 4, 5, and 8 each yield F1 below 0.15. Label 5 (Neurotic) is never predicted, consistent with only 13 dev instances and chronic underrepresentation across all augmentation scales.

DMRS Mechanism Activation Patterns. Activation values are uniformly low (range: -0.8 to -1.7), reflecting the implicit nature of defensive language. Despite this, differential patterns across classes, particularly on *Autistic Fantasy*, *Undoing*, and *Affiliation*, confirm that the Defense Profile carries discriminative signal that complements the contextual encoder in the fusion architecture.

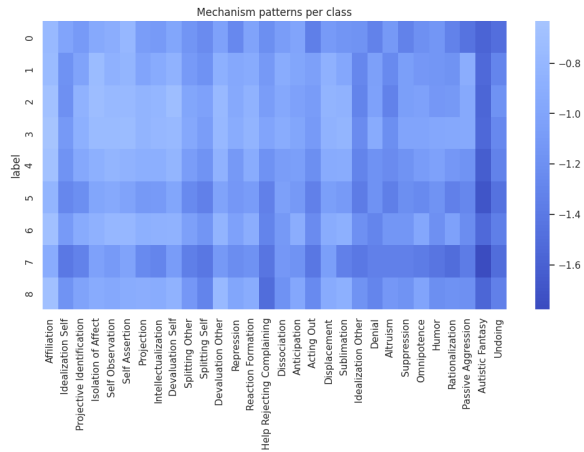


Figure 12: Mean NLI-inferred DMRS mechanism activation per defense class (log-entailment scores). All values are negative due to log-probability scaling. Differential gradients on *Autistic Fantasy*, *Undoing*, and *Affiliation* provide discriminative signal for the hybrid fusion model despite uniformly low absolute scores.

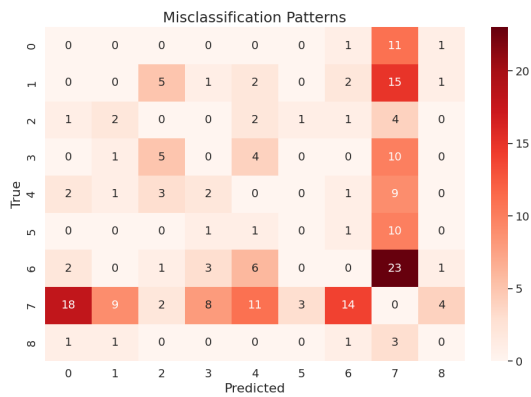


Figure 13: Off-diagonal misclassification counts (best setting, $\times 2$). Label 7 acts as a prediction sink across all classes. The Label 6 \rightarrow 7 confusion (23 errors) is the largest single off-diagonal cell, reflecting clinical proximity between obsessional and high-adaptive defenses.

Misclassification Patterns. The off-diagonal error analysis (Fig. 13) reveals that Label 7 is the dominant prediction sink: it absorbs the largest share of errors from every other class, accounting for 11, 15, 10, 9, 10, 23, and 3 misclassified samples from Labels 0-8 respectively. This is not random confusion but a systematic bias toward the majority class. Notably, Label 6 (Obsessional) is misclassified as Label 7 in 23 of 44 cases (52%), suggesting high clinical proximity between obsessional and high-adaptive functioning, a distinction that requires deeper contextual modeling to resolve.

A.3 Prompt Template for Synthetic Data Generation

We employ a theory-driven prompting strategy to generate synthetic utterances conditioned on stressors, dialogue history, and clinical defense mechanisms. The template used for generation is shown below:

Llama3 Prompt

```
prompt = f"""
### TASK: Generate Synthetic Psychological Defense Examples
You are simulating a seeker in a mental health support chat.

### CONTEXTUAL GROUNDING:
STRESSOR: {stressor}
DIALOGUE HISTORY:
{history}

### DEFENSE TO SIMULATE:
Mechanism: {mechanism_name} (Level {level})
Definition: {definition}
Pattern: {pattern_description}

### REFERENCE STYLE (Few-Shot):
1. "{example_1}"
2. "{example_2}"
3. "{example_3}"

### GOAL:
Generate 5 NEW seeker utterances for the NEXT TURN using the {mechanism_name} defense.
Ensure they follow the history and react to the stressor.

### OUTPUT FORMAT:
1 string.
No explanation, no markdown, no code fences.
"""
```

A.4 Prompt Template for Stressor Identification

Llama3 Prompt

```
prompt = f"""
### TASK: Clinical Stressor Identification
Identify the "Salient Stressor" causing psychological conflict in the Target Utterance.

### DIALOGUE CONTEXT:
{history}

### TARGET UTTERANCE:
"{target_turn}"

### OUTPUT FORMAT:
1. Stressor Category: (e.g., Interpersonal Conflict, Self-Esteem Threat, External Crisis)
2. Description: (One sentence explaining the threat)
"""
```

Diverse Transformer Ensemble with Majority Voting for Medical Decision Extraction at MedExACT 2026

Rishik Kondadadi

konda052@umn.edu

Abstract

This paper describes our system for the MedExACT 2026 shared task on extracting and classifying medical decisions from ICU discharge summaries. We frame the task as BIO token classification and train 25 diverse transformer models spanning 13 distinct architectures, including Longformer, DeBERTa, RoBERTa, BioBERT, SciBERT, and others. Each model is trained with category-aware oversampling, focal loss, and demographic-group-aware sampling to address class imbalance and promote fairness across patient subgroups. At inference time, we aggregate predictions via text-normalized majority voting, retaining spans agreed upon by at least 6 of 25 models. Our best submission achieves a final score of 0.5554 on the test set, demonstrating that a simple vote-based ensemble over architecturally diverse models outperforms more complex filtering approaches. We find that architectural diversity is a key driver of ensemble quality and that cross-validation is essential for reliable model selection on small clinical datasets.

1 Introduction

Clinical discharge summaries from intensive care units (ICUs) contain critical medical decisions that guide patient care transitions. Automatically extracting and categorizing these decisions can support clinical decision support, quality assurance, and retrospective analysis. The MedExACT 2026 shared task (Elgaar et al., 2026) challenges systems to identify decision spans in MIMIC-III (Johnson et al., 2016) discharge summaries and classify them into nine categories from the DICTUM taxonomy (Elgaar et al., 2024): (1) Contact related, (2) Gathering information, (3) Defining problem, (4) Treatment goal, (5) Drug, (6) Therapeutic procedure, (7) Evaluating test result, (8) Deferment, and (9) Advice and precaution.

A distinguishing feature of this shared task is its fairness-aware evaluation. The final score is

computed as the average of a base score (mean of span-level and token-level F1) and the Worst Group Score across nine demographic subgroups defined by sex, race, and language proficiency. This formulation penalizes systems that perform well on average but poorly on underrepresented populations.

We approach the task using a diverse ensemble of 25 transformer-based token classifiers. Our key findings are: (1) simple majority voting over diverse architectures outperforms learned ensemble methods, (2) architectural diversity matters more than model count within a single architecture family, and (3) cross-validation provides reliable model selection signals on small clinical datasets, while held-out validation scores are prone to overfitting.

2 System Description

2.1 Task Formulation

We formulate the task as BIO sequence labeling with 19 labels: O (outside), B- k (beginning of category k), and I- k (inside of category k) for $k \in \{1, \dots, 9\}$. Special tokens and padding positions are assigned the ignore index -100 during training.

2.2 Model Architectures

Our ensemble comprises 25 models spanning 13 distinct pretrained architectures, chosen to maximize prediction diversity:

- **Longformer variants (11 models):** We use Longformer-base (Beltagy et al., 2020) and clinical Longformer variants trained with different hyperparameter configurations (learning rates from 2×10^{-5} to 5×10^{-5} , training durations from 5 to 50 epochs, random seeds 42 and 7). Longformer’s sliding-window attention natively handles the long discharge summaries in the dataset, where approximately 32% exceed 4096 tokens.

- **DeBERTa-v3 (3 models):** We train DeBERTa-v3-base and DeBERTa-v3-large (He et al., 2023) with differential learning rates (encoder: 2×10^{-5} , classifier head: 1×10^{-4}) and sliding-window inference (window size 512, stride 128).
- **Domain-specific models (6 models):** BioBERT (Lee et al., 2020), Bio_ClinicalBERT (Alsentzer et al., 2019), BiomedBERT-large (Gu et al., 2022), SciBERT (Beltagy et al., 2019), BlueBERT (Peng et al., 2019), and PubMedBERT (Gu et al., 2022). These models are pretrained on biomedical or clinical corpora and bring domain-specific vocabulary knowledge.
- **General-purpose models (5 models):** RoBERTa-large (Liu et al., 2019), ELECTRA-large (Clark et al., 2020), ALBERT-xxlarge (Lan et al., 2020), XLNet-large (Yang et al., 2019), and BERT-large (Devlin et al., 2019). These models provide complementary representations learned from general-domain text.

2.3 Training Procedure

All models are fine-tuned as token classifiers using the HuggingFace Transformers library (Wolf et al., 2020). We employ several techniques to handle the challenges of the dataset:

Sliding-Window Chunking. For models with limited context windows (512 tokens), we chunk documents using a sliding window with a stride of 128 tokens. During training, overlapping windows share labels derived from the same gold annotations. During inference, predictions from overlapping windows are merged by retaining the highest-confidence prediction for each token position. Longformer models use a window size of 4096 with stride 1024.

Focal Loss. We use focal loss (Lin et al., 2017) with $\gamma = 2.0$ to address the severe class imbalance in the dataset, where the O label dominates:

$$\mathcal{L}_{\text{focal}} = -(1 - p_t)^\gamma \log(p_t) \quad (1)$$

where p_t is the predicted probability of the true class. Additionally, per-class weights computed via inverse square-root frequency provide further emphasis on rare BIO labels.

Category-Aware Oversampling. Categories 2 (Gathering information), 4 (Treatment goal), and 8 (Deferment) are underrepresented in the training data. We assign training chunks containing these categories a sampling weight of $10\times$ through weighted random sampling, ensuring the model sees these rare categories more frequently during training.

Group-Aware Sampling. To promote fairness across demographic subgroups, we compute per-sample weights inversely proportional to the frequency of each document’s demographic group, defined by the combination of sex, race, and language. This ensures underrepresented demographic groups receive proportionally more training signal.

Hyperparameters. We train with AdamW optimization, a warmup ratio of 0.1, weight decay of 0.01, and learning rates in the range $[2 \times 10^{-5}, 5 \times 10^{-5}]$. Models are trained for 5–50 epochs depending on convergence behavior, with gradient accumulation to achieve effective batch sizes of 16–32. The best checkpoint for each model is selected based on the official shared-task final score computed on the validation set at each epoch.

2.4 Inference

At inference time, each model processes a document using overlapping sliding windows as described above. BIO label sequences are decoded into spans by grouping consecutive B and I tags of the same category. Near-identical spans (same category, overlapping character offsets) are merged. Span boundaries are then refined by expanding to full word boundaries and trimming leading/trailing punctuation using the NLTK TreebankWordTokenizer, which matches the preprocessing applied in the official evaluation script.

2.5 Ensemble via Majority Voting

Given predictions from all 25 models for each document, we aggregate them using text-normalized majority voting:

1. For each model’s predictions on a document, we refine span boundaries and extract the low-ercased span text.
2. Predictions across all models are grouped by the key (category, normalized text).
3. A span is retained in the final output only if it receives votes from at least τ models. We set

$\tau = 6$ (approximately 24% of models).

- For each retained span, the character offsets are taken from the contributing model with the highest confidence score.

This approach is deliberately simple. We experimented with more sophisticated aggregation methods—per-category voting thresholds, confidence-weighted voting, IoU-based span matching, and a gradient-boosted machine (GBM) filter trained to classify ensemble candidates as true or false positives—but found that uniform majority voting consistently produced the best and most robust results on cross-validation.

3 Experimental Setup

3.1 Data

The MedExACT dataset (Elgaar et al., 2024) consists of discharge summaries from the MIMIC-III database with over 56,000 expert-labeled decision spans across nine DICTUM categories. We use the provided data splits, with approximately 200 documents for training and 53 for validation. Each document is linked to patient demographic metadata (sex, race, language proficiency) used for the fairness-aware evaluation.

3.2 Evaluation Metrics

Systems are evaluated on a composite final score:

$$\text{Final} = \frac{\text{Base} + \text{Worst Group Score}}{2} \quad (2)$$

where $\text{Base} = (\text{Span_F1} + \text{Token_F1})/2$ and the Worst Group Score is the minimum base score across nine demographic subgroups (Female, Male, White, African American, Hispanic, Asian, Other, English, Non-English).

Span F1 matches predictions to gold annotations by comparing tuples of (category, refined lowercased text), where span boundaries are preprocessed by expanding to word boundaries and trimming punctuation. Token F1 assigns token-level labels based on span overlap, computes macro-averaged F1 per document across the nine categories, and then averages across all documents.

4 Results

4.1 Test Set Results

We submitted three system configurations to the shared task. Table 1 summarizes the test set results.

System	Final Score
Organizer baseline (RoBERTa-base)	0.5111
Run 1: GBM filter on 25-model ensemble	0.4968
Run 2: Hybrid voting + LLM augmentation	0.4800
Run 3: 25-model vote ≥ 6	0.5554

Table 1: Test set final scores for our three submitted runs and the organizer-provided RoBERTa-base baseline (Elgaar et al., 2026). The baseline achieves Span F1 of 0.4363, Token F1 of 0.6238, Base of 0.5301, and Worst Group Score of 0.4922.

Our best submission (Run 3) exceeds the organizer-provided RoBERTa-base baseline (Elgaar et al., 2026) by 4.4 points absolute (0.5554 vs. 0.5111) and ranked 12th out of 37 submitted systems on the official leaderboard. The simple voting ensemble (Run 3) outperforms the GBM filter (Run 1) by 5.9 points absolute. The GBM filter was trained on the 53-document validation set to classify ensemble candidate spans as true or false positives using features such as vote count, confidence scores, and span length. Despite achieving a high validation score of approximately 0.75, the GBM overfits to the small validation set, resulting in substantially lower test performance. The hybrid approach (Run 2) supplemented the ensemble predictions on minority-demographic documents with spans extracted by a large language model (LLM) in a few-shot setting. This performed worst because the LLM extractions produced systematic boundary mismatches with the gold annotation conventions, yielding a validation score of only 0.277.

4.2 Subgroup Analysis

Our best submission (Run 3) attains its lowest base score on the Asian subgroup, which is notably absent from both the training and validation data. The vote-based ensemble handles unseen demographics reasonably well because it contains no learned parameters that could encode demographic bias—it simply counts agreement among independently trained models on the span text.

4.3 Ablation: Ensemble Size and Diversity

We compare our 25-model diverse ensemble against a 16-model ensemble restricted to Longformer and DeBERTa architectures only. On 5-fold cross-validation over the training set, the 25-model ensemble achieves a final score of approximately 0.50 versus 0.47 for the 16-model ensemble, confirming that architectural diversity—not simply model count—drives ensemble quality. Models

pretrained on different corpora and using different architectures (e.g., BioBERT vs. ELECTRA vs. XLNet) produce complementary errors, improving the signal-to-noise ratio of majority voting.

4.4 Cross-Validation vs. Validation Overfit

A critical finding from our experiments is the importance of cross-validation for model selection on this dataset. The 53-document validation set is too small for reliable evaluation: the GBM filter achieved 0.75 on validation but only 0.50 on test. In contrast, 5-fold cross-validation on the training set predicted final scores of approximately 0.49–0.50, which closely matched the actual test performance of 0.50–0.56. We used cross-validation scores for all ensemble configuration decisions, including the voting threshold τ and model selection.

5 Analysis and Discussion

Why simple voting wins. Learned ensemble methods (GBM filters, per-category thresholds) have sufficient capacity to memorize patterns in the small validation set. Majority voting has a single hyperparameter (τ), which limits its ability to overfit. The text-normalization step (lowercasing, boundary refinement) is crucial: it allows models with different tokenizers to agree on the same span despite minor offset differences.

Architectural diversity. Different pretrained architectures encode different inductive biases. Longformer captures long-range dependencies across the full document; DeBERTa uses disentangled attention representations; domain-specific models like BioBERT and ClinicalBERT bring biomedical vocabulary knowledge. Their disagreements are informative: spans confirmed across diverse architectures are more likely to be correct than spans supported by multiple variants of the same architecture.

Fairness through simplicity. Our group-aware sampling during training and the parameter-free nature of the voting ensemble contribute to relatively uniform performance across demographic subgroups. Since the ensemble aggregation does not use any document-level or demographic features, it cannot learn to behave differently for different patient populations.

LLM extraction limitations. Our attempt to use a large language model for few-shot span extraction scored poorly due to systematic boundary mis-

matches with the gold annotations. The LLM tended to paraphrase or extend spans beyond what the annotation guidelines specify, producing text that did not match after boundary refinement. This suggests that LLM-based extraction requires careful calibration to match dataset-specific annotation conventions.

6 Conclusion

We presented a 25-model diverse transformer ensemble for the MedExACT 2026 shared task on medical decision extraction. Our system aggregates predictions from 13 distinct pretrained architectures using text-normalized majority voting, achieving a final score of 0.5554. The key takeaways are: (1) architectural diversity in the ensemble matters more than model count or sophisticated aggregation methods, (2) majority voting provides natural regularization against overfitting on small clinical datasets, and (3) cross-validation is essential for reliable model selection when the validation set is small.

Limitations

Our system has several limitations. First, the voting threshold $\tau = 6$ was tuned via cross-validation on approximately 200 training documents and may not generalize to other clinical text domains or dataset sizes. Second, our approach requires training and running inference with 25 separate models, which is computationally expensive. Third, performance on demographic subgroups absent from training (e.g., Asian) remains lower, suggesting that additional techniques may be needed for true zero-shot fairness. Finally, our BIO formulation cannot represent overlapping or nested decision spans, which may occur in clinical text.

Ethics Statement

This work uses the MIMIC-III database (Johnson et al., 2016), which contains de-identified electronic health records. All data access was conducted under an approved PhysioNet credentialing agreement. Our system processes only de-identified text and does not attempt to re-identify patients. The fairness-aware evaluation framework of MedExACT highlights an important direction for equitable clinical NLP systems.

References

- Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78. Association for Computational Linguistics.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3615–3620. Association for Computational Linguistics.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024. **MedDec: A dataset for extracting medical decisions from discharge summaries**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16442–16455, Bangkok, Thailand. Association for Computational Linguistics.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Mehrnaz Sadrolashrafi, Mitra Mohtarami, Adrian Wong, Hadi Amiri, and Leo A. Celi. 2026. **Overview of medical decision extraction, analysis, and classification task (MedExACT) 2026**. In *The 25th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, San Diego, California, USA. Association for Computational Linguistics.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*.
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32.

FBK-NLP at ClinSkill QA 2026: Improving Temporal Reasoning via Keypoint-Augmented Inputs

Pedro Gabriel Campana^{1,2}, Alberto Lavelli², Bernardo Magnini²,

¹University of Padova / Via Trieste 63, Padova, Italy

²Fondazione Bruno Kessler / Via Sommarive 18, Povo, Trento, Italy

Correspondence: pedro.campana@phd.unipd.it

Abstract

Understanding procedural skills from visual data is a key challenge in medical AI, especially for tasks that require reasoning over temporal sequences. We report on FBK-NLP’s participation at the ClinSkill QA 2026 shared task, which requires models to arrange shuffled key frames into a coherent sequence of clinical actions and provide explanations for the resulting order. We conduct a systematic study of prompting and reasoning strategies using an open and easily deployable vision-language model (VLM). The central finding of our study is that incorporating keypoint-based representations of people’s body parts substantially improves temporal reasoning behind frame ordering. Furthermore, we show that model performance is highly sensitive to prompt design and to seemingly minor factors such as filename ordering and the inclusion of domain information.

1 Introduction

Recent advances in VLMs show strong capabilities in multimodal reasoning. In the medical domain, such models could meaningfully support clinical training and assessment by helping experts interpret procedural videos and verify compliance with standardized workflows. However, their behavior on structured temporal tasks remains poorly understood, as these demand not only visual recognition but also structured reasoning about how actions unfold over time. The ClinSkill QA 2026 shared task directly targets this challenge by requiring models to infer the correct chronological order of keyframes from clinical procedures.

We investigate how an open VLM can be instructed to perform temporal reasoning through systematic prompt design and input augmentation. Our approach is fully training-free, does not rely on

any task-specific fine-tuning, validation, or adaptation, and reveals non-obvious factors that significantly impact performance. Our main contribution is the use of keypoint-derived predicates as a complementary source of information. They encode body dynamics (e.g., wrist positions) and provide the model with explicit cues about motion and progression. To the best of our knowledge, this is the first work to explore keypoint information in the context of keyframe ordering. We demonstrate that this approach leads to significant performance gains — a non-obvious result, as coordinate-based representations could be expected to be out-of-distribution for VLMs.

In addition, we show that seemingly minor factors — such as the alphabetical ordering of image filenames — can introduce strong biases that degrade model predictions. First, we find that increasing reasoning complexity in prompts does not necessarily improve performance and can, in fact, make models more prone to such biases. Second, we demonstrate that incorporating domain-specific few-shot examples improves robustness and ordering accuracy. Third, we explore the role of domain knowledge by rerunning our best experiments without any domain-specific information in either the prompt or the few-shot examples to explore the task’s generalizability.

2 Related Work

Recent work has highlighted the difficulty that VLMs face when reasoning over ordered sequences of actions. In the textual domain, Wang et al. (2023) and Anika and Miah (2025) introduce benchmarks for step-wise and global procedural reasoning, showing how Large Language Models (LLMs) struggle to maintain correct orderings, particularly as sequence length and disorder increase. In multimodal settings, Song et al. (2025) show that models often fail to build persistent temporal representa-

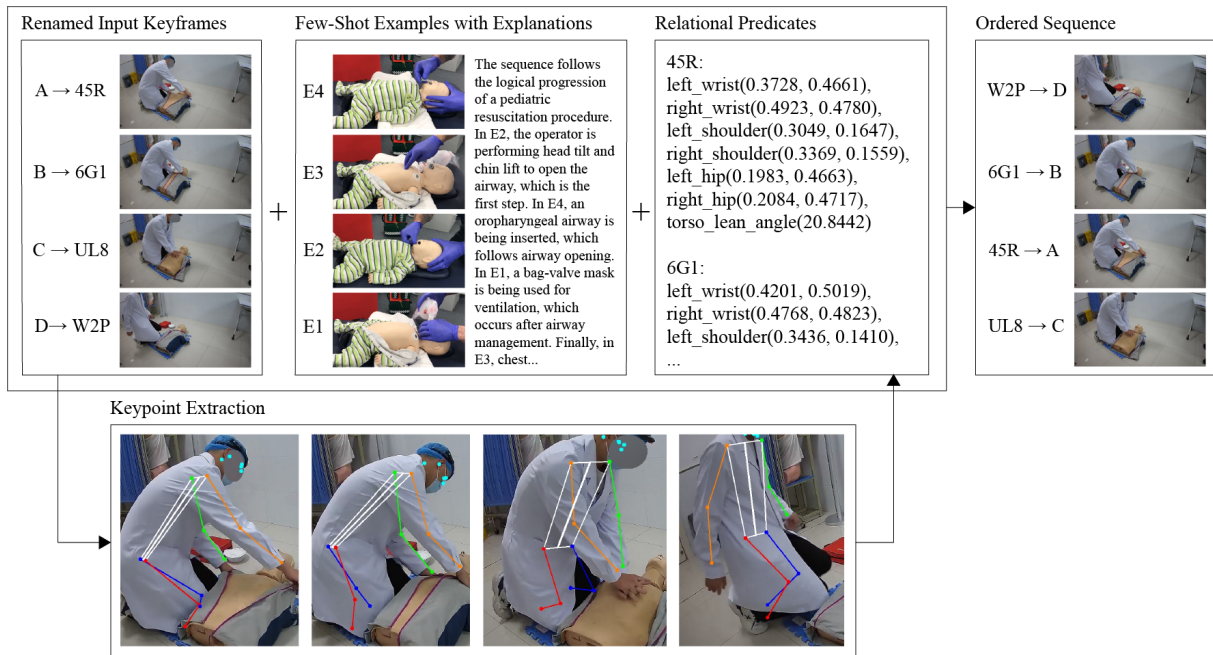


Figure 1: Overview of our most complete pipeline for clinical keyframe ordering. Unordered input frames are renamed to mitigate alphabetical bias and augmented with few-shot demonstrations. Human pose keypoints are extracted using an off-the-shelf model and converted into structured relational predicates, which are also incorporated into the prompt. The vision-language model then predicts the temporal ordering along with a textual explanation.

tions of visual sequences and instead rely on superficial cues, with performance dropping when visual information is removed.

Ma et al. (2024) also use temporal ordering as a proxy for deeper understanding in robotic settings by estimating progress from shuffled keyframes. They emphasize how ordering constraints can induce reasoning about causality, progression, and state transitions. Other studies further suggest that temporal reasoning is not only a representation issue but also a modeling and reasoning challenge: Chen et al. (2023) find that temporal information is already present in visual encodings but is not fully exploited by language models, while Ko et al. (2023) show that LLMs can exhibit temporal and causal reasoning in video QA, although in a fragile and prompt-sensitive manner.

Taken together, these works highlight both the limitations and partial successes of modern models in temporal and procedural reasoning. They motivate our focus on keyframe ordering in clinical workflows and on understanding how prompting strategies and structured representations can improve performance in a training-free setting.

3 Problem Formulation

ClinSkill QA 2026 is a shared task designed to evaluate multimodal reasoning over clinical pro-

cedures, built on the dataset introduced by Huang et al. (2026). It consists of two tasks: (1) arranging shuffled keyframes into a coherent sequence of clinical actions, and (2) providing explanations for the resulting order. Only a test set with 200 samples is provided. Each sample contains four to six keyframes depicting different stages of a basic life support (BLS) procedure. The ground-truth labels are not given to ensure fairness.

Evaluation considers both the correctness of the predicted sequence and the quality of the respective explanations. Ordering performance is measured using task accuracy, which requires exact sequence matching, and pairwise accuracy, which captures partial correctness by evaluating adjacent frame relationships. In addition, explanation quality is assessed using BERTScore and an LLM-based evaluator (G-Eval). These metrics are combined into an overall score used for leaderboard ranking. We use it as our primary metric, since the individual metrics are not provided for all submissions.

The task presents several challenges: (i) it requires relational reasoning, since the correct position of each frame depends on comparisons with the others rather than on independent classification; (ii) many temporal cues are implicit, forcing the model to infer ordering from subtle changes in posture, object configuration, or interaction dy-

namics; and (iii) the presence of textual context introduces opportunities for spurious correlations, where the model may rely on superficial signals (e.g., filename structure or prompt wording) instead of visual evidence.

4 Methodology

Our methodology examines how prompting strategies, input representations, and auxiliary cues influence zero-shot and few-shot performance in temporal ordering. Our pipeline is shown in [Figure 1](#).

4.1 Prompting

We explore different prompting strategies, including simple formulations that request the final ordering directly and more structured prompts that require intermediate reasoning steps. They correspond to the following baselines:

Order–then–explanation. We ask the model to produce the final ordering before the explanation.

Explanation–then–order. We first ask the model for the explanations and then the ordering of the key frames, as models may revise or improve their answers during the explanation process.

Intermediate reasoning. We increase the complexity of the prompt and explicitly request intermediate reasoning steps, including frame-wise descriptions and pairwise comparisons, before the final ordering and the explanations.

4.2 Input Representation

Beyond prompting, we explore how input representations affect model predictions. In particular, we discovered that filenames can introduce unintended biases, as models may exploit alphabetical ordering as a shortcut for sequence prediction. To mitigate this effect, we use randomly assigned filenames to ablate our three initial baselines, ensuring that filenames do not encode any implicit ordering signal. This allows us to better isolate the model’s reliance on visual and semantic information.

4.3 Few-Shot Examples

To provide additional task guidance, we then incorporate a few-shot demonstration into our prompts. Each demonstration consists of a set of input frames paired with correct orderings and corresponding explanations. We study both in-domain and out-of-domain demonstrations to evaluate how domain alignment affects performance, and to what

extent the model can generalize from unrelated procedural contexts. More information about these examples can be found in [Appendix A](#).

4.4 Keypoint-Derived Predicates

Our main contribution is the introduction of keypoint-derived predicates as auxiliary inputs for temporal reasoning. We extract human pose keypoints from each frame using the off-the-shelf YOLOv8 ([Jocher et al., 2023](#)) pose estimation model and convert them into structured textual descriptions. In particular, we retain the 2D coordinates of the left and right wrists, shoulders, and hips, and derive an additional torso-lean predicate from the angle between the shoulder-center and hip-center vectors (see [Appendix B](#)).

Unlike raw pixel data, these predicates offer a higher-level representation of action dynamics, encoding spatial relationships between body parts and providing explicit cues about posture and movement. The process is entirely training-free: the pose estimator is not adapted to the task, and the predicates are generated automatically without manual annotation. Furthermore, they are provided as complementary information rather than primary inputs. The model is explicitly instructed not to rely primarily on them, as keypoint extraction may be noisy or inaccurate.

5 Experiments

As a backbone VLM we use Qwen3-VL-32B-Instruct ([Team, 2025](#)) with its default hyperparameter configurations. Results can be seen in [Table 1](#).

5.1 Prompting Baselines and Filename Bias

Among the three prompting strategies, the simple order–then–explanation approach consistently outperformed the others. In addition, intermediate reasoning yielded considerably lower scores, suggesting that more complex and demanding instructions can introduce noise and overthinking rather than improve temporal inference.

As for the filenames, random representations led to consistent improvements (approximately 3 points), confirming that models are sensitive to superficial ordering cues such as alphabetical filenames. This is supported by the fact that the number of sequences ordered as the filename alphabetic pattern, such as ABCD or ABCDEF, decreased across all baselines (see [Appendix A](#)).

Baseline	Renaming	Few-Shot Examples	Predicates	Domain Information	Score
Order-then-explanation	No	No	No	Yes	46.51
Order-then-explanation	Yes	No	No	Yes	49.30
Explanation-then-order	No	No	No	Yes	41.84
Explanation-then-order	Yes	No	No	Yes	45.47
Intermediate reasoning	No	No	No	Yes	39.60
Intermediate reasoning	Yes	No	No	Yes	42.93
Order-then-explanation	Yes	No	No	Yes	49.30
Order-then-explanation	Yes	Yes	No	Yes	51.13
Order-then-explanation	Yes	Yes	Yes (Subset)	Yes	56.71
Order-then-explanation	Yes	Yes	Yes	Yes	58.79
Order-then-explanation	Yes	No	Yes	Yes	59.45
Order-then-explanation	Yes	No	No	No	44.41
Order-then-explanation	Yes	No	Yes	No	52.34
Order-then-explanation	Yes	No	Yes	Yes	59.75
Order-then-explanation	Yes	Yes	No	Yes	51.13
Order-then-explanation	Yes	Yes	No	No	44.20
Order-then-explanation	Yes	Yes	Yes	No	54.47

Table 1: Summary of all experiments analyzing the impact of prompting strategy, random file renaming, few-shot demonstrations, keypoint-derived predicates, and domain information. The table is divided into three blocks: (top) comparison of prompting strategies with and without renaming, highlighting the effect of alphabetical bias; (middle) incremental improvements from few-shot examples and predicates; and (bottom) analysis of domain information and generalization.

5.2 Examples and Predicates

Adding three in-domain few-shot examples to the prompt improved the best experiment we had from 49.30 to 51.13. The gain, however, was relatively modest, suggesting that although models can benefit from a small number of task-relevant demonstrations, these provided limited additional gains for temporal ordering.

The inclusion of keypoint-derived predicates, on the other hand, led to substantial improvements. With the full predicate set, we reached a score of 58.79. We reran the experiment with only wrists and torso predicates to test whether the others introduced noise or redundancy, but this variant reached a slightly lower score of 56.71. Interestingly, our best result was obtained with the full predicate set without any few-shot examples (59.45).

5.3 Domain Information

When removing domain information only from the prompt, the score remained unchanged at 51.13, suggesting that examples can preserve domain strengths. When both the prompt and the examples were out-of-domain, however, performance dropped to 44.20, confirming that the few-shot benefit is domain-sensitive. This drop was expected, but it was interesting to see how the keypoint predicates still increased the score to 54.47, i.e., they carry meaningful task structure even without domain-specific instructions.

5.4 Discussion

Our results highlight several important aspects of temporal reasoning in VLMs beyond raw performance. First, increasing the complexity of prompting does not necessarily improve the quality of reasoning. In fact, more elaborate instructions that encourage step-by-step analysis and pairwise comparisons underperform simpler formulations. Second, random file renaming consistently improves results, revealing a non-trivial alphabetical bias in the model’s behavior. Third, while in-domain examples help guide the model toward the desired output format and reasoning pattern, they are not sufficient to substantially improve performance on their own. Finally, explicit pose information provides strong guidance for temporal reasoning even when no domain information is given.

6 Conclusion

Our main contribution is the introduction of keypoint-derived predicates as structured auxiliary inputs. Extracted using an off-the-shelf pose estimator and integrated at inference time, these predicates provide explicit cues about body configuration and motion, leading to significant improvements in ordering performance.

7 Acknowledgments

This work has been partially funded by the European Union under the Horizon Europe eCREAM Project (Grant Agreement No.101057726).

References

- Adrita Anika and Md Messal Monem Miah. 2025. Evaluating LLMs’ reasoning over ordered procedural steps. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2259–2267.
- Junwen Chen, Jie Zhu, and Yu Kong. 2023. Atm: Action temporality modeling for video question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4886–4895.
- Xiyang Huang, Jiawei Lin, Keying Wu, Jiaxin Huang, Kailai Yang, Renxiong Wei, Jiayi Xiang, Ziyang Kuang, Min Peng, Qianqian Xie, and 1 others. 2026. Siming-bench: Evaluating procedural correctness from continuous interactions in clinical skill videos. *arXiv preprint arXiv:2604.09037*.
- Glenn Jocher, Jing Qiu, and Ayush Chaurasia. 2023. [Ultralytics yolo](#).
- Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. 2023. [Large language models are temporal and causal reasoners for video question answering](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4316, Singapore. Association for Computational Linguistics.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Yecheng Jason Ma, Joey Hejna, Ayzaan Wahid, Chuyuan Fu, Dhruv Shah, Jacky Liang, Zhuo Xu, Sean Kirmani, Peng Xu, Danny Driess, Ted Xiao, Jonathan Tompson, Osbert Bastani, Dinesh Jayaraman, Wenhao Yu, Tingnan Zhang, Dorsa Sadigh, and Fei Xia. 2024. [Vision language models are in-context value learners](#). *Preprint*, arXiv:2411.04549.
- Yingjin Song, Yupei Du, Denis Paperno, and Albert Gatt. 2025. Burn after reading: Do multimodal large language models truly capture order of events in image sequences? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24316–24342.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Weizhi Wang, Hong Wang, and Xifeng Yan. 2023. Steps: A benchmark for order reasoning in sequential tasks. *arXiv preprint arXiv:2306.04441*.

A Prompting and Input Design

This appendix provides additional details to support the reproducibility of our experiments. We report the most relevant prompts used in our study,

describe input preprocessing choices, and provide further information on examples and analyses.

A.1 Prompts

We report below the exact prompt used for the **order-then-explanation** baseline.

Without domain information:

You are analyzing a set of images depicting a sequence of actions.

You will receive several frames from the same sequence. The frames are presented in random order.

Your task is to determine the correct chronological order of the frames and then provide a short explanation describing the visual cues that determine the ordering.

Focus on cues such as changes in object positioning, visibility of elements, interaction with items, and overall progression of the scene.

In cases of ambiguity, assume that the sequence is moving forward in a natural progression, such as moving closer to a subject rather than away, or revealing more of a scene rather than concealing it.

When explaining your reasoning, include the reason for which a frame should precede or succeed the other.

You must output ONLY a valid JSON object in the following format and nothing else:

```
{
  "predicted_order": ["<filename of the first frame in the chronological order>", "<filename of the second frame in the chronological order>", ...],
  "order_rationale": "<brief explanation of the visual reasoning and progression>"
}
```

With domain information:

You are an expert clinician analyzing training images from a medical procedure.

You will receive several frames from the same procedure. The frames are presented in random order.

Your task is to determine the correct chronological order of the frames and then provide a short explanation describing the visual cues that determine the ordering.

Focus on cues such as operator actions, exposure of anatomical regions, hand placement, interaction with equipment, and progression of the procedure.

In cases of ambiguity, assume that the operator is progressing into the procedure, for example, moving toward the person or manikin rather than away and exposing the chest rather than covering it. Another example: a frame in which pads are already placed on the chest should occur later than frames in which they are not.

When explaining your reasoning, include the reason for which a frame should precede or succeed the other.

You must output ONLY a valid JSON object in the following format and nothing else:

```
{
  "predicted_order": ["<filename of the
    first frame in the chronological order>",
    "<filename of the second frame in the
    chronological order>", ...],
  "order_rationale": "<brief explanation
    of the visual reasoning and procedural
    progression>" }
```

A.2 Sequential Frame Presentation

We also experimented with an alternative prompting strategy in which frames were provided sequentially within the same conversation, rather than all at once. At each step, the model received a new frame and was expected to update its internal representation of the sequence.

This approach was motivated by the hypothesis that incremental exposure could facilitate temporal reasoning by accentuating pairwise comparisons. However, even with the keypoint-based predicates, it resulted in substantially worse performance (46.06) than all other prompting strategies. We hypothesize that this is due to the model’s limited ability to maintain and revise a coherent global representation of the sequence across messages. Unlike the standard setup, where all frames are available simultaneously, this one requires persistent memory and consistent updating, which current models appear to handle poorly.

A.3 Filename Bias

We adopted the following filename-renaming scheme:

```
A → 45R   D → W2P
B → 6G1   E → T93
C → UL8   F → 7WH
```

It removes alphabetical bias while avoiding the numerical bias that would arise from renaming files to IMG_1, IMG_2, and so on. Because each sample corresponds to a different conversation, we applied the same mapping across all samples, ensuring that no filename-based patterns or associations

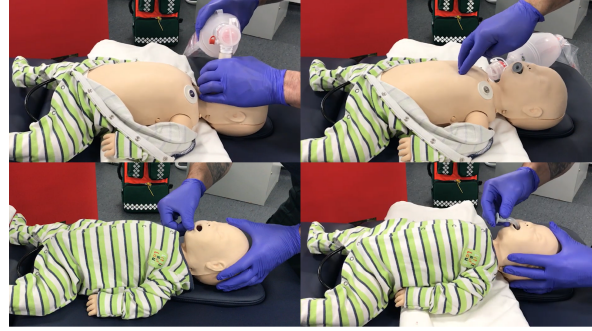


Figure 2: Example of an in-domain set of video frames collected.



Figure 3: Example of an out-of-domain set of video frames collected.

were possible. Under this scheme, the number of sequences whose ordering matched the original alphabetical pattern (ABCD or ABCDEF) decreased as follows:

- **Order-then-explanation:** 37 → 35
- **Explanation-then-order:** 58 → 32
- **Intermediate reasoning:** 76 → 50

This indicates that, even in a purely visual task, the model may still exploit irrelevant textual structure in the input representation.

A.4 Few-Shot Examples

In-domain examples were obtained from YouTube videos by querying "basic life support" and filtering for content available under a Creative Commons license. Out-of-domain examples were obtained using the same procedure but with other queries, such as "skateboarding", "baking", and "yoga". Some of these examples can be seen in Figures 2 and 3. We always provided three examples, each presented to the model along with its correct temporal ordering and a detailed explanation written by us.

B Keypoint-Derived Predicate Construction

This appendix describes and discusses in more detail the implementation of our predicate extraction pipeline.

B.1 Pose Estimation and Keypoint Selection

The pretrained YOLOv8 pose estimator predicts the 17 COCO (Lin et al., 2014) body keypoints for every detected person. We assume the operator is the person detected with higher confidence in the frame. The detected keypoints are then normalized by image width and height to obtain scale-invariant coordinates in the range $[0, 1]$. From the complete COCO skeleton, we retain the upper-body landmarks that were empirically most informative for procedural ordering:

- left/right shoulders (indices 5 and 6),
- left/right wrists (indices 9 and 10),
- left/right hips (indices 11 and 12).

These landmarks are converted into textual predicates of the form `left_wrist(x, y)`, where x and y denote normalized image-space coordinates.

B.2 Torso Orientation

In addition to raw landmark coordinates, we derive a coarse torso orientation descriptor from shoulder and hip positions. Let

$$s = \frac{p_{ls} + p_{rs}}{2}$$

be the midpoint between the left and right shoulders, and

$$h = \frac{p_{lh} + p_{rh}}{2}$$

the midpoint between the hips. We define the torso vector as $v = s - h$. The torso lean angle is then computed relative to the upward vertical direction $u = (0, -1)$:

$$\theta = \cos^{-1} \left(\frac{v \cdot u}{\|v\|} \right).$$

The resulting value is encoded as the predicate `torso_lean_angle(θ)`, providing an estimate of body posture and forward leaning, which can correlate with different stages of clinical procedures.

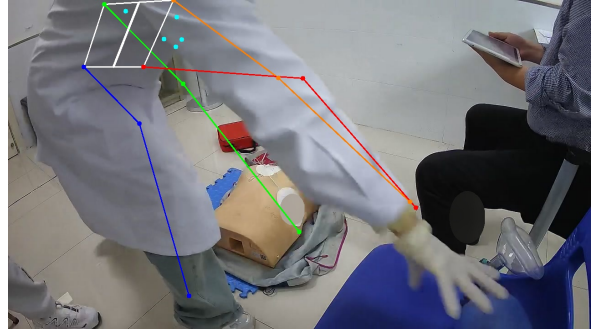


Figure 4: Example of an occluded frame where the operator is only partially captured in the camera view, resulting in an unreliable keypoint detection.



Figure 5: Example of a failure case where an observer is detected instead of the operator.

B.3 Error Analysis

As only overall submission-level scores were provided and no per-sample metrics were available, we restrict our error analysis to the keypoint-derived predicates. As expected, most errors were attributed to occlusions — both in frames where parts of the operator were deliberately masked and in others where the operator was only partially visible in the camera view. Figures 4 and 5 illustrate two representative failure cases. It would be interesting to explore whether an alternative detection strategy could mitigate them.

transformer_1376 at PsyDefDetect: A QLoRA-Based Generative Framework for Context-Aware Psychological Defense Mechanism Detection

Pritha Saha¹, Shuvodwip Saha¹, Anik Mahmud Shanto^{2,*}

¹ Chittagong University of Engineering & Technology, Chittagong-4349, Bangladesh

² Southeast University, Dhaka, Bangladesh

{priitha.saha, shuvodwipsaha}@gmail.com,

anik.mahmudshanto@seu.edu.bd

Correspondence: anik.mahmudshanto@seu.edu.bd

Abstract

Psychological defense mechanisms play a crucial role in shaping human responses during emotionally charged conversations, yet remain underexplored in natural language processing. In this work, we address the PSYDEFCONV shared task, which involves classifying defense mechanisms in multi-turn dialogues using clinically grounded annotations based on the Defense Mechanism Rating Scales (DMRS). We propose a generative supervised fine-tuning framework that reformulates the task as conditional text generation. A pre-trained causal language model (Gemma-2-2B) is adapted using parameter-efficient fine-tuning (PEFT) with 4-bit quantization, enabling efficient training under limited computational resources. To handle class imbalance, we apply random oversampling, and we design a prompt-based input representation to incorporate conversational context effectively. Experimental results demonstrate that our generative approach is competitive with discriminative baselines while offering improved flexibility in modeling subtle and context-dependent defensive behaviors. The findings highlight the potential of generative large language models for psychologically grounded dialogue understanding tasks.

1 Introduction

Psychological defense mechanisms are vital adaptive or maladaptive strategies that shape how individuals regulate emotions and express distress. In emotional support dialogues, identifying these mechanisms—such as denial or intellectualization—is essential for understanding a speaker’s mental state and providing appropriate support. However, defensive functioning remains largely underexplored in natural language processing (NLP) compared to surface-level tasks like sentiment analysis (Na et al., 2025).

The PsyDefDetect@BioNLP 2026 shared task (Na et al., 2026a) addresses this gap by

requiring the classification of target utterances into nine categories based on the Defense Mechanism Rating Scales (DMRS). This task is particularly challenging due to the subtle, context-dependent nature of defenses, the nuanced hierarchical label space, and the presence of ambiguous cases where context is insufficient for classification.

To address these challenges, we propose a generative framework that reformulates defense mechanism identification as a conditional text generation problem rather than standard multi-class classification. We utilize parameter-efficient fine-tuning (PEFT) on a pre-trained causal language model to adapt to the task with minimal computational overhead. Our approach leverages the reasoning capabilities of large language models to capture the long-range dependencies and implicit cues necessary for identifying nuanced psychological constructs.

Contributions. Our main contributions are as follows:

- We introduce a parameter-efficient adaptation strategy that enables effective training under limited computational resources.
- We demonstrate the effectiveness of our approach on the PsyDefDetect shared task, highlighting its ability to capture context-dependent and nuanced defensive behaviors.

2 Related Work

Psychological Analysis in NLP. Research in computational psychology has traditionally focused on emotion recognition and sentiment analysis (Porcia et al., 2017; Zhang et al., 2018). While datasets like ESConv (Liu et al., 2021) facilitate emotional support modeling, they often focus on surface-level affect. The PSYDEFCONV dataset addresses this by providing annotations based on the Defense Mechanism Rating Scales (DMRS)

(Perry, 1990), enabling a deeper analysis of cognitive processes in dialogue.

Context-Aware Dialogue Understanding. Effective dialogue modeling requires capturing dependencies across multiple turns. While early methods used recurrent architectures (Serban et al., 2016), modern transformer-based approaches leverage self-attention to encode context (Devlin et al., 2019; Wolf et al., 2019). However, identifying implicit psychological phenomena remains a challenge, as these constructs depend on subtle nuances often missed by standard encoding strategies.

Generative Approaches for Classification. While discriminative models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are standard for classification, recent work explores reformulating these tasks as text generation (Raffel et al., 2020; Brown et al., 2020). Conditional generation via instruction-based prompting has shown superiority in complex scenarios requiring reasoning and contextual interpretation, making it highly suitable for defense mechanism detection.

Parameter-Efficient Fine-Tuning. Fine-tuning large language models (LLMs) is computationally demanding. Parameter-efficient fine-tuning (PEFT) techniques, such as LoRA (Hu et al., 2021), allow for effective model adaptation by updating only a fraction of the total parameters. These methods significantly reduce resource requirements while maintaining performance levels comparable to full fine-tuning, providing a scalable solution for specialized NLP tasks.

3 Dataset

We conduct our experiments on the PSYDEFCONV dataset (Na et al., 2026b), released as part of the shared task on psychological defense understanding in conversations. The dataset is constructed from a subset of the ESConv corpus (Liu et al., 2021) and contains multi-turn emotional support dialogues annotated using the Defense Mechanism Rating Scales (DMRS) (Perry, 1990).

Statistic	Value
Total Dialogues	200
Total Utterances	4,709
Annotated Utterances	2,336
Number of Classes	9

Table 1: Overview of the PSYDEFCONV dataset.

The task is formulated as a 9-class classification problem, where each instance consists of a target utterance and its preceding dialogue context. The labels correspond to hierarchical levels of defensive functioning, ranging from no defense to high-adaptive defenses, along with an additional category for ambiguous cases requiring more context.

Following the shared task setup, we have used the provided train, validation, and test split. The task is particularly challenging due to its context dependency, subtle inter-class variations, and the presence of ambiguous instances.

4 Methodology

We propose a generative supervised fine-tuning (SFT) framework as shown in figure 1 to model psychological defense classification as a conditional text generation task. A pre-trained causal language model (Gemma-2-2B) is adapted using parameter-efficient fine-tuning (PEFT) to predict defense labels given conversational context.

4.1 Data Processing

To preserve class distribution, we perform a stratified 80/20 train-validation split. Due to class imbalance, we apply random oversampling to minority classes up to a fixed target size, while retaining all samples from majority classes. This strategy improves representation of rare defense categories without discarding useful data.

4.2 Model and Fine-Tuning

We adopt a 4-bit quantized QLoRA setup for efficient training. Low-rank adapters are injected into all linear projection layers, enabling task adaptation with minimal parameter updates. Quantization significantly reduces memory usage, allowing fine-tuning on limited hardware while maintaining performance.

4.3 Generative Formulation

Unlike conventional classifiers, we formulate the task as next-token prediction. Each input is constructed using a chat-style template containing the system prompt, dialogue context, and target utterance. The model is trained to generate a single token corresponding to the defense label.

To ensure proper supervision, tokens corresponding to the input prompt are masked, and loss is computed only on the generated label token. For long sequences, a left-truncation strategy is applied

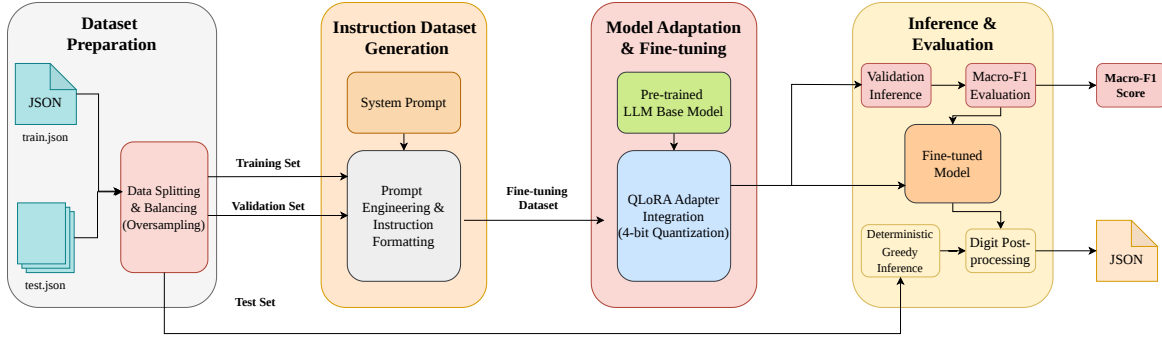


Figure 1: Our proposed framework for the PsyDefDetect task

to preserve the target utterance and label within the maximum sequence length. Since Gemma-2’s chat template does not support a dedicated system role, the system instruction is prepended directly into the user turn. The full prompt template is provided in Appendix B.

4.4 Training Strategy

We implement a manual training loop using PyTorch with gradient accumulation to achieve an effective batch size of 16. Mixed precision training is employed to improve efficiency. A linear warmup followed by cosine decay is used for stable optimization.

4.5 Inference and Evaluation

During inference, we use greedy decoding to generate the predicted label deterministically. A simple post-processing step extracts the label token from the generated output. Following the shared task protocol, we report Macro-F1 score over classes 1–8, excluding the majority “No Defense” class to ensure balanced evaluation.

5 Experimental Analysis

5.1 Experimental Setup

All experiments are conducted on cloud-based GPU platforms (Kaggle and Google Colab) using NVIDIA T4 GPUs. The hardware constraints motivate our use of 4-bit quantization and parameter-efficient fine-tuning throughout.

We fine-tune Gemma-2-2B-IT via 4-bit QLoRA using the Unsloth framework. LoRA adapters are injected into all attention and feed-forward projection layers. Training uses a manual PyTorch loop with AdamW optimization, cosine decay scheduling with linear warmup, and gradient clipping. To

address class imbalance, minority classes are over-sampled to a fixed target count prior to training.

To isolate the contribution of the generative reformulation, we additionally train DeBERTa-v3-base as a discriminative sequence classifier under identical data conditions. The dialogue context and target utterance are encoded as a sentence pair with left-truncation applied to the context side. This model serves as an internal comparison point and is distinct from the organizer-provided official baseline.

Both systems are trained and evaluated on the same stratified 80/20 split of the official training data, with final evaluation performed on the official blind test set. Full hyperparameters for both systems are listed in Appendix A.

5.2 Results

Our system achieves an official Macro-F1 of 0.2475 on the blind test set (LB1), ranking 12th out of 21 registered teams. Table 2 presents a full comparison across all evaluation metrics.

Table 2: Comparison of model performance metrics (Accuracy, Precision, Recall and F1-score) on the official test set.

Model	Acc.	Prec.	Rec.	F1
Official Baseline*	0.6483	0.3397	0.3045	0.3148
DeBERTa-v3-base	0.5500	0.1202	0.1636	0.1376
Proposed Framework	0.5508	0.2669	0.2351	0.2475

*Organizer-provided baseline from official leaderboard.

The proposed generative system substantially outperforms the discriminative DeBERTa-v3-base encoder across all metrics, with a Macro-F1 gain of +0.11, suggesting that reformulating defense classification as conditional text generation better captures the implicit and context-dependent nature of psychological defenses. The gap relative to the official baseline reflects the inherent difficulty of

the task under limited computational resources, and points toward larger generative models as a promising direction for future work.

6 Conclusion

In this work, we presented a generative supervised fine-tuning framework for psychological defense mechanism classification in conversational data. By reformulating the task as conditional text generation and leveraging parameter-efficient fine-tuning with 4-bit quantization, we successfully adapted a causal language model to capture context-dependent psychological cues.

Experimental results demonstrate that the proposed method is competitive with standard discriminative baselines while offering improved flexibility in modeling implicit and nuanced defense behaviors. Future work will explore larger-scale models, improved context modeling strategies, and more robust decoding mechanisms to further enhance performance on psychologically grounded conversational tasks.

7 Limitations

Our approach has several limitations. First, the model relies on generative decoding, which introduces slight variability in outputs and may require strict post-processing to ensure valid label prediction. Second, the use of a relatively small causal language model (Gemma-2-2B) limits the capacity to capture highly complex conversational nuances compared to larger models. Third, class imbalance handling via oversampling may introduce redundancy and potential overfitting to minority classes. Finally, the reliance on fixed prompt templates may reduce generalization to unseen dialogue structures.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of](#)

[deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#).

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.

J. Christopher Perry. 1990. *Defense Mechanism Rating Scales (DMRS)*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. [Context-dependent sentiment analysis in user-generated videos](#). In *Proceedings of the*

55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. [Building end-to-end dialogue systems using generative hierarchical neural network models](#). *Preprint*, arXiv:1507.04808.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. [Transfertransfo: A transfer learning approach for neural network based conversational agents](#). *Preprint*, arXiv:1901.08149.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. [Deep learning for sentiment analysis : A survey](#). *Preprint*, arXiv:1801.07883.

A Hyperparameters

Table 3 details the hyperparameters for reproducibility, including the LoRA adapters and the 4-bit quantization setup.

Table 3: Detailed training and model configurations for reproducibility.

Hyperparameter	DeBERTa-v3-base	Gemma-2-2B + QLoRA
Learning rate	2e-5	1e-4
Effective batch size	8	16
Epochs	5	3
Weight decay	0.01	0.01
Max sequence length	512	1024
Quantization	—	4-bit NF4
LoRA rank	—	16
LoRA alpha	—	32
LoRA dropout	—	0.05
Warmup ratio	—	0.05
Grad clip norm	—	1.0
Oversampling target	—	400/class
Precision	fp32	bf16/fp16
Random seed	42	42

B Prompt Template

The following template is used to construct the input for both training and inference. During training, the assistant turn contains the gold label digit; during inference, the generation prompt cue is appended and the assistant turn is left empty for the model to complete.

```
[User]:
You are a clinical psychology expert specializing in defense mechanism analysis. Classify the psychological defense level of the FINAL seeker utterance based on the conversation history provided.
```

```
Defense Level Reference:
0 = No Defenses
1 = Action Defenses (Passive Aggression, Help-Rejecting Complaining, Acting Out)
2 = Major Image-Distorting (Splitting, Projective Identification)
3 = Disavowal Defenses (Denial, Rationalization, Projection, Autistic Fantasy)
4 = Minor Image-Distorting (Devaluation, Idealization, Omnipotence)
5 = Neurotic Defenses (Repression, Dissociation, Reaction Formation, Displacement)
6 = Obsessional Defenses (Isolation of Affect, Intellectualization, Undoing)
7 = High-Adaptive Defenses (Affiliation, Altruism, Anticipation, Humor, Self-Assertion, Sublimation, Suppression)
8 = Needs More Information
```

```
Rules:
- Read the full conversation for context.
- Focus on the TARGET utterance’s defensive style, not the supporter’s.
- Respond with ONLY a single digit (0-8). No explanation, no punctuation.
```

```
- Conversation History -
[Speaker]: [utterance]
[Speaker]: [utterance]
...

- Target Utterance [Seeker] -
[target utterance text]
```

```
What is the defense level? Respond with ONLY a single digit (0-8):
```

```
[Assistant]:
{label digit}
```

Explainers at PsyDefDetect: Hierarchical Prompting and Representation-Based Classification for Psychological Defenses

Liudmila Babakova
Moscow State
Pedagogical University
babakovalyuda@gmail.com

Christopher Luongo-Vázquez
Universidad de La Rioja
chluongo@unirioja.es

Ilia Stepin
Universidad Autónoma
de Madrid
ilia.stepin@uam.es

Abstract

Psychological defense detection is one of essential present-day challenges in clinical practice. The state-of-the-art natural language processing (NLP) tools aim to automate this task. However, their potential and efficiency remain largely unexplored. This manuscript attempts to address this problem from various perspectives: it first explores the efficiency of direct large language model (LLM)-prompting. Then, it applies NLP techniques for LLM fine-tuning applied to the psychological defense classification task. Finally, it attempts to generate states of mind based on the speaker’s psychological state. The results show that the complexity of the task requires further improvement of the software solutions used.

1 Introduction

One of the present-day psychotherapy challenges consists in automating psychological defense detection. Large language models (LLMs) have been shown to have great potential to revolutionize psychotherapy. Nevertheless, their effectiveness, when applied to treating mental health issues, remains underexplored, as they have only been shown to offer static solutions to single mental health-related problems (Na et al., 2025).

This manuscript presents (1) a description of the methodological aspects, (2) experimental design as well as (3) the empirical results obtained while participating in the shared task PsyDefDetect (Na et al., 2026a). Essentially, it explores the potential of some of the most widely used LLMs applied to a corpus of multi-turn dialogues that are labeled to highlight a possible psychological defense mechanism (Na et al., 2026b).

The remainder of the manuscript is structured as follows. Section 2 describes the methodology applied as well as the experimental settings. Section 3 reports the results obtained. Finally, section 4 concludes the manuscript.

2 Methodology

Four series of experiments have been carried out as part of this work. First, prompting-related experiments took place for an LLM to directly classify the given (non-)defensive utterances (see Section 2.1 for details). Subsequently, encoder fine-tuning experiments were conducted to estimate the effect of hyperparameter fine-tuning of the given language models (see Section 2.2 for details). Afterwards, a novel technique for state-of-mind generation was proposed and tested in the context of psychological defense classification (see Section 2.3 for details). Finally, LLM fine-tuning makes part of the concluding series of experiments. The corresponding methodology is described in Section 2.4.

All the experiments were carried out using the PsyDefConv dataset (Na et al., 2026b) containing nine classes (0 – no defense, 1 – action defenses, 2 – major image-distorting, 3 – disavowal defenses, 4 – minor image-distorting, 5 – neurotic defenses, 6 – obsessional defenses, 7 – high-adaptive defenses, 8 – needs more information).

2.1 LLM prompting

The prompting experiments were designed to include three internal stages: (1) direct prompting, (2) few-shot prompting, and (3) two-step psychological defense classification. Four baseline models have been used in all the corresponding experiments. Namely, these are DeepSeek-v3.2 (DeepSeek-AI et al., 2025), Qwen3 (Qwen3-235B-A22B) (Team, 2025), Qwen2.5-72B-Instruct (Team, 2024), and GLM-5 (GLM-5-Team et al., 2026). First, all the aforementioned models were prompted to predict the type of psychological defense mechanism (if any). Subsequently, they were enhanced with several input examples to enable in-context learning. Finally, two-step predictions were made. First, all the utterances underwent a binary classification task (defensive vs.

non-defensive). Second, all the utterances claimed defensive were subsequently assigned labels of the corresponding defensive dataset classes. To do so, both binary and multiclass classification requests were enhanced with few-shot prompts generated in stage (2).

2.2 Encoder fine-tuning

In the second phase, we evaluate encoder-based models on a simplified binary formulation of the task aimed at improving recognition of non-defensive utterances. We compare two labeling strategies. In the *Unmerged* setting, models are trained on the original multiclass labels and later evaluated with respect to the binary distinction. In the *Merged* setting, label 0 (*No Defense*) is preserved, while all non-zero labels are collapsed into a single positive class representing the presence of any defense mechanism.

Several transformer encoders were used in the corresponding experiments. Namely, those are BERT (Devlin et al., 2018), ModernBERT (Warner et al., 2024), RoBERTa (Liu et al., 2019), and DeBERTa-v3 (He et al., 2021). For each model and labeling setup, we explore the effect of setting multiple learning rates (2×10^{-4} , 2×10^{-5} , and 2×10^{-6}) in order to assess optimization sensitivity and identify robust configurations. The best-performing binary classifier from this phase is later incorporated as the filtering component in our hybrid retrieval-based system.

2.3 State-of-mind generation

As part of the original pipeline, we additionally propose a hybrid framework that combines LLM-based semantic abstraction, instance-based classification, and binary filtering. First, each dialogue context together with the target utterance is transformed into a structured *state-of-mind* representation using GPT-5.2. Rather than operating directly on raw dialogues, the model generates a compact description of the speaker’s psychological state, intentions, and coping behavior, providing a normalized representation in which defensive patterns become more comparable across instances.

We then perform classification in this representation space using a k-nearest neighbors (KNN) strategy. The generated state-of-mind descriptions are embedded with Gemma (Schechter Vera et al., 2025)¹. As for each test example, we retrieve the

nearest training representations ($k \in 3, 5, 7$). The final label is assigned by majority voting over the retrieved neighbors. This is argued to allow the system to leverage semantic similarity between psychologically related situations instead of relying only on surface lexical cues.

Finally, to better address the strong class imbalance, we incorporate the best-performing binary classifier from Section 2.2 (RoBERTa *Merged*, learning rate= 2×10^{-5}), which distinguishes *No Defense* (Level 0) from *Any Defense* (Levels 1–8). In the final hybrid setup, utterances predicted as *No Defense* are directly assigned label 0, while all the remaining instances are classified using the KNN module. In this way, the encoder serves as a precision-oriented gate for trivial non-defensive cases, while the retrieval component handles finer-grained defense level prediction.

2.4 LLM fine-tuning

In the concluding phase of the experiments, a single baseline LLM (Llama-3.2-3B) (AI@Meta, 2024) was fine-tuned to assess the corresponding effect on psychological defense mechanism classification. The quantized low-rank adaptation technique (QLoRA) (Detmers et al., 2023) was used to fine-tune the baseline LLM. Two learning rates (1×10^{-4} and 1×10^{-5}) were employed in the corresponding experiments.

3 Experimental results

This section reports the experimental results of all the four groups of the experiments whose methodology is described in section 2. Table 1 reports precision, recall, and F1-scores for all the baseline models tested within all the prompting experiment stages. It is worth noting that DeepSeek performance increases, as few-shot prompting and two-step classification are incorporated at later stages of the experiment whereas this effect is not observed in the case of all the other baseline models used.

Table 2 overviews the encoder fine-tuning experimental results. Remarkably, all the LLMs tested show better performance in the *Merged* setting.

Table 3 presents the state-of-mind generation experimental results as a function of the number of k-nearest neighbors. It can be seen that the model’s performance slightly increases along with the increase in the number of neighbors.

Table 4 reports the effects of the QLoRA-based LLM fine-tuning. These do not appear to signifi-

¹huggingface.co/google/embeddinggemma-300m

Model	Precision	Recall	F1-score
Stage 1: Direct prompting (baseline models)			
DeepSeek-v3.2	0.182	0.213	0.155
Qwen3	0.238	0.226	0.180
Qwen2.5-72B	0.192	0.204	0.152
GLM-5	0.348	0.216	0.178
Stage 2: Few-shot prompting			
DeepSeek-v3.2	0.369	0.259	0.194
Qwen3	0.328	0.248	0.203
Qwen2.5-72B	0.344	0.221	0.175
GLM-5	0.302	0.209	0.172
Stage 3: Defense classification			
DeepSeek-v3.2	0.276	0.133	0.083
Qwen3	0.204	0.137	0.071
Qwen2.5-72B	0.073	0.112	0.031
GLM-5	0.247	0.148	0.088

Table 1: First-phase experimental results (prompting).

Labeling	Learning rate	F1-score
Model: <i>BERT</i>		
	2e-4	0.076
Unmerged	2e-5	0.221
	2e-6	0.138
Merged	2e-4	0.454
	2e-5	0.836
	2e-6	0.794
Model: <i>ModernBERT</i>		
Unmerged	2e-4	0.159
	2e-5	0.234
	2e-6	0.131
Merged	2e-4	0.857
	2e-5	0.853
	2e-6	0.774
Model: <i>RoBERTa</i>		
Unmerged	2e-4	0.076
	2e-5	0.215
	2e-6	0.122
Merged	2e-4	0.454
	2e-5	0.913
	2e-6	0.796
Model: <i>DeBERTa-v3</i>		
Unmerged	2e-4	0.030
	2e-5	0.030
	2e-6	0.076
Merged	2e-4	0.144
	2e-5	0.144
	2e-6	0.454

Table 2: Second-phase experimental results (encoder fine-tuning).

# neighbors	Precision	Recall	F1-score
3	0.174	0.173	0.170
5	0.175	0.176	0.171
7	0.190	0.177	0.171

Table 3: Third-phase experimental results (LLM-based state-of-mind representation).

Learning rate	Precision	Recall	F1-score
1e-4	0.139	0.201	0.164
1e-5	0.130	0.135	0.116

Table 4: Fourth-phase experimental results (LLM fine-tuning).

cantly differ depending on the changes in the learning rate employed.

Based on the insights from the experiments carried out, the final, 7-neighbor state-of-mind generation model was passed on to the shared task organizers so that the test results could be calculated. These are presented in Table 5.

4 Conclusion

The predominantly negative results obtained emphasize the complexity of the psychological defense mechanism classification task. They highlight, for instance, the need for extensive use of data preprocessing techniques. It is particularly necessary when using the provided dataset due to a high degree of class imbalance among the data instances. In addition, better elaborated fine-tuning mechanisms are needed to improve the performance of the models tested.

As part of future work, a more thorough comparative analysis of various language models seems required. The experimental results do not allow us to conclude how effective the presented conceptualization of LLM-based state-of-mind representations is. That said, they pave the way toward further refining of the methodological basis of the experiments presented in this manuscript.

Limitations

Regarding the first-phase experiments (prompting), these only include a limited set of LLMs. A com-

Accuracy	Precision	Recall	F1-score
0.6144	0.2366	0.1660	0.1612

Table 5: Test results.

parison against a wider set of models as well as the use of alternative techniques reshaping the original prompts might help improve the performance.

As for the second-phase experiments (encoder fine-tuning), only a limited number of models and hyperparameters has been used for fine-tuning. To be precise, only the learning rate for four LLMs has been modified while running the corresponding experiments. In addition, the experimental results reported do not allow us to make conclusions on their statistical significance.

When it comes to the third-phase experiments (state-of-mind generation), these employ only one classic machine learning classifier. It is therefore necessary to explore the potential of state-of-the-art classifiers to ensure adequate classification of, for example, defensive vs. non-defensive cases. In addition, the results reported indicate that improved data preprocessing might help improve the model performance, particularly when dealing with class imbalance.

Finally, the fourth-phase experiments (LLM fine-tuning) include only one preselected LLM. In both experiments carried out, only the learning rate was changed as part of fine-tuning. This may be argued not to be sufficient to improve the discriminatory capacity of the corresponding classifier.

References

- AI@Meta. 2024. [Llama 3 model card](#).
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. [Deepseek-v3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: efficient finetuning of quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *Preprint*, arXiv:1810.04805.
- GLM-5-Team, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, and 167 others. 2026. [GLM-5: from Vibe Coding to Agentic Engineering](#). *Preprint*, arXiv:2602.15763.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing](#). *Preprint*, arXiv:2111.09543.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A Robustly Optimized BERT Pretraining Approach](#). *Preprint*, arXiv:1907.11692.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.
- Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 69 others. 2025. [EmbeddingGemma: Powerful and Lightweight Text Representations](#). *Preprint*, arXiv:2509.20354.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Qwen Team. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference](#). *Preprint*, arXiv:2412.13663.

A LLM prompts

This appendix reports the LLM prompts used across the first- and third-phase experiments.

In the first-phase experiments, the following prompt was used: “You are a strict classifier. Choose exactly one label: 0 or 1. Use the following guide: 0 - No Defenses: Functional utterances that maintain conversational flow without engaging conflict. 1 - Action Defenses: Passive Aggression, Help-Rejecting Complaining, Acting Out. Distress is released by acting on the environment instead of reflecting. 2 - Major Image-Distorting: Splitting (self/other), Projective Identification. Reduces anxiety via all-good/all-bad distortions of self or other. 3 - Disavowal Defenses: Denial, Rationalization, Projection, Autistic Fantasy. Rejects threatening reality by denying, excusing, blaming, or fantasizing. 4 - Minor Image-Distorting: Devaluation/Idealization (self or other), Omnipotence. Softer distortions temporarily inflate or deflate self-esteem. 5 - Neurotic Defenses: Repression, Dissociation, Reaction Formation, Displacement. Keeps unacceptable motives out of awareness; feelings surface indirectly. 6 - Obsessional Defenses: Isolation of Affect, Intellectualization, Undoing. Uses excessive logic or symbolic acts to separate feelings from events. 7 - High-Adaptive Defenses: Affiliation, Altruism, Anticipation, Humor, Self-Assertion, Self-Observation, Sublimation, Suppression. Mature coping that integrates emotion and thought to channel affect constructively. 8 - Needs More Information: Label used when an utterance is too ambiguous or lacks context. Return only JSON in one line: {“label”: <0 or 1>, “reason”: “short reason”}.”

In the third-phase experiments, the following prompt was used: “You are a psychologist, skilled in Cognitive Behavioral Therapy (CBT), Eye Movement Desensitization and Reprocessing (EMDR) protocol, Acceptance and Commitment Therapy, and Mindfulness practices, providing empathetic guidance in psychotherapy. Your task is to briefly describe the current state of mind of the seeker based on the ongoing conversation between seeker and supporter. Focus on features most relevant to identifying psychological defense mechanisms — specifically: how the seeker relates to distressing emotions (avoids, intellectualizes, acts out, suppresses, projects, denies, or integrates them); whether their perception of self or others appears distorted (idealized, devalued, split into all-

good/all-bad); how they respond to help (accept, resist, reject); the degree of self-awareness and reflective capacity; and any signs of fantasy, displacement, rationalization, or mature coping such as humor or self-assertion. Describe in 2-4 sentences.”

PerceptionLab at PsyDefDetect: Overcoming Extreme Response Bias in LLMs via Rubric-Grounded Retrieval and Supervised Clinical Reasoning Distillation for Fine-Grained Ordinal Classification

Tamjid Hasan Fahim, Syed Asif Johan, Saad Bin Maksud

Department of Computer Science & Engineering
Rajshahi University of Engineering and Technology
Rajshahi-6204, Bangladesh

{tamjidhfahim, connect.syedasifjohan, sbinmaksud}@gmail.com

Abstract

Automating the classification of psychological defense mechanisms is a critical yet challenging frontier in clinical natural language processing. General-purpose Large Language Models (LLMs) struggle to apply fine-grained ordinal frameworks like the Defense Mechanism Rating Scales due to the implicit nature of clinical cues and a fundamental clinical reasoning gap. These models exhibit severe extreme response bias, systematically gravitating toward the scale’s endpoints while failing to resolve nuanced, mid-level defenses. In this paper, we present our third-place system for the PSYDEFDETECT Shared Task at BioNLP 2026, designed specifically to overcome this failure mode. We propose a hybrid architecture that synergizes label-flattened generative retrieval with an LLM classifier fine-tuned via the distillation of supervised clinical reasoning traces. This dual approach, grounding decisions in rubric criteria while leveraging task-specific supervision, successfully mitigates the observed bias, achieving an accuracy of 67.37% and a macro-F1 of 39.56%. Our work provides empirical evidence that tightly integrating targeted clinical supervision with dynamic rubric-grounded retrieval significantly outperforms the raw parameter scale of un-tuned foundation models.

1 Introduction

Understanding how individuals psychologically defend against emotional distress is central to psychotherapeutic practice. The *Defense Mechanism Rating Scales* (DMRS; Di Giuseppe and Perry 2021) provide a validated framework for categorizing these unconscious processes across seven ordinal maturity levels (1–7). While essential for clinical assessment, manual scoring requires specialized training and does not scale to the large volumes of data generated in digital mental health interventions.

The PSYDEFDETECT Shared Task at BioNLP 2026 (Na et al., 2026a) formalizes the challenge of automating this process. This task presents three fundamental challenges. First, defense mechanisms exhibit subtle overlap between adjacent levels, where labels share significant semantic and pragmatic features, challenging standard classifiers that treat them as discrete categories rather than an ordered spectrum. Second, general-purpose LLMs exhibit a clinical reasoning gap. They lack the domain-specific reasoning required to apply the rigid criteria of a professional rubric, leading to a disconnect between surface-level sentiment and the underlying unconscious processes. Finally, defense mechanisms are expressed through implicit and context-dependent cues, resulting in a low signal-to-noise ratio. Accurate classification therefore depends less on explicit lexical indicators and more on latent patterns.

In this paper, we present a highly competitive hybrid system designed to address these inherent challenges, securing third place among 21 registered teams in the official shared task evaluation. We introduce a teacher-student distillation framework that encodes structured chain-of-thought clinical reasoning into a smaller LLM, eliminating the need for long-context rubric-handbook prompting at inference time. We further augment this supervised fine-tuning (SFT) pipeline with a dynamic DMRS-Q item retrieval framework to ground predictions in the rubric’s nuanced criteria. Crucially, by flattening the clinical rubric during retrieval and removing ordinal identifiers from the prompt and output space, the system mitigates the model’s tendency toward extreme response bias. Ultimately, our results demonstrate that automating nuanced psychological assessment depends less on continuously scaling foundation models and more on tightly integrating targeted clinical supervision with dynamic, rubric-grounded retrieval.

2 Related Work

The application of Natural Language Processing to mental health has expanded rapidly, moving from depression and suicide risk detection (Coppersmith et al., 2018; Gkotsis et al., 2017) to modeling clinical dialogue dynamics like empathy and therapeutic intent (Pérez-Rosas et al., 2019). However, Na et al. (2025) highlight fragmentation in LLM research on psychotherapy and limited theoretical integration, suggesting the automated analysis of defense mechanisms remains a highly significant yet underexplored frontier. Most prior psycholinguistic analyses of therapy transcripts have relied on lexicon-based features or shallow classifiers (Miner et al., 2022; Kim et al., 2025). These approaches struggle with the deep pragmatic context and subtle cues of the DMRS framework, where signals are encoded in implicit narrative structure rather than explicit lexical content.

A primary obstacle in modeling these mechanisms is fine-grained ordinal classification (Zhang et al., 2017). Defense mechanisms lie on a continuum, and adjacent levels often differ only by subtle shifts in intent, affect, or defensive strategy. In this setting, direct prompting of LLMs often reveals *extreme response bias* — a tendency to choose extreme categories on an ordinal rating scale, leading to overly extreme or highly confident judgments. A related failure mode is *extreme label bias*, where predictions disproportionately collapse toward the most salient extreme classes rather than reflecting the full graded structure of the rubric (Reif and Schwartz, 2024; Wang et al., 2024). This behavior exposes a deeper clinical reasoning gap: general-purpose LLMs do not reliably apply the nuanced, rule-bound distinctions required to separate highly overlapping ordinal states.

To bridge this gap, recent advancements in chain-of-thought prompting (Wei et al., 2022) and knowledge distillation (Ho et al., 2023; Magister et al., 2023) offer a pathway to teach models the requisite intermediate clinical logic. Concurrently, Retrieval-Augmented Generation (RAG; Lewis et al. 2020) has proven effective at grounding models in complex texts, often improving downstream reasoning and classification performance (Shi et al., 2024). By retrieving matching patterns directly from the clinical handbook, models can bypass the need to classify utterances in a vacuum. Our work synthesizes these paradigms, demonstrating that LLMs with task-specific supervision through distilled rea-

soning traces, combined with rubric-grounded retrieval, and chain-of-thought generation, provide a robust solution to the complexities of clinical NLP.

3 Methodology

We frame defense mechanism classification as a supervised sequence classification task over the PSYDEFCONV (Na et al., 2026b) dataset. Following the shared-task setup, each seeker utterance is assigned one of nine task labels (0–8): the seven DMRS maturity levels plus two auxiliary categories. We first establish traditional encoder-only baseline, followed by a progression of seven LLM-based systems designed to systematically address the specific challenges of this clinical classification task.

3.1 Approach 1: Encoder-Based BERT Variants

To assess whether a simpler encoder-only architecture could solve the task, we fine-tuned two BERT-family models: the larger *bert-large-uncased* (Devlin et al., 2018), and *mental-bert-base-uncased* (Ji et al., 2022), a smaller but domain-adapted variant pre-trained on mental health-related Reddit posts. We represent the dialogue history as a single sequence using explicit speaker-role markers ([SEEKER] and [SUPPORTER]), concatenate the context and target utterance with the [SEP] token, and apply left-side truncation to preserve the most recent content when the input exceeds 512 tokens. We split the training data into train and validation subsets using a stratified 85/15 split to preserve the label distribution, and optimize the models with class-weighted cross-entropy to address class imbalance. Both models are trained for 5 epochs with a learning rate of 2×10^{-5} .

3.2 Approach 2: Direct Prompting with Strong LLM

In this few-shot baseline, we provide the full DMRS handbook ($\sim 12k$ tokens) and explicitly include three examples for each of the defense levels. We instruct Gemini 3.1 Pro Preview (Temperature = 0.1) to predict the defense level by generating a structured, four-step chain-of-thought rationale (See Appendix A.): (i) *Context Analysis* (situating the utterance within the dialogue history); (ii) *Goal Identification* (determining the communicative intent of the target utterance); (iii) *DMRS Alignment* (mapping clues to the handbook criteria); and (iv) *Differential Analysis* (ruling out adja-

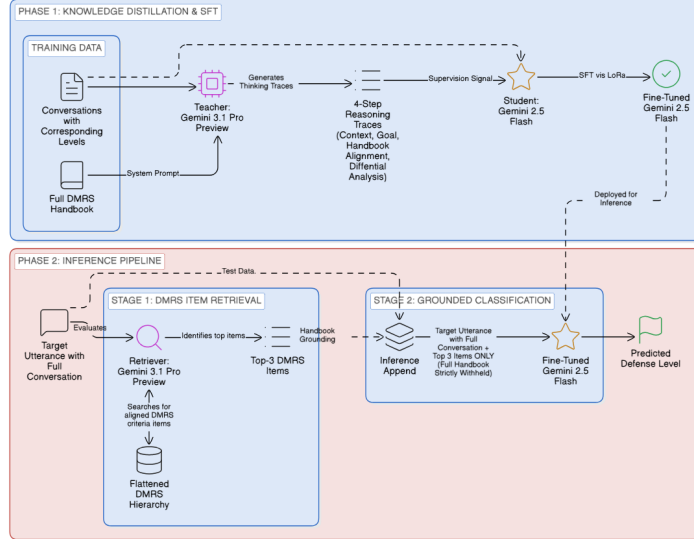


Figure 1: Methodology Diagram of the Best Performing System (Approach 7).

cent levels). Despite the model’s one-million-token context window, it produces a highly skewed prediction distribution, concentrating near the ordinal extremes, especially levels 0 (*No Defenses*) and 7 (*High-Adaptive Defenses*). We treat this extreme response bias as a key failure mode of direct long-context prompting, motivating the development of the subsequent approaches.

3.3 Approach 3: SFT of Small LLM with Knowledge Distillation

We introduce task supervision by fine-tuning *Ministral-8B-Instruct-2410* via a teacher-student distillation paradigm. A teacher model (Gemini 3.1 Pro Preview) with access to the full DMRS handbook and ground-truth label generates structured reasoning traces for each training example, following the four-step rationale from Approach 2. We then fine-tune the student model on these traces using a balanced and trimmed training set. By distilling the rubric’s knowledge directly into the generated reasoning traces, the handbook and the few-shot examples can be withheld at inference time, eliminating the massive prompt overhead while successfully preserving the required clinical logic.

3.4 Approach 4: SFT of Strong LLM

To evaluate the effect of model scale on task performance, we apply the same knowledge distillation framework from Approach 3 to Gemini 2.5 Flash, a substantially larger model. The model is fine-tuned via Google Vertex AI (LoRA rank = 16; 5 epochs; learning_rate_multiplier=1) on the full train-

ing set of the teacher-generated reasoning traces. This experiment tests whether a stronger model can better resolve the subtle distinctions within the DMRS hierarchy than the 8B-parameter student model used in Approach 3.

3.5 Approach 5: DMRS-Q Items Retrieval with Weighted Voting

To address the needle-in-a-haystack problem of locating relevant passages within a long clinical rubric, we reframe the classification task as extraction. We observe that a foundation LLM, despite poor ordinal classification performance, is a capable candidate *extractor*. We flatten the DMRS hierarchy by removing all explicit ordinal identifiers, including level numbers, hierarchy markers, and level-specific groupings, and prompt Gemini 3.1 Pro Preview to perform generative retrieval with explicit justification, identifying the top-3 DMRS-Q items most aligned with a target utterance within its conversational context. This flattening technique removes the ordinal target structure from the retrieval space, preventing the model from triggering its extreme response bias during extraction. The retrieved items are subsequently mapped back to their parent levels (0–8), and a final prediction is made via weighted voting. We evaluated two heuristic voting schemes: a strict top-1-dominant setting [0.7, 0.3, 0.1], and a more balanced alternative [0.45, 0.35, 0.20] that allows the combined weight of the second and third retrievals to outweigh the first when they agree.

System	Description	Accuracy	Precision	Recall	Macro-F1
1a	bert-large-uncased Finetuning	45.6	21.1	20.6	20.2
1b	mental-bert-base-uncased Finetuning	35.0	26.0	24.7	22.3
2	Direct Prompting (gemini-3.1-pro-preview)	38.3	37.6	14.9	20.5
3	SFT Small LLM (Ministral-8B-Instruct-2410)	42.6	25.3	28.4	25.3
4	SFT Strong LLM (gemini-2.5-flash)	52.1	33.0	23.7	26.5
5a	Retrieval + Voting [0.7, 0.3, 0.1]	64.8	35.0	37.3	34.5
5b	Retrieval + Voting [0.45, 0.35, 0.20]	65.0	39.0	39.6	37.0
6	Retrieval + Pro Re-ranking	64.7	41.1	37.2	35.7
7	System 4 + Retrieval	67.4	42.6	40.9	39.6
8a	System 7 + Pro Re-ranking (18 samples)	66.9	41.2	40.4	38.6
8b	System 7 + Pro Re-ranking (131 samples)	66.9	42.0	37.0	35.9

Table 1: Performance of all evaluated systems corresponding to the approaches described in Section 3. All values are percentages. Bold denotes the best result per column.

3.6 Approach 6: Retrieval + Stronger LLM Re-ranking

Static voting cannot effectively resolve cases where the top-3 retrieved items map to contradictory levels, as it lacks any mechanism for reasoning over conflicting evidence. We therefore propose a two-stage pipeline in which the retriever from Approach 5 produces candidate levels, and Gemini 3.1 Pro Preview re-evaluates them with explicit chain-of-thought reasoning. We passed *only* the corresponding handbook sections rather than the full rubric as system prompt. This narrows the model’s attention to at most 3 candidate levels, directly mitigating the needle-in-a-haystack problem.

3.7 Approach 7: Supervision + Retrieval (Best System)

Our best performing method synthesizes the strengths of task-specific supervision (Approach 4) and dynamic handbook grounding (Approach 5).

We prompt the fine-tuned Gemini 2.5 Flash classifier (from Approach 4) with the full conversational context, the target utterance, and the dynamically retrieved top-3 DMRS-Q items (from Approach 5), as illustrated in Figure 1. The full DMRS handbook is strictly withheld from the system instruction to eliminate inference overhead and avoid context dilution.

We designed this hybrid architecture to test whether the model could leverage its internalized SFT clinical reasoning while grounding its prediction in the narrow, dynamic context of the retrieved items.

3.8 Approach 8: Selective Re-ranking via Stronger LLM

Building on our best-performing system (Approach 7), we test whether a stronger foundation model can serve as an effective tie-breaker in cases

of severe retrieval contradiction. While the fine-tuned Flash model excels generally, we hypothesized that the superior reasoning capacity of Gemini 3.1 Pro Preview could better resolve edge cases, despite lacking task-specific fine-tuning. Predictions from Approach 7 are retained when all top-3 retrievals agree ($\sim 75\%$ of cases). When retrievals conflict, the stronger Pro model is invoked to resolve the prediction using explicit chain-of-thought reasoning over the handbook sections of the candidate levels. We evaluate this cascade under two disagreement thresholds: cases where all three retrievals differ (18 samples), and cases where at least two differ (131 samples).

4 Results

Table 1 reports Accuracy, Precision, Recall, and macro-F1 for all systems on the PSYDEFCONV test set (472 examples), with metrics computed over the positive classes (1–8) only. The detailed class-wise behavior of the best system (System 7) is shown in Appendix Figure 2.

Encoder baselines are weak lexical benchmarks.

Although encoder-based models establish a foundation, they lack the capacity to explicitly reason over subtle cues and the complex DMRS rubric. Notably, the larger bert-large-uncased model underperforms its smaller but domain-adapted counterpart, mental-bert-base-uncased, suggesting that domain adaptation is more beneficial than model scale for this task.

Supervision is essential; scale is not a substitute.

The state-of-the-art LLM (System 2) completely fails despite accessing the full DMRS handbook within its large context window, achieving performance comparable to BERT-based baselines (System 1). The foundation model exhibits severe

extreme response bias, collapsing into confident, plausibly reasoned, yet clinically ungrounded predictions at the extreme ends of the scale (levels 0 and 7). This confirms that fine-grained ordinal classification fundamentally requires task-specific SFT.

Furthermore, comparing our supervised models reveals the nuanced impact of model scale. Although the substantially larger Gemini 2.5 Flash (System 4) improves accuracy substantially over the 8B-parameter student (System 3), its macro-F1 gain is surprisingly modest. This suggests that scale alone does not resolve the minority-class and ordinal ambiguity challenges. Together, these results demonstrate that the clinical supervision signal is the true bottleneck; prompt engineering and raw parameter scale alone cannot overcome the clinical reasoning gap.

Retrieval is the single most impactful intervention. Reframing the task as generative extraction (System 5) is the most effective isolated intervention, yielding a substantial performance leap without any additional training. As hypothesized, removing the ordinal identifiers from the prompt and output spaces via rubric flattening successfully neutralizes extreme response bias while mitigating the needle-in-a-haystack problem for downstream inference.

Interestingly, comparing the voting weight configurations in Systems 5a and 5b reveals that down-weighting the top-1 item improves overall performance. This suggests that while the foundation LLM is an excellent candidate extractor, its raw initial ranking is not a sufficiently reliable proxy for direct classification.

Supervision and retrieval are strongly complementary. Our best system (System 7) combines dynamically retrieved DMRS-Q items with the fine-tuned Gemini 2.5 Flash classifier, producing a substantial +13.1 macro-F1 improvement over using SFT alone (System 4). Furthermore, we find that the fine-tuned model demonstrates strong resilience to retrieval noise¹, successfully correcting predictions even in cases where all three retrieved candidates map to incorrect levels. This suggests that retrieval and SFT capture robust, mutually reinforcing signals: retrieval surfaces localized DMRS evidence to ground the prediction and narrow the decision space, while SFT provides the clinical

reasoning needed to interpret, weigh, and sometimes override that evidence under noisy retrieval conditions.

General-purpose cascaded re-ranking amplifies retrieval noise. Systems 6 and System 8 expose the strict limitations of relying on un-tuned foundation models to evaluate retrieved candidates. Replacing static voting with LLM re-ranking (System 6) actively degrades performance relative to heuristic voting (System 5). Because both systems evaluate the exact same retrieved items, this demonstrates that general-purpose reasoning lacks the clinical logic necessary to accurately weigh candidate labels—proving the bottleneck is the reasoning process itself, not just upstream retrieval accuracy. Furthermore, invoking this stronger foundation model as a selective tie-breaker for contradictory retrievals (System 8) underperforms our SFT-driven hybrid approach (System 7). In both cascade scenarios, the explicit clinical logic internalized by the fine-tuned model proves far more reliable at resolving ambiguities and recovering from retrieval noise. Ultimately, these results reiterate that task-specific supervision is strictly superior to raw parameter scale, even at the re-ranking stage.

5 Conclusion

Our results show that raw parameter scale alone cannot overcome the clinical reasoning gap inherent in foundation models for fine-grained ordinal defense mechanism classification. Instead, task-specific supervised fine-tuning and retrieval-based grounding provide complementary signals that substantially improve performance. A key caveat is that our best system (System 7) has train-inference asymmetry: retrieval is injected only at inference time, while the fine-tuned classifier is trained without retrieved DMRS-Q items. This means the student learns from reasoning traces alone and then adapts to appended handbook evidence at test time, which may limit its ability to fully exploit the retrieved context. A natural next step is retrieval-augmented fine-tuning, where retrieved items are included during training so the model can learn to integrate them end-to-end. More broadly, future work in clinical NLP should focus on improving upstream retrieval precision, refining knowledge distillation for supervision, and ultimately deepening the synergy between targeted clinical supervision and rubric-grounded retrieval.

¹We provide a qualitative case study in Appendix B.

Limitations

Our study has several limitations. First, our best system exhibits train-inference asymmetry, since retrieved DMRS-Q items are provided only at inference time and not during fine-tuning. We discuss this as a major direction for retrieval-augmented training. Second, our distillation as well as retrieval pipeline depends on proprietary models, which introduces reproducibility concerns and non-trivial API costs, limiting accessibility for resource-constrained research groups. Third, due to compute constraints, the Ministral-8B model in Approach 3 was trained for only 2 epochs on a trimmed, balanced subset, and Gemini 2.5 Flash SFT was not extended beyond 5 epochs despite still improving, suggesting our results are a lower bound on what supervised fine-tuning can achieve. Finally, our retrieval step relies on a single generative model without re-ranking diversity, making the pipeline vulnerable to systematic retrieval errors that downstream models cannot always recover from.

References

- Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. [Natural language processing of social media as screening for suicide risk](#). *Biomedical Informatics Insights*, 10:1178222618792860.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.
- Mariagrazia Di Giuseppe and J. Christopher Perry. 2021. [The hierarchy of defense mechanisms: Assessing defensive functioning with the defense mechanisms rating scales q-sort](#). *Frontiers in Psychology*, Volume 12 - 2021.
- George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J. P. Hubbard, Richard J. B. Dobson, and Rina Dutta. 2017. [Characterisation of mental health conditions in social media using informed deep learning](#). *Scientific Reports*, 7(1).
- Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. [Large language models are reasoning teachers](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882. Association for Computational Linguistics.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. [Mental-BERT: Publicly available pretrained language models for mental healthcare](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.
- Samuel Kim, Oghenemaro Imieye, and Yunting Yin. 2025. [Interpretable depression detection from social media text using llm-derived embeddings](#). *Preprint*, arXiv:2506.06616.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. [Teaching small language models to reason](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.
- Adam S. Miner, Scott L. Fleming, Albert Haque, Jason A. Fries, Tim Althoff, Denise E. Wilfley, W. Stewart Agras, Arnold Milstein, Jeff Hancock, Steven M. Asch, Shannon Wiltsey Stirman, Bruce A. Arnow, and Nigam H. Shah. 2022. [A computational approach to measure the linguistic characteristics of psychotherapy timing, responsiveness, and consistency](#). *npj Mental Health Research*, 1(1).
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. [Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations](#). In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. [You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.
- Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. [What makes a good counselor? learning to distinguish between high-quality](#)

- and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.
- Yuval Reif and Roy Schwartz. 2024. **Beyond performance: Quantifying and mitigating label bias in LLMs**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6784–6798, Mexico City, Mexico. Association for Computational Linguistics.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. **Replug: Retrieval-augmented black-box language models**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 8371–8384. Association for Computational Linguistics.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. **Large language models are not fair evaluators**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. **Chain-of-thought prompting elicits reasoning in large language models**. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. **Ordinal common-sense inference**. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Appendix

A Prompt Listings

```
class ClinicalReasoning(BaseModel):
    context_trigger: str = Field(
        description="Description of the
        stressor in the dialogue."
    )
    psychological_goal: str = Field(
        description="What the speaker is
        trying to achieve/avoid."
    )
    handbook_alignment: str = Field(
        description="Specific evidence
        from the handbook that justifies the
        label."
    )
    differential_diagnosis: str = Field(
        description="Why this isn't a
        higher or lower-level defense."
    )

class DefensePrediction(BaseModel):
    clinical_reasoning:
    ClinicalReasoning
    defense_level: int = Field(
        description="The numeric defense
        level (0-8).")
    label: str = Field(description="The
    specific name of the defense
    mechanism used.")

SYSTEM_INSTRUCTION = f"""
You are an expert clinical psychologist
and data annotator. Your task is to
analyze dialogues and classify the
psychological defense mechanism used
in the 'current_text_to_classify'
based on the Defense Mechanisms
Rating Scales (DMRS) hierarchy.

You must assign exactly one label from
the list below:
{LABEL_DESCRIPTIONS}

Your classification must be grounded in
the DMRS hierarchy given below. This
following comprehensive HANDBOOK
serves as your core classifying
guideline:
{HANDBOOK_TEXT}

Here are some examples of how to reason
through the task:
{FEW_SHOT_EXAMPLES}

CORE INSTRUCTIONS:
1. Primacy of Context: Always read the
preceding dialogue to understand
what triggered the '
current_text_to_classify'.
2. Function-Oriented: Ask yourself, "
What psychological goal is the
speaker trying to achieve?"
3. Handbook Grounded: Match the behavior
to the specific criteria in the
DMRS Handbook. Reason through why
specific criteria are met.
```

```
4. You must maintain hierarchical
integrity explicitly reason
through why the classification does
not drift into higher (more adaptive
) or lower (more pathological)
levels by verifying that all
exclusionary criteria for the
selected level are met.
5. Distinguish Emotion from Defense:
Saying "I am sad" is Level 0. A
defense requires distortion,
avoidance, or transformation.
6. Always pick the single most accurate
label (0-8) from the LABEL
REFERENCE above.
7. Output strict JSON matching the
requested schema.
"""
```

Listing 1: Base System Prompt primarily utilized in Approach 2. Subsequent approaches adapt this template by entirely withholding the handbook and few-shot examples during inference for supervised models, substituting the handbook with dynamically retrieved top-3 DMRS-Q items, or isolating specific candidate sections of the handbook for re-ranking.

```
class ClinicalReasoning(BaseModel):
    context_trigger: str = Field(
        description="Description of the
        stressor in the dialogue."
    )
    psychological_goal: str = Field(
        description="What the speaker is
        trying to achieve/avoid."
    )
    handbook_alignment: str = Field(
        description="Specific evidence
        from the handbook that justifies the
        defense level (0-8).")
    differential_diagnosis: str = Field(
        description="Why this isn't a
        higher or lower-level defense."
    )

SYSTEM_INSTRUCTION = f"""
You are an expert clinical psychologist
and data annotator. Your task is to
analyze dialogues and generate the
exact clinical reasoning (thought
trace) that perfectly justifies the
PROVIDED psychological defense
mechanism level for the '
current_text_to_classify', based on
the Defense Mechanisms Rating Scales
(DMRS) hierarchy.

The correct true label level will be
given to you in the prompt. You must
explain *why* it is the correct
level.
{LABEL_DESCRIPTIONS}

Your classification reasoning must be
grounded in the DMRS hierarchy given
below. This following comprehensive
```

```

HANDBOOK serves as your core
classifying guideline:
{HANDBOOK_TEXT}

CORE INSTRUCTIONS:
1. Primacy of Context: Always read the
preceding dialogue to understand
what triggered the '
current_text_to_classify'.
2. Function-Oriented: Ask yourself, "
What psychological goal is the
speaker trying to achieve or avoid?"
3. Handbook Grounded: Match the behavior
to the specific criteria in the
DMRS Handbook. Reason through why
specific criteria are met pointing
towards the provided correct label.
4. Hierarchical Integrity: You must
maintain hierarchical integrity
explicitly reason through why the
classification does not drift into
higher or lower levels by verifying
that all exclusionary criteria for
the selected level are met.
5. Emulate the Correct Path (CRITICAL):
Provide the thought process as if
you independently arrived at the
provided correct label.
6. Output strict JSON matching the
requested schema, ensuring your
output 'defense_level' matches the
provided correct level.
"""

```

Listing 2: System Prompt for the Reasoning Trace Generation by the Teacher Model.

```

SYSTEM_INSTRUCTION = f"""
You are an expert clinical psychologist
and data annotator.

Below is the Psychological Defense
Mechanism Coding Handbook:
{FLATTEN_HANDBOOK_TEXT}

YOUR TASK:
You will be provided with a dialogue and
a 'current_text_to_classify'.
You must SCAN THROUGH THE HANDBOOK and
pick out the MOST MATCHING 3 DMRS-Q
items for the given text.

CRITICAL INSTRUCTIONS:
1. FOCUS EXPLICITLY ON THE DMRS-Q ITEMS.
Compare the dialogue's scenario
against the specific behaviors
described in the DMRS-Q items.
2. Identify the corresponding defense
mechanism (such as 'Splitting', '
Acting Out', 'Humor', 'Repression').
3. Base your selections strictly on the
Definitions, Given Scenarios,
Examples, and Distinctions provided
in the handbook.
4. Provide exactly 3 matches. If one
defense is blatantly obvious,
provide the top 3 DMRS-Q items that
best capture the nuances of the
interaction.
"""

```

Listing 3: System Prompt for DMRS-Q Items Retrieval.

```

class DMRSItemMatch(BaseModel):
    dmrs_q_item: str = Field(
        description="The exact DMRS-Q
        Item number and text from the
        handbook (e.g., 'ITEM 45: At times
        when expressing...')."
    )
    match_justification: str = Field(
        description="Detailed
        explanation of why this specific
        DMRS-Q item matches the conversation
        based on the handbook's definition,
        scenario, examples, and
        distinctions."
    )
    sublevel_name: str = Field(
        description="The specific
        defense mechanism name from the
        handbook (e.g., 'Passive Aggression',
        'Rationalization', 'Devaluation'
        etc.). Do NOT use the 0-8 level
        numbers."
    )

class ExtractionResult(BaseModel):
    top_3_matches: list[DMRSItemMatch] =
    Field(
        description="Exactly 3 most
        matching DMRS-Q items and their
        details."
    )

```

B Qualitative Case Study: SFT as a Retrieval Error-Correction Mechanism

To examine why supervised fine-tuning combined with retrieval (System 7) improves performance beyond direct prompting with a large foundation model (System 2), retrieval-based weighted voting (System 5), and retrieval-based re-ranking with an unsupervised LLM (System 6), we analyze a representative test case in which all three non-supervised prediction approaches described above fail, while the supervised model correctly predicts the gold label.

This analysis is especially important because the training dataset is highly imbalanced. Table 2 shows the label distribution in the dataset. Within the seven DMRS maturity-level classes, Level 5 is the smallest class, with only 48 examples. Therefore, correct prediction of a Level 5 case by the supervised model is meaningful: **the model is not simply following the majority-class tendency, but is able to identify a minority mid-level defense that all the other systems miss.**

Label	Count
0	296
1	108
2	61
3	99
4	84
5	48
6	172
7	968
8	28

Table 2: Label distribution of the training set. The distribution is highly imbalanced, with Level 7 being the dominant class and Level 5 being the smallest class among the seven DMRS maturity-level labels.

System	Predicted level	Correct?
Gold	5	–
System 2	0	No
System 5	7	No
System 6	7	No
System 7	5	Yes

Table 3: Prediction outcomes for test_00377, a Level 5 minority-class case where Systems 2, 5, and 6 fail, while System 7 recovers the correct label.

In test sample test_00377, where the true gold label is Level 5, the un-tuned foundation model (System 2, Gemini 3.1 Pro Preview) misinterprets the clinical nuance as a neutral conversational response and predicts Level 0, a textbook illustration of the severe extreme response bias that plagues direct prompting strategies on ordinal scales. For the retrieval-only framework (System 5), the top-3 retrieved DMRS-Q items map to levels 7, 7, and 5 respectively, causing the static heuristic voting scheme to collapse toward the incorrect majority, outputting Level 7. Even when advanced general-purpose reasoning is applied via LLM re-ranking (System 6), the un-tuned foundation model fails to resolve the conflicting evidence and yields Level 7 with plausible but ungrounded reasoning. In sharp contrast, our best hybrid architecture (System 7) successfully demonstrates its resilience to upstream retrieval noise. By harmonizing its internalized supervised reasoning with dynamic handbook grounding, it successfully corrects the misleading retrieval weights to accurately predict the gold Level 5 label.

Specifically, the supervised model recognizes that the target utterance immediately succeeds an unresolved, emotionally charged dialogue regarding abandonment, interpersonal loss, and financial insecurity. Rather than misinterpreting this context as a neutral conversational shift, the model utilizes its internalized clinical logic to identify a Level 5 defensive strategy (*Displacement*), wherein the

speaker defuses acute emotional distress by pivoting to a safer, tangential topic. Crucially, the model’s differential diagnosis proves that it understands why the top-retrieved Level 7 items are incorrect: it explicitly recognizes the speaker is merely sharing a casual anecdote, not actively applying collaborative advice (*Affiliation*) or utilizing an established coping mechanism (*Sublimation*).

Importantly, while the correct DMRS-Q item is surfaced during the initial retrieval phase, its position as the third-ranked candidate causes it to be overridden by conflicting evidence within both the heuristic voting (System 5) and un-tuned re-ranking (System 6) pipelines. This diagnostic outcome strongly substantiates our central finding: **although retrieval-augmented frameworks excel at capturing relevant rubric clauses, task-specific supervised reasoning remains indispensable for reliably interpreting and weighing that evidence when candidate items are noisy or highly competitive.**

The complete dialogue transcript for the test instance test_00377 along with the corresponding reasoning trace outputs for Systems 2, 5, 6, and 7 are presented below. To maximize readability, superficial formatting artifacts, such as escaped new-line characters and markdown emphasis markers, have been removed, while the verbatim wording of the model generations has been strictly preserved.

B.1 Full Conversation and Target Utterance

```

=== test_00377 ===
Supporter: Hello how are you?
Seeker: hello im looking for someone to talk to
im fine how are you
Supporter: I'm doing ok I'm glad you are good. Is
it snowing by you? Merry Christmas!
Seeker: thats great and no its not snowing its
very cold thow merry christmas to you also
Supporter: How can I help you today?
Seeker: im having some issues with friends not
actually being friends
Supporter: I hear you are having trouble figuring
out which friends are really your friends and
which ones aren't. Is that about right?
Seeker: yes
Supporter: I understand that... Did something
happen that makes you feel like that?
Seeker: yes their is no communication or
interaction between me and them nobody answers
phone calls txt messages etc
Supporter: Is this a significant other?
Seeker: no friends
Supporter: Are they ignoring you?
Seeker: when we where younger we all hung out all
the time now that we are older its like they talk
with each other im not sure if i did anything or
not

```

Supporter: Are you sure they are okay? Or not away with family for the holidays?
 Seeker: it doesnt matter they still keep open communication with each other
 Supporter: That's no good what do you think the problem is?
 Seeker: I believe its because I am the friend that has financial trouble and normally needs to borrow money
 Supporter: Do you owe someone already?
 Seeker: no anything i owe i always pay back perhaps they are tired of being asked
 Supporter: That can get in the way of a friendship sometimes, but I try not to get ahead of myself and jump the gun. They may have lost a phone or something may have come up... That very well could be... Are you working?
 Seeker: no im not currently I also think sometimes to myself that they don't answer because they think I'm asking for money
 Supporter: I see, Have you tried for unemployment?
 Seeker: yes i dont qualify
 Supporter: I understand that. What if that is the case? You may need to talk to them and let them know how you feel about that. How long has this been going on?
 Seeker: I'm nervous about asking them why and it happens on and off like when they feel lie it i guess i will say maybe 4 months
 Supporter: I understand that, but if you want to get to a point that you feel better about it then you will have to speak to them.
 Seeker: yes i know i hope everything be immediately resolved and not lead into other problems and end of friendships
 Supporter: Do you see them everyday? I understand that, are you able to go over there.
 Seeker: no i dont live in the same city anymore when i go out of town i can visit but with covid19 rules
 Supporter: Well I can see that you have put the effort in to try and keep the relationship so it sounds like the ball is in their court if they want to do the same.
 Seeker: yes
 Supporter: Well that is about all that can be done and wait for them to come around...Do you have any hobbies?
 Seeker: yes i like to read and play music
 Supporter: I like to go on runs every morning. It helps to clear my mind and makes it easier to brush things off when they bother me. What instrument do you play?
 Seeker: my sister goes on runs in the morning time and says it relax her for the day i think i migh i played the piano when i was in highschool

Target Utterance: my sister goes on runs in the morning time and says it relax her for the day i think i migh i played the piano when i was in highschool

B.2 System 2 Direct Prompting Reasoning Trace

```
{
  "clinical_reasoning": {
```

```
    "context_trigger": "The Supporter shifted the conversation from the Seeker's primary stressor (distant friends) to potential coping mechanisms, specifically mentioning running and asking what instrument the Seeker plays.",
    "psychological_goal": "The Seeker aims to maintain the conversational flow by answering the Supporter's direct question about the instrument and acknowledging their comment about running.",
    "handbook_alignment": "This utterance is a factual and functional response to the Supporter's prompt. According to the handbook, utterances that serve a purely conversational or social function and do not engage with emotional conflict or psychological content are classified as Level 0. Furthermore, reporting factual details of an event or engaging in social small talk on neutral topics are explicitly listed as situations that are typically not defenses.",
    "differential_diagnosis": "This is not a highly adaptive defense like Sublimation because the Seeker is merely recalling a past hobby (playing piano in high school) and considering running, rather than actively channeling their current emotional distress into a creative or athletic pursuit. It does not qualify as any lower-level defense because there is no evidence that a difficult emotion or internal conflict is being avoided, distorted, or transformed."
  },
  "defense_level": 0,
  "label": "No Defense / Neutral Utterance"
}
```

B.3 System 5 Retrieval Trace

```
{
  "dmrs_q_item": "ITEM 44: When the subject describes seeking help from others, there is a sense of having learned something from the interchange.",
  "match_justification": "The Seeker is actively engaging with the Supporter's shared experience about running to clear the mind. By stating 'i think i migh' (meaning they might start running), the Seeker demonstrates that they are taking away a potential new coping strategy and learning from the supportive interchange.",
  "sublevel_name": "Affiliation",
  "defense_level": 7
},
{
  "dmrs_q_item": "ITEM 97: Following experiences of emotional distress or conflict, the subject engages in sports or other physical activities which are an invigorating outlet for any lingering frustrations.",
  "match_justification": "In the context of feeling distressed and anxious about their friendships, the Seeker considers taking up running after hearing that it helps relax the mind and brush things off. This shows an inclination toward using physical activity as a healthy, sublimatory outlet for their current emotional frustrations.",
```

```

    "sublevel_name": "Sublimation",
    "defense_level": 7
  },
  {
    "dmrs_q_item": "ITEM 69: When confronting emotionally charged topics, the subject tends not to address concerns directly and fully but wanders off to tangentially related topics that are emotionally easier for the subject to discuss or prefers to pay attention to someone else dealing with a similar situation. This can include preferring to read or watch a film portraying people dealing with similar problems.",
    "match_justification": "The conversation has shifted from the highly distressing topic of the Seeker's failing friendships and financial insecurities to the safe, neutral topic of hobbies (running, playing the piano). The Seeker readily participates in this shift, focusing on these easier, tangential subjects rather than dwelling on the immediate emotional conflict.",
    "sublevel_name": "Displacement",
    "defense_level": 5
  }
}

```

- Sublimation (Level 7): Sublimation involves actively channeling distressing feelings into a socially acceptable, creative, or physical outlet. While running and playing the piano are classic examples of sublimatory activities, the Seeker is only considering running in the future ("i think i migh") and played the piano in the past ("when i was in highschool"). Because they are not currently engaging in the activity to process their immediate distress, Sublimation is not the best fit.

- Affiliation (Level 7): Affiliation involves turning to others for support and collaboratively finding solutions or enhancing coping skills. By listening to the Supporter's personal experience with running, validating it, and stating "i think i migh" [try it], the Seeker demonstrates that they are actively learning from the supportive interchange. They are taking away a tangible, adaptive coping strategy from the conversation, which perfectly aligns with the criteria for Affiliation (specifically ITEM 44: "there is a sense of having learned something from the interchange").

4. Verdict:

The Seeker is utilizing the supportive relationship to adopt a new, healthy coping strategy, making Affiliation the most accurate defense mechanism.

Final Selected Defense Level

7

B.4 System 6 Foundation LLM Re-ranking Trace

1. Context:

The Seeker has been discussing their distress over failing friendships and financial insecurities. The Supporter, recognizing that there is little more to be done about the friends at this moment, shifts the topic to hobbies as a way to cope. The Supporter mentions that they go on runs to clear their mind and asks what instrument the Seeker plays.

2. Function:

The Seeker responds to both parts of the Supporter's message. They validate the Supporter's coping strategy by mentioning their sister's similar experience with running, express an openness to trying it themselves ("i think i migh"), and answer the direct question about the instrument they play. The psychological function here is engaging with the support offered and considering a new coping mechanism.

3. Evaluation:

- Displacement (Level 5): Displacement involves the subject unconsciously shifting the focus from a distressing topic to a safer, tangential one to avoid emotional discomfort. However, in this dialogue, the Supporter initiated the topic shift to hobbies. The Seeker is simply following the conversational lead and answering direct questions. Therefore, the Seeker is not actively using displacement.

B.5 System 7 Supervised Reasoning Trace

Context Trigger: The Supporter has just concluded the discussion about the Seeker's distressing friendship issues, suggesting that the Seeker has done all they can and the ball is in the friends' court. The Supporter then asks about the Seeker's hobbies.

Psychological Goal: To avoid the lingering emotional pain, anxiety, and helplessness associated with the unresolved friendship conflict by readily shifting the conversational focus to a neutral, safe, and positive topic.

Handbook Alignment: The utterance aligns with Displacement (Level 5). The DMRS handbook defines Displacement as keeping unacceptable motives out of awareness, where feelings surface indirectly. Specifically, ITEM 69 states: 'When confronting emotionally charged topics, the subject tends not to address concerns directly and fully but wanders off to tangentially related topics that are emotionally easier for the subject to discuss...' The Seeker was just discussing a highly emotionally charged topic (feeling ignored and abandoned by friends due to financial issues). The Supporter then shifts the topic to hobbies. The Seeker readily engages in this new, emotionally easier topic, talking about their sister's running and their past piano playing, thereby displacing the focus from the distressing conflict to a safe, tangential subject.

Differential Diagnosis: This is not a Level 0 (Neutral Utterance) because, in the immediate context of a highly distressing and unresolved emotional conflict, the abrupt and enthusiastic shift to a completely unrelated topic serves a defensive function to escape the pain. It is not Level 7 (Affiliation or Sublimation) because the Seeker is not actively using the Supporter's advice to collaboratively solve the friendship problem (Affiliation), nor are they describing an active, established coping mechanism to channel their frustration (Sublimation); they are merely agreeing with the Supporter's hobby and sharing a casual, unrelated anecdote. It is not Level 6 (Isolation of Affect) because the Seeker is not discussing the distressing event in a detached manner; they have completely changed the subject to avoid the event altogether.

Defense Level: 5

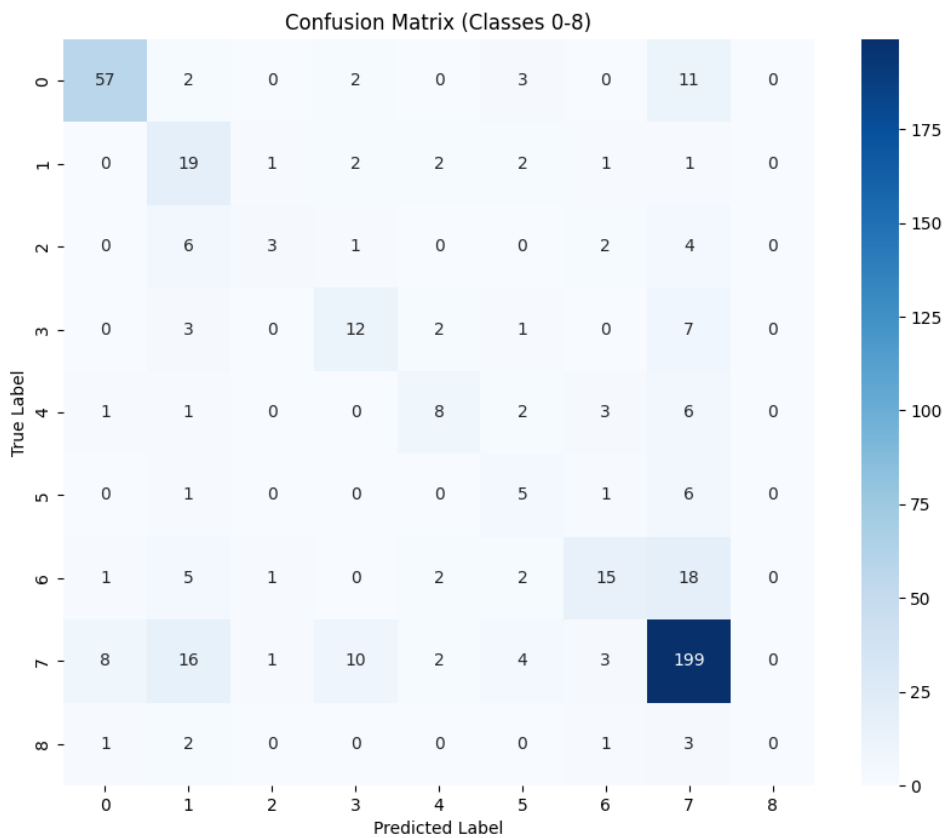


Figure 2: Confusion Matrix of the best performing system (Approach 7). The system demonstrates high accuracy on the most frequent classes, though it exhibits a slight prediction bias towards Level 7 when evaluating ambiguous mid-level defenses.

LinguITics at PsyDefDetect: Iterative Imbalance-Aware Fine-tuning of Qwen3-8B for Psychological Defense Mechanism Classification

Shefayat E Shams Adib*, Ahmed Alfey Sani*, Md Hasibur Rahman Alif*,
Ajwad Abrar*

Department of Computer Science and Engineering,
Islamic University of Technology, Dhaka, Bangladesh
{shefayatadib, ahmedalfey, hasiburrahman21, ajwadabbrar}@iut-dhaka.edu

*All authors contributed equally to this work.

Abstract

Detecting psychological defense mechanisms in conversational text remains a challenging clinical NLP problem. For the PsyDefDetect 2026 shared task (9-class utterance classification evaluated via macro F1), our team LinguITics¹ achieves a macro F1-score of 0.3917 on the official positive-class leaderboard, ranking 4th out of 21 registered teams and improving over the Ministral-8B task baseline (31.48 macro F1) by +7.7 absolute points (+24.4% relative). BERT-family encoders and zero-shot LLMs proved ineffective on rare classes due to severe class imbalance, leading us to QLoRA fine-tuning of Qwen3-8B. We leverage three key strategies: grouped stratified cross-validation (preventing leakage), minority-class round-robin lexical augmentation, and a post-processing pipeline with logit bias tuning and ensemble blending. Together, these components close much of the validation–leaderboard gap and substantially improve minority-class recall, driving the critical “Unclear” class (Level 8) from near-zero performance to $F1 = 0.797$.

1 Introduction

Automatic detection of psychological defense mechanisms (unconscious strategies to mitigate distress under the DMRS framework (Perry and Henry, 2004)) helps mental health platforms flag maladaptive coping and improves empathetic conversational agents (Liu et al., 2021; Na et al., 2025). The PsyDefDetect 2026 shared task (Na et al., 2026a,b) challenges participating systems to classify seeker utterances into nine DMRS levels. This poses a major obstacle, that is extreme class imbalance (He and Garcia, 2009), with the frequency gap between majority (“High-Adaptive”, 51.9%) to minority (“Unclear”, 1.5%) classes at about 34.6 times respectively. Because evaluation uses

¹Code and resources are available at <https://github.com/Shefwef/LinguITics-PsyDefDetect-BIONLP26>

macro-averaged F1, optimizing on accuracy leads to majority-class collapse and task failure.

To address this, we followed up with an iterative development process. Standard single-fold PEFT (parameter-efficient fine-tuning) on Qwen3-8B (Yang et al., 2025) suffered a massive generalization gap (0.34 validation vs. 0.24 leaderboard F1) due to limited low-rank capacity and majority-class overfitting. By systematically upgrading model capacity, loss functions, and inference, we established a robust pipeline. Our key contributions are:

1. A leakage-safe cross-validation scheme at the level of groups, where synthetic augmentations are in a set with their source utterances. This leads to an order of magnitude smaller generalization gap between out-of-fold and leaderboard batches.
2. An oversampling method that preserves the original psychological signal (Wei and Zou, 2019). This is achieved by expanding specific minority classes by 3 times in a round-robin lexical mutation approach.
3. A post-processing pipeline that combines OOF based logit bias tuning, that is guarded using v2 decoding, and multi-seed probability blending.

2 Task and Dataset

The PsyDefDetect 2026 task (Na et al., 2026a) classifies seeker utterances into nine psychological defense levels, as defined by the DMRS framework (Perry and Henry, 2004), and evaluated by macro averaged F1-score. The PsyDefConv dataset (Na et al., 2026b) contains 2,336 utterances across 200 dialogues from ESConv (Liu et al., 2021). For 5-fold CV, we merged train and validation sets (1,864 training examples and 472 test). It exhibits very

high class imbalance with a 34.6 times frequency gap (Table 1).

L	Defense Mechanism	N	%
0	No Defense / Neutral	296	15.9
1	Action Defenses	108	5.8
2	Major Image-Distorting	61	3.3
3	Disavowal	99	5.3
4	Minor Image-Distorting	84	4.5
5	Neurotic	48	2.6
6	Obsessional	172	9.2
7	High-Adaptive	968	51.9
8	Unclear / Needs More Info	28	1.5
Combined Train		1,864	100.0

Table 1: PsyDefConv combined training class distribution (Train + Val splits). Level 7 vs. Level 8 ratio: 34.6×.

3 Implementation Process

Our development went through three iterative stages, each revealing a fundamental limitation that directly inspired the next architectural transition. Table 10 follows the complete leaderboard path from R0 (F1 = 0.240) to our final submission (F1 = 0.392). The detailed system run log is provided in Table 2.

System / Variant	OOF F1	LB F1
R0: MentalBERT	N/A	0.240
R1: MentalBERT+RoBERTa ensemble	N/A	0.240
R2: MentalRoBERTa Focal + EMD	0.314	N/A
R3: DeBERTa-v3-base 5-fold	0.307	0.236
R4: RoBERTa-base	N/A	0.269
R5: Qwen3-8B 1-fold $r=64$ baseline	0.345	0.249
R6: Qwen3-8B 5-fold $r=128$ + weighted CE	0.361	0.329
R7: Qwen3-8B v2	0.372	0.355
R8: Qwen3-8B v2 microplus	0.372	0.354
R9: Qwen3-8B seed A only + v2 decode	0.431	N/A
R10: Qwen3-8B old + seed A blend + v2 decode	0.437	0.392

Table 2: Complete system run log. LB = CodaBench leaderboard.

3.1 Stage 1: BERT-Family Encoder Baselines

We evaluated MentalBERT, MentalRoBERTa, DeBERTa-v3-base, and RoBERTa-base (Devlin et al., 2019; Ji et al., 2022; He et al., 2021; Liu et al., 2019) for multiple context windows, loss functions (cross-entropy, Focal (Lin et al., 2017), EMD), and ensemble strategies (full scores in Appendix E, Table 10). Best validation macro F1 was 0.314 (MentalRoBERTa + Focal/EMD + Hungarian remapping), with leaderboard peak at 0.240. Importantly, F1 for Classes 3, 5 and 8 was still zero across *all* variants, establishing an *encoder*

capacity ceiling with a 51.8% majority-class prior using $n \leq 50$ minority examples. From the general formula:

$$\mathcal{L}_{\text{Focal}} = -(1 - p_t)^\gamma \log p_t \quad (1)$$

where p_t is the probability of predicting the correct class, logarithmically compensated by strength constant γ .

3.2 Stage 2: Zero-Shot Evaluation

Qwen3-8B, Llama 3.1-8B, and Ministral-8B evaluated zero-shot with explicit DMRS label definitions produced 8–16% macro F1 (Table 10), confirming that task knowledge cannot be prompt-engineered.

3.3 Stage 3: Diagnostic LLM Fine-Tuning

Ministral-8B fine-tuned with 4-bit NF4 quantization achieved 64.71% accuracy but only 14.74 macro F1 (Table 10). This illustrates that standard cross-entropy collapses to majority-class prediction under severe imbalance. Furthermore, the accuracy is an actively misleading metric in this setting.

3.4 Stage 4: Final System (Qwen3-8B LoRA Pipeline)

As such, with the three lessons above guiding us, our final pipeline consists of five components: model architecture, imbalance-aware training objective, data augmentation, leakage-safe cross-validation and a post-processing ensemble, each targeting a specific failure mode identified in the earlier stages.

3.4.1 Model Architecture

We fine-tune Qwen3-8B via QLoRA (Detmeters et al., 2023) with 4-bit NF4 quantization, reducing peak GPU memory from ~ 32 GB to ~ 8 GB on a single NVIDIA RTX 3090 Ti. LoRA adapters (Hu et al., 2021) target all attention and feed-forward layers (q, k, v, o, gate, up, down) and the score head ($r = 128$, $\alpha = 256$, dropout = 0.1), yielding ≈ 31 M trainable parameters (0.4% of the 8B base). Increasing rank from $r = 64$ to $r = 128$ delivered a +24.9% fold-level F1 gain, critical for separating psychologically similar classes (e.g., Level 4 vs. Level 5). Full training hyperparameters for this configuration are summarised in Table 8 in Appendix B.

3.4.2 Input Representation

Each prompt has three parts: (1) the DMRS Label Guide with 9-class clinical schema; (2) Conversational Context of the last 30 dialogue turns

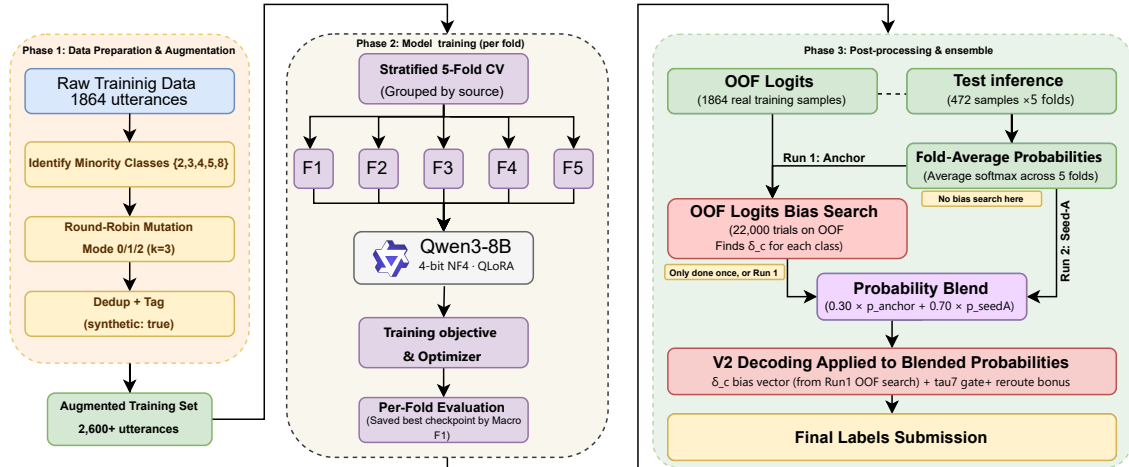


Figure 1: Full system pipeline. **Phase 1:** raw training data is preprocessed and minority classes (Levels 2, 3, 4, 5, 8) are oversampled via round-robin lexical mutation, creating an expanded training set of 2,600+ utterances. **Phase 2:** two independent grouped stratified 5-fold QLoRA fine-tuning runs of Qwen3-8B the *Anchor* (seed = 42) and *Seed-A* (seed = 20260407) sharing identical architecture and hyperparameters, with dialogues grouped across folds to prevent data leakage. **Phase 3:** a class-specific logit bias vector (δ_c) is optimised on Anchor OOF predictions and locked for reuse; test probabilities from both runs are blended before a final guarded decode step that maximises minority-class recall without sacrificing majority-class precision.

prefixed by SEEKER:/HELPER: tags and (3) an Output Instruction directing the model to emit a single integer (0–8). Inputs are tokenised to 1,024 tokens with dynamic padding to multiples of 8, covering >95% of samples without truncation.

3.4.3 Imbalance-Aware Training Objective

This dataset is, to a large extent, imbalanced towards Level 7. To mitigate the problem of majority-class collapse, we apply two stabilization techniques.

Inverse-Square-Root Class Weighting. Per-class weights

$$w_c = \frac{1/\sqrt{n_c}}{\sum_{i=0}^8 1/\sqrt{n_i}} \quad (2)$$

boost the most under-represented classes (e.g., $w_8 = 1.67$, $w_5 = 1.29$) while dampening the majority class ($w_7 = 0.28$), instead of inverse-frequency weighting that can lead to gradient instabilities.

Label Smoothing & Optimization Schedule. By using label smoothing (Szegedy et al., 2016) ($\epsilon = 0.05$), we avoid early logit saturation for Level 7. This prevents gradients from being dominated by the majority class and allows minority-class signals to contribute more effectively during back-propagation. We employ AdamW with a peak learning rate of 1.2×10^{-4} , cosine annealing (8% warmup), 10 epochs per fold, batch size 16 (2×8

accumulation), and gradient clipping at 0.3 (see Appendix B).

3.4.4 Data Augmentation

For the rarest classes (Levels 2, 3, 4, 5, and 8), there are between 28 and 84 training examples. This is not enough for an 8B capacity model to learn sufficiently reliable decision boundaries. In order to fix this, we perform **round-robin lexical mutation** to generate $k = 3$ synthetic variants per source utterance in these classes, cycling through three surface-level rewriting modes:

- **Mode 0:** Contraction replacement (e.g., *I am* → *I'm*) plus a hedging prefix (*Honestly, ...*).
- **Mode 1:** Vocabulary style-shift (e.g., *maybe* → *perhaps*) plus a trailing filler (*... I guess.*).
- **Mode 2:** Hesitation markers (e.g., periods → ellipses; ? → ??).

Mutations target only the seeker utterance to preserve the psychological signal; after deduplication, minority class counts increased from 28–84 to 65–252 examples (see Appendix C). This targeted minority-class oversampling strategy is consistent with prior findings that augmenting only underrepresented classes yields more effective and stable performance improvements than augmenting all classes equally (Sani et al., 2026).

3.4.5 Grouped Five-Fold Cross-Validation

Random splitting risks leakage across the 200 source dialogues, making dialogue-level isolation essential. We therefore apply grouped stratified cross-validation (grouped CV, implemented as StratifiedGroupKFold with $k = 5$) using dialogue_id as the grouping key:

- **Zero Leakage Guarantee** (0 leaked dialogues confirmed): All utterances and their synthetic variants are kept entirely within one fold.
- **Reliable Validation Signal** (OOF leaderboard gap reduced from 9.6 to 1.7-4.5 points): Strong rank-correlation between OOF and leaderboard gains enables safe threshold tuning.
- **Ensemble Foundation** (5 checkpoints per seed): Five-fold training yields pure OOF predictions for post-processing calibration and reduces inference variance.

Reliable validation behaviour across folds is shown in the per-fold OOF metrics in Table 3.

Fold	Acc.	Mac. F1	Mac. P	Mac. R
1	0.6193	0.3804	0.3836	0.3925
2	0.6247	0.3701	0.3902	0.3617
3	0.6408	0.3899	0.4276	0.4000
4	0.6300	0.3553	0.3753	0.3514
5	0.5968	0.3326	0.3396	0.3298
OOF	0.6223	0.3716	0.3817	0.3675

Table 3: Per-fold CV results (grouped-clean augmented run, before seed blending).

3.4.6 Post-Processing and Ensemble Strategy

Despite class-weighted training, raw probabilities remain heavily biased towards the majority class (Level 7). To rectify this and recover rare classes without compromising precision, we implement a three-stage post-training pipeline (v2decode).

Stage A: OOF Bias Optimization. Using logit adjustment for long-tail learning (Menon et al., 2021), we search for class-specific probability offsets (δ_c) that maximize the OOF macro F1 score:

$$\hat{y} = \arg \max_c [\log p_c + \delta_c] \quad (3)$$

We evaluate approximately 22,000 randomly sampled bias vectors on OOF predictions to identify a configuration that balances majority precision with minority recall. The best locked vector applies a

negative penalty to Level 7 ($\delta_7 < 0$) and substantial positive bonuses to minority classes like Level 8 ($\delta_8 > 0$).

Stage B: Multi-Seed Blending. We run a second identical 5-fold training pipeline, denoted *Seed-A*, using a different random seed (seed = 20260407) and the same architecture and hyperparameters. We combine the test-set probabilities of the original Anchor and Seed-A using a 30/70 weighted average:

$$p_{\text{blend}} = 0.30 \cdot p_{\text{anchor}} + 0.70 \cdot p_{\text{seedA}} \quad (4)$$

The ratio was tuned using real-only OOF F1, combining the Anchor’s high precision with Seed-A’s strong minority recall.

Stage C: The τ_7 -Gate Decoding. A confidence safeguard is applied from the locked δ_c bias vector onto p_{blend} to prevent precision collapse:

- **τ_7 -Protection Gate:** The prediction is locked to Level 7 and δ_c offsets are not applied if $p_{\text{blend},7} \geq 0.69$.
- **Minority Rerouting:** If $p_{\text{blend},7} < 0.69$, δ_c offsets are applied, rerouting ambiguous samples into the highest-probability minority class.

So, minority labels are aggressively recovered when the model is uncertain. This increases the minority recall without affecting the precision.

4 Experiments and Results

4.1 Cross-Paradigm Comparison

BERT-family encoders struggled to break above 0.314 macro F1 due to capacity limits, across three paradigms (Table 10 in Appendix E). Both zero-shot LLMs and standard LLM fine-tuning collapsed to majority-class predictions (near 15% F1). In contrast, our imbalance-aware Qwen3-8B pipeline resolved these issues, reaching 39.17% macro F1.

4.2 Comparison with SOTA Baselines

Table 4 compares the results of our systems against the task baselines (Na et al., 2026a,b). Our final pipeline surpassed the stated SOTA, Ministral-8B fine-tuned baseline (31.48 macro F1) +7.7 absolute points, corresponding to a +24.4% relative improvement in macro F1.

System	Acc. (%)	Macro F1 (%)
GPT-5 zero-shot (task paper)	52.75	19.53
Gemini 2.5 Pro zero-shot	56.36	25.99
DeepSeek-V3.2 zero-shot (CoT)	55.72	26.17
Llama 3.1-8B fine-tuned	62.92	30.51
InternLM3-8B fine-tuned	63.98	30.53
Ministral-8B fine-tuned (SOTA)	64.83	31.48
DeBERTa-v3-base (5-fold)	59.11	23.58
RoBERTa-base	51.27	26.97
Qwen3-8B LoRA (Baseline Finetuned)	54.45	24.91
Qwen3-8B LoRA (Grouped CV + Bias Tuning)	58.43	35.48
Qwen3-8B LoRA (SeedA Ensemble + v2decode)	64.19	39.17

Table 4: Comparison with task paper baselines.

4.3 Ablation Study

Table 5 analyses each component’s contribution. Increasing LoRA rank to $r = 128$ produced the highest boost (+24.9% fold-level F1), supporting

System Configuration	Macro F1
R0: 1-fold, $r=64$, no weighting	0.249
+ 5-fold CV, $r=128$	0.284 [†]
+ Weighted CE + label smoothing	0.329 [†]
+ Grouped-clean 5-fold	0.355
+ Data augmentation (RR-k3)	0.355
+ Seed-A blend (30/70) + v2 decode	0.392

Table 5: Ablation: each component’s contribution. [†]Metrics for these rows are single-fold estimates from the 5-fold setup, included as indicative rather than full OOF results.

that model capacity was indeed the primary bottleneck. Grouped CV, data augmentation, and post-processing decode rules contributed incrementally, securing the final +3.69 F1 points.

4.4 Per-Class Analysis

Per-class performance is shown in Figure 2 (Appendix D) and Table 6. Level 8 (“Unclear”) saw the most improvement, climbing from near-zero recall to 0.797 F1 via augmentation and bias tuning, with Levels 2 and 3 also gaining substantially. On the other hand, Levels 4 and 5 continue to be difficult (0.25–0.27 F1) due to high linguistic overlap with the majority class. Importantly, optimising for minority classes did not compromise the majority class (Level 7), which still resulted in a solid F1 of 0.709.

5 Conclusion

We showed that the data-centric imbalance mitigation methods (grouped CV, weighted loss, round-robin lexical augmentation, and dynamic OOF bias tuning with ensembling) that we used were much

L	Mechanism	P	R	F1
0	Neutral	0.747	0.858	0.799
1	Action	0.242	0.398	0.301
2	Major Img-D	0.480	0.451	0.465
3	Disavowal	0.401	0.402	0.401
4	Minor Img-D	0.317	0.211	0.254
5	Neurotic	0.398	0.214	0.278
6	Obsessional	0.203	0.267	0.231
7	High-Adaptive	0.693	0.726	0.709
8	Unclear	0.797	0.797	0.797
Macro		0.431	0.436	0.426

Table 6: Per-label OOF metrics, final blended system with v2 decode. Level 8 improved from ≈ 0 to 0.797 via augmentation and bias tuning.

more important than raw model capacity for psychological defense classification. We achieved a macro F1-score of 0.3917 on the official positive-class leaderboard, ranking 4th out of 21 registered teams. This corresponds to a +7.7 macro F1-score improvement (+24.4% relative) over the Ministral-8B fine-tuned baseline. In future, we plan to add more effective paraphrase-based data augmentation, use losses better suited to imbalanced classes, and evaluate on more datasets.

Limitations

These decode rules and OOF bias vectors are calibrated specifically to this dataset and so requires recomputation for unseen domains. The grouped CV protocol keeps mutant variants within the same source group to reduce leakage, but the risk of leakage cannot be fully eliminated. Lastly, we were limited to PEFT on models with 8B parameters or less and by hardware constraints (24 GB VRAM).

Acknowledgments

We thank the PsyDefDetect 2026 shared task organizers (Na et al., 2026a) for providing the PsyDefConv dataset (Na et al., 2026b) and evaluation infrastructure through CodaBench.

References

- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). *arXiv preprint arXiv:2305.14314*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

- Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Haibo He and Edwardo A. Garcia. 2009. [Learning from imbalanced data](#). *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: Decoding-enhanced BERT with disentangled attention](#). In *International Conference on Learning Representations*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *arXiv preprint arXiv:2106.09685*.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. [MentalBERT: Publicly available pretrained language models for mental healthcare](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. [Focal loss for dense object detection](#). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, Venice, Italy.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *Preprint*, arXiv:1907.11692.
- Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2021. [Long-tail learning via logit adjustment](#). In *International Conference on Learning Representations*.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.
- John Perry and Melissa Henry. 2004. [Studying defense mechanisms in psychotherapy using the defense mechanism rating scales](#). *Advances in Psychology*, 136.
- Ahmed Alfey Sani, Kazi Akib Zaoad, Shefayat E Shams Adib, Md Abdul Muqtadir, and Ajwad Abrar. 2026. [Addressing data scarcity in bangla fake news detection: An llm-based dataset augmentation approach](#). *arXiv preprint arXiv:2605.01292*.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. [Rethinking the inception architecture for computer vision](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, Nevada.
- Jason Wei and Kai Zou. 2019. [EDA: Easy data augmentation techniques for boosting performance on text classification tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

A Full DMRS Label Definitions

Table 7 provides the complete clinical descriptions for all nine psychological defense levels used in our classification prompt.

B Full Hyperparameter Table

Hyperparameter	Value
Base model	Qwen/Qwen3-8B
Quantization	4-bit NF4 + double quant
LoRA rank / alpha	128 / 256
LoRA dropout	0.1
LoRA target modules	q/k/v/o/gate/up/down/score
Trainable parameters	$\approx 31\text{M}$ (0.4%)
Max sequence length	1024
Optimizer	AdamW
Learning rate	1.2×10^{-4}
Weight decay	0.01
LR scheduler	Cosine annealing
Warmup ratio	8%
Per-device batch size	2
Gradient accumulation	8 (eff. batch = 16)
Gradient clip norm	0.3
Epochs per fold	10
Label smoothing ε	0.05
Class weight formula	Inverse-sqrt (Eq. 2)
Hardware	NVIDIA RTX 3090 Ti 24 GB
Mixed precision	bf16

Table 8: Complete training hyperparameters for the final Qwen3-8B LoRA system.

C Data Augmentation Examples

Below are three mutation modes applied to a single Level-3 (Disavowal) sample. All mutations target only the seeker utterance; the supporting dialogue context is unchanged.

D Confusion Matrices

Figure 2 illustrates error distributions across minority and majority classes, highlighting grouping and filtering improvements.

E Comprehensive Model Comparison

Table 10 reports all systems evaluated during our development, organised by model family and experimental stage. Empty cells indicate the model was not evaluated under that paradigm. Rows marked † are external baselines from the task paper (Na et al., 2026b); all others are our own internal tuning experiments.

L	Name	Clinical Description
0	No Defense	Phatic or factual exchange with no active defense mechanism
1	Action	Emotional discharge through behavior (passive aggression, complaining, impulsive action)
2	Major Img-D	Extreme cognitive distortion: projection, splitting, all-or-nothing thinking
3	Disavowal	Avoiding unpleasant reality: denial, rationalization, minimization
4	Minor Img-D	Subtler distortion: devaluation, idealization, omnipotence
5	Neurotic	Unconscious anxiety management: repression, displacement, reaction formation
6	Obsessional	Over-intellectualization, isolation of affect, undoing
7	High-Adaptive	Mature coping: humor, altruism, insight, self-assertion, sublimation
8	Unclear	Insufficient context for reliable DMRS classification

Table 7: DMRS label definitions used in the classification prompt.

Mode	Utterance Text
Original (Level 3)	<i>“It is not really that bad honestly, I have been through worse situations before.”</i>
Mode 0 (Contractions + Hedging)	<i>“Honestly, it’s not really that bad, I’ve been through worse situations before.”</i>
Mode 1 (Style Shift + Filler)	<i>“It is not quite that bad, I have been through worse situations before I guess.”</i>
Mode 2 (Hesitation)	<i>“It is not really that bad honestly... I have been through worse situations before...”</i>

Table 9: Example of round-robin lexical mutations for a Disavowal (Level-3) seeker utterance. The core signal (minimization (“not that bad”) and historical comparison (“been through worse”)) is preserved across all mutations, ensuring label validity.

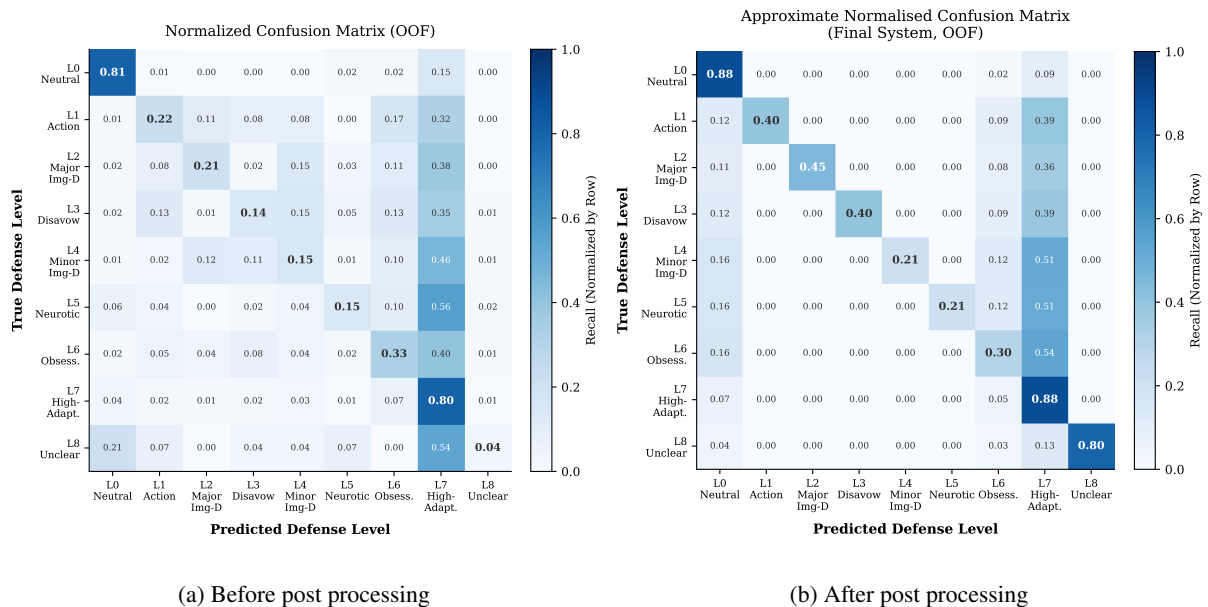


Figure 2: Approximate row-normalised confusion matrices (OOF). Post-processing successfully shifts residual majority-class prediction bias away from L7, noticeably improving minority recall along the diagonal.

Model	BERT-family (Supervised)				Zero-Shot LLM Inference				LLM Fine-Tuning (LoRA)			
	Acc	macro F1	Input / Loss	Ep	Acc	Macro Prec	Macro Recall	macro F1 (1-8)	Acc	Precision	Recall	F1
MentalBERT-base	n/a	0.240 ✓	Flat SEP $k=5$, Wtd CE, LR $2e-5$	5								
MentalRoBERTa-base	n/a	0.2200	Flat SEP $k=10$, Wtd CE, LR $1e-5$	10								
MentalBERT+RoBERTa (ens.)	n/a	0.240 ✓	Flat SEP $k=10$, Wtd CE, LR $2e/1e-5$	5/8								
DeBERTa-v3-base (5-fold)	0.591	0.2358	[CTX]/[TGT], Wtd CE, LR $1.5e-5$	8								
RoBERTa-base	0.513	0.2697	[CTX]/[TGT], Wtd CE, LR $2e-5$	6								
RoBERTa-base (len=320)	0.479	0.2763	[CTX]/[TGT], Wtd CE, LR $1.5e-5$	6								
RoBERTa-base + OS (unif.)	0.618	0.2893	[CTX]/[TGT], Std CE + OS(120), LR $2e-5$	6								
RoBERTa-base + OS (tgt.)	0.614	0.2430	[CTX]/[TGT], Std CE + tgt. OS, LR $2e-5$	8								
Qwen3-8B					0.3422	0.1773	0.1670	0.1536				
Llama 3.1-8B					0.3036	0.1958	0.1999	0.1584				
Ministral-8B					0.2406	0.2517	0.1481	0.0841				
Mistral-7B-v0.3 (Kaggle)									0.1444	n/a	n/a	0.1023 (0.2054 [†])
Ministral-8B (Local)									0.6471	n/a	n/a	0.1474
Qwen3-8B ★									0.6419	0.4003	0.3958	0.3917 ★

Table 10: Consolidated results across all model families. Empty cells indicate the model was not evaluated under that paradigm. ✓ denotes selected ensemble members. ★ denotes the best result. Rows marked † are external baselines from the task paper (Na et al., 2026a); all others are our own internal tuning experiments.

TONI-NLP at PsyDefDetect: Defense Mechanism Detection via LLM-based Ensemble Methods

Durjoy C. Paul¹, Callum Chan², Arshitha Basavaraj³, Veronica Perez-Rosas¹
Diana Inkpen², Francisco Pereira⁴, Juan Antonio Lossio-Ventura⁴

¹Texas State University, USA, ²University of Ottawa, Canada,

³International Institute of Information Technology, Bangalore, India,

⁴National Institute of Mental Health, National Institutes of Health, USA

{xei29, vperezr}@txstate.edu, {cchan073, diana.inkpen}@uottawa.ca,

arshitha.basavaraj@iiitb.ac.in, {francisco.pereira, juan.lossio}@nih.gov

Abstract

This system paper presents the approach of Team TONI-NLP to the PsyDefDetect 2026 shared task. The objective of the task was to classify utterances from helper–seeker conversations into nine categories: seven labels representing progressively higher levels of defensive maturity, one label indicating the absence of a defense mechanism, and one label for cases requiring additional information. We investigated several modern NLP approaches, including prompt engineering, fine-tuning, hierarchical modeling and classification using text embeddings derived from transformer-based models as well as classical embeddings such as TF-IDF. Our results show that ensemble methods performed best among our submitted systems, achieving a macro-F1 score of 0.320 and ranking 9th in the shared task out of 21 teams.

1 Introduction

The field of Natural Language Processing (NLP) has seen a surge in research dedicated to extracting mental health insights from textual data. While most of these efforts have leveraged large-scale social media data for psychological analysis (Garg, 2023; Skaik and Inkpen, 2020), recent work has begun to shift the focus to conversational data, particularly in the context of emotional and mental health support dialogues (Na et al., 2025).

This paper describes our team’s participation in the PsyDefDetect 2026 shared task at BioNLP 2026. This task aims to advance the understanding of defense mechanisms expressed in emotional support dialogues. To address the complexities of identifying these mechanisms, we propose a multifaceted approach utilizing Large Language Models (LLMs) and other traditional NLP methods.

2 Shared Task Description

The PsyDefDetect 2026 shared task focuses on identifying psychological defense mechanisms

from multi-turn emotional support dialogues (Na et al., 2026a), involving two participants: a help-seeker (the person sharing their difficulties) and a supporter (also called helper). The task is grounded in the Defense Mechanism Rating Scales (DMRS) (Perry and Henry, 2004) and uses the PSYDEFCONV dataset of 200 seeker–supporter dialogues (Na et al., 2026b).

Dataset. For this task, 1,864 help-seeker turns were provided as training and 472 as test sets, totaling 2,336 turns. Hereafter, we use *utterance* to refer to a help-seeker turn. The defense labels include: No Defense, Action, Major Image-Distorting, Disavowal, Minor Image-Distorting, Neurotic, Obsessional, High-Adaptive, and Needs More Information. The training set is highly imbalanced and skewed towards the *High-Adaptive* label.

Train/Validation Split. Since no official validation set was provided, we partitioned the 1,864 training samples into training and validation sets at the conversation level to avoid data leakage: all turns from a given dialogue were assigned exclusively to one split.

From 200 unique dialogues, we selected 20 conversations (10%) for validation using a greedy scoring strategy that maximizes label diversity, approximates the global class distribution, and penalizes longer conversations. This yields 1,592 training samples (180 conversations) and 272 validation samples (20 conversations), an approximately 85/15 split. Table 1 reports the per-label distribution across the training and validation splits. Label proportions in the validation set closely mirror those in the full dataset, with validation percentages ranging from 11% (label 7, *High-Adaptive*) to 29% (label 8, *needs more information*), reflecting the inherent difficulty of balancing rare classes under conversation-level constraints. No dialogue leakage was detected between splits. Figure 1 provides a visual summary of these distributions.

Label	Defense Type	Train	Val	Total
0	No Defense	248 (83.8%)	48 (16.2%)	296
1	Action	89 (82.4%)	19 (17.6%)	108
2	Major Image-Distorting	49 (80.3%)	12 (19.7%)	61
3	Disavowal	77 (77.8%)	22 (22.2%)	99
4	Minor Image-Distorting	66 (78.6%)	18 (21.4%)	84
5	Neurotic	37 (77.1%)	11 (22.9%)	48
6	Obsessional	145 (84.3%)	27 (15.7%)	172
7	High-Adaptive	861 (89.0%)	107 (11.0%)	968
8	Needs More Information	20 (71.4%)	8 (28.6%)	28
Total		1,592	272	1,864

Table 1: Per-label sample counts and percentages across the training and validation splits. Percentages indicate the proportion of each label’s total samples assigned to each split.

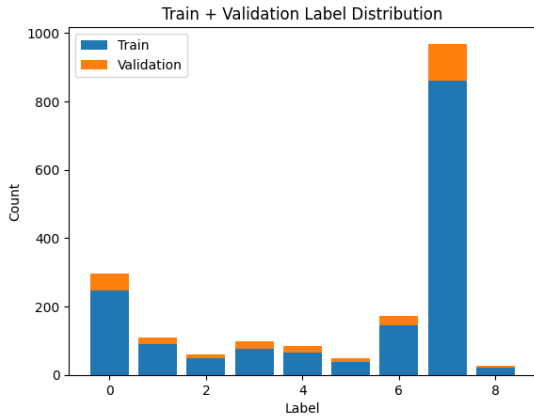


Figure 1: Label distribution across training and validation splits. The validation set closely mirrors the overall distribution, though the strong class imbalance is evident, particularly for the High-Adaptive class (label 7), which accounts for over half of all samples.

3 Methodology

We investigated four approaches for defense mechanism classification: prompting, hierarchical classification, supervised fine-tuning, and ensemble methods. These approaches were evaluated within a unified experimental setup using training and validation splits derived from the training data, as described in the previous section, enabling systematic comparison across methods.

3.1 Prompting

We conducted zero-shot and few-shot prompting experiments with instruction-tuned LLMs, following the prompt structure of Chan et al. (2025) and Lossio-Ventura et al. (2025). We evaluated prompt optimization, random few-shot prompting, and dynamic example retrieval across three model families: Llama (Llama-3.1-8B-Instruct, Llama-3.3-70B-Instruct), Claude (claude-opus-4-6 and claude-

sonnet-4-6), and GPT (gpt-5.2).

3.1.1 Zero-shot experiments

We started with the zero-shot prompt provided by the task organizers and iteratively refined it based on classification performance. This process, informed by the DMRS annotation guidelines (Di Giuseppe and Perry, 2021), involved adding additional context, such as detailed label definitions, along with keysigns, workflow guidelines and disambiguation rules to capture nuanced distinctions between classes. All models were prompted to respond with a single digit (0–8) corresponding to the predicted defense level.

3.1.2 Few-shot experiments

We applied the context strategies described above to few-shot prompting and included labeled examples in the prompt. We experimented with few-shot prompting using randomly sampled examples (random few-shot) and also with examples selected based on semantic similarity (dynamic few-shot). For dynamic few-shot, we retrieved examples using cosine similarity over sentence embeddings, experimenting with $k \in \{5, 10, 20, 25\}$ examples per class to assess the effect of example coverage on classification performance.

Implementation details. All models received the same prompt structure, consisting of a task description, the DMRS annotation handbook, the label definitions, the full dialogue context, and the target utterance. We experimented with both the full and condensed handbook versions. Since Llama and Claude cap temperature at 1.0 while GPT allows up to 2.0, we sampled $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ and $\{0.4, 0.8, 1.2, 1.6, 2.0\}$ respectively, covering each model’s temperature range. All models were queried with $\text{top}_p=0.7$.

3.2 Hierarchical Approach

We also performed experiments with a hierarchical classification approach inspired by the Defense Mechanism Rating Scales (DMRS) hierarchical organization of defensive categories (Di Giuseppe and Perry, 2021). Initially, we distinguished between four aggregated defense categories: Mature Defenses, Neurotic Defenses, Immature Defenses, and No Defense. We then performed a secondary categorization step using the fine-grained subcategories shown in Table 2. We used a Multinomial Logistic Regression (MLR) classifier with TF-IDF features combined with contextualized embeddings.

These embeddings are obtained using the simplified SBERT-WK (Wang and Kuo, 2020) pooling strategy to MentalBERT (Ji et al., 2022) by averaging the final four hidden layers.

Aggregated Category	Specific Defense Label
No Defense	0 – No Defense
	8 – Needs More Information
Mature Defenses	7 – High-Adaptive
Neurotic Defenses	5 – Neurotic
	6 – Obsessional
Immature Defenses	1 – Action
	2 – Major Image-Distorting
	3 – Disavowal
	4 – Minor Image-Distorting

Table 2: Mapping between the four aggregated categories and their corresponding specific defense labels.

Similar to our prompting methods, we experimented with context augmentation strategies. We conducted two experiments: (1) using the target utterance only, and (2) using the target utterance plus context, consisting of the previous conversational turn¹. For comparison, we also ran a 9-way classification task for all categories using the target utterance as the sole input.

3.3 Fine-tuning LLMs

We explored several fine-tuning strategies for the classification task, ranging from encoder-only discriminative models to decoder-only generative LLMs, with varying approaches to class imbalance mitigation, input representation, and parameter efficiency. We organized our fine-tuning experiments into three groups: (i) encoder-based discriminative models, (ii) decoder-only LLMs with a classification head, and (iii) generative LLMs trained in an instruction-following format. Table 3 summarizes the key design choices across all approaches.

3.3.1 Encoder-based Discriminative Models

We fine-tuned a set of encoder-only transformer models such as BERT (base and large), RoBERTa, and domain-specific mental variants, given the clinical nature of the task, such as MentalBERT and MentalRoBERTa (Ji et al., 2022) for direct sequence classification. All models operated on the *target utterance* only, with a maximum sequence length of 512 tokens. A grid search over learning rates

¹We experimented with other context sizes but did not observe a significant improvement

$\{2e-5, 3e-5, 4e-5, 5e-5, 4e-4\}$, batch sizes $\{8, 16, 32\}$, and weight decay $\{0.01, 0.05, 0.1\}$ was conducted over 10 epochs. To address class imbalance, we explored two strategies:

(1) *Loss-based balancing*: inverse-frequency class weights were applied to the cross-entropy loss.

(2) *Hybrid balancing (data + loss)*: majority classes were undersampled (e.g., label 7 capped at 200 samples, label 0 at 150), and class weights recomputed on the balanced dataset.

3.3.2 Decoder LLMs with Classification Head

We adapted causal LLMs for classification by attaching a linear head to the final token representation of a LoRA-adapted model. We evaluated several configurations differing in parameter efficiency, input context, and training strategy:

(1) *QLoRA (8B, 70B)*: 4-bit quantized fine-tuning (Detmers et al., 2023) with LoRA ($r = 16$, $\alpha = 32$) applied to attention projections. The input consisted of the target utterance only (512 tokens). The optimization used cosine scheduling with warmup and `paged_adamw_8bit`.

(2) *LoRA with dialogue context (8B)*: full-precision BFloat16 training with LoRA applied to all projection layers. We evaluated configurations where ($r \in 8, 16, 32, 64$) and (α) is set to ($2r$). Inputs included full dialogue plus target utterance (up to 1,400 tokens).

(3) *LoRA with 5-fold cross-validation*: stratified folds were built on the initial training data, with class weights recomputed per fold. The configurations were ($r \in 8, 16, 32, 64$) and ($\alpha = 2r$).

(4) *One-vs-rest (OvR)*: nine binary classifiers were trained with LoRA, combining class-weighted loss and minority-class oversampling. Predictions were aggregated via softmax renormalization.

3.3.3 Generative Instruction-Tuned LLMs

We reformulate the task as instruction-based generation, where the model predicts the label digit as the output token. The loss was computed only over the generated label token, with prompt tokens masked. We explored two fine-tuning strategies:

(1) *Full fine-tuning (8B)*: using the llama-cookbook FSDP framework. Minority classes were oversampled to a minimum of 300 samples, yielding approximately 2,700 training instances. Training used learning rates of $\{5e-5, 4e-4\}$ for 10 epochs, with the checkpoint achieving the lowest validation loss selected for evaluation.

(2) *LoRA fine-tuning (70B)*: using parameter-

Group	Approach	Context	Imbalance Handling
Encoder	Weighted loss + Undersampling	Utterance Utterance	Loss-based (class weights) Data + Loss (undersampling + weights)
Decoder + Head	QLoRA (8B, 70B)	Utterance	Loss-based (class weights)
	LoRA (8B)	Utterance / Dialogue	Loss-based (class weights)
	LoRA 5-Fold (8B)	Utterance / Dialogue	Loss-based (class weights)
	OvR (8B)	Utterance / Dialogue	OvR + Oversampling
Generative	Full FT (8B)	Dialogue	Oversampling
	LoRA (70B)	Dialogue	Oversampling

Table 3: Summary of fine-tuning approaches. “Context” indicates whether only the target utterance or the full dialogue was used as input. “Imbalance Handling” distinguishes between loss-based, data-level, and hybrid strategies. Note that the encoder group was based only on BERT/roBERTa families. **8B** refers to Llama-3.1-8B-Instruct and **70B** refers to Llama-3.3-70B-Instruct.

efficient adaptation with LoRA ($r = 16, \alpha = 32$). Training used learning rates of $\{5e-5, 4e-4\}$. Despite scalability, validation loss analysis showed early overfitting, with the optimal checkpoint occurring at epoch 1.

3.4 Ensemble Methods

Our final approach used ensemble methods to combine predictions from multiple models. Ensemble methods improve performance by aggregating models, helping generalization to new data (Dietterich, 2000) and have proven effective across different NLP tasks (Zhang and Shafiq, 2024). After official results were released, we conducted post-hoc ensemble analyses over our submitted systems. We compared majority voting with weighted averaging using post-release system-level macro F1 scores. These analyses were not used to tune or select final submissions or guide model selection.

We created two different ensembles. The first (*Across approaches*) combined our best submissions from prompting, hierarchical classification, and fine-tuning. The second (*Best submissions*) combined our three strongest systems: (*Decoder*) *LoRA (8B) - utterance*, (*Decoder*) *LoRA (8B) - dialogue*, and (*Generative*) *Full FT (8B)*. For majority voting, ties were resolved using the Decoder LoRA (8B) utterance model, our best individual system. For weighted averaging, each system was weighted by its post-hoc macro F1 score-based performance.

4 Results

Table 4 presents results on the test data. The table reports prompt-engineering results, followed by hierarchical approaches, fine-tuned LLMs, and ensembles. We report macro-averaged precision, recall, and F1, following the shared-task metric.

For prompting methods, dynamic few-shot prompting achieves the best macro F1 (0.197), followed by zero-shot prompting (0.193), while random few-shot prompting performs worse (0.078). The best configuration uses $k = 20$ examples per class retrieved via cosine similarity, with performance decreasing for both $k = 25$ and $k = 5$. For hierarchical methods, the best performance is obtained using a two-step configuration restricted to the utterance (macro F1 = 0.238). This setup outperforms both the single-step baseline (0.236) and the two-step variant using the previous conversational turn (0.217). For fine-tuned models, the decoder-only LoRA (8B) with utterance-only input gets the highest macro F1 (0.303), slightly exceeding the dialogue-context variant (0.300) and outperforming encoder-based models (MentalBERT: 0.252; MentalRoBERTa: 0.263). Overall, ensemble methods achieve the best performance. The best submission (voting) reaches a macro F1 of 0.320 by combining our three strongest systems, outperforming the across-approaches voting ensemble (0.300). Averaging-based ensembles perform poorly (0.212 and 0.105), suggesting that majority voting is more robust for this task. These results place our system 9th out of 21 teams on the leaderboard.

5 Discussion

In our prompting experiments, dynamic few-shot prompting using the top 20 most similar examples outperformed structured zero-shot prompting with the annotation handbook as additional context. In contrast, randomly selected few-shot examples performed worse than both approaches, likely due to poorer class coverage and lower example relevance. Performance also varied with the number of retrieved examples and the embedding model,

Approach	Method	Accuracy	Precision	Recall	Macro F1
Prompting	Zero-shot (Baseline)	0.364	0.500	0.134	<u>0.193</u>
	Random Few-shot	0.277	0.131	0.145	0.078
	Dynamic Few-shot	0.511	0.198	0.219	0.197
Hierarchical	Two-step (Utterance + Previous Turn)	0.532	0.213	0.224	0.217
	Two-step (Utterance)	0.534	0.241	0.242	0.238
	Single-step (Utterance)	0.519	0.225	0.267	<u>0.236</u>
LLM Fine-tuning	(Encoder) MentalBERT	0.532	0.254	0.264	0.252
	(Encoder) MentalRoBERTa	0.568	0.356	0.258	0.263
	(Decoder) LoRA (8B) (Utterance)	0.532	0.293	0.325	0.303
	(Decoder) LoRA (8B) (Dialogue)	0.657	0.367	0.278	<u>0.300</u>
	(Generative) Full FT (8B)	0.657	0.375	0.260	<u>0.289</u>
Ensemble	Across approaches (Voting)	0.591	0.299	0.304	<u>0.300</u>
	Across approaches (Averaging)	0.439	0.265	0.231	0.212
	Best submissions (Voting)	0.672	0.470	0.284	0.320
	Best submissions (Averaging)	0.193	0.098	0.176	0.105

Table 4: Results for team TONI-NLP on the PsyDefDetect 2026 shared task test set, grouped by approach. Within each group, **bold** denotes the best macro F1 and underline denotes the second best. **8B** refers to Llama-3.1-8B-Instruct. A simple voting over our best submissions achieved the highest score. Only systems included in the submitted/tested result set are shown.

indicating that both retrieval quality and class coverage are important. While $k = 25$ showed signs of prompt saturation, $k = 5$ provided insufficient coverage, with $k = 20$ the best balance.

The two-step hierarchical approach produced only marginal gains over the single-step baseline using the utterance alone, while adding one prior turn reduced test performance. The two-step design reduces the classification space by routing instances to a subset-specific classifier. However, this introduces error propagation from the first stage. In contrast, the single-step MLR shows a bias toward majority classes (Appendix Figure 2).

Across all fine-tuned models, using only the target utterance matched or outperformed full dialogue context, consistent with findings from the hierarchical approach. Among encoder models, MentalRoBERTa and MentalBERT outperformed the non-specialized BERT models. All encoder models were fully fine-tuned given their small size. Full fine-tuning of Llama-3.1-8B-Instruct was feasible but computationally expensive, while Llama-3.3-70B-Instruct could only be fine-tuned using LoRA due to GPU memory constraints. LoRA fine-tuning of the 8B model got the best result (0.303), suggesting that parameter-efficient adaptation is both more scalable and more effective for this task.

Voting outperformed averaging-based ensembles, suggesting that hard voting is more robust to poorly calibrated confidence scores. The best ensemble (0.320) combined decoder LoRA (utterance), decoder LoRA (dialogue), and generative

full fine-tuning. Their diversity in input context and training strategy likely produced complementary error patterns and gains over any single model.

6 Conclusion and Future Work

We described TONI-NLP’s participation in PsyDefDetect 2026, comparing prompting, hierarchical classification, fine-tuning, and ensemble strategies for defense mechanism detection. Our results showed that ensemble methods achieved the strongest overall performance, while target-utterance-only configurations performed well across both hierarchical and fine-tuned models. These findings suggest that reducing input complexity and combining complementary systems may be useful strategies for this task. In future work, we plan to further investigate hierarchical approaches using LLMs and explore data balancing strategies at each step of the classification pipeline.

7 Limitations

Our experiments were limited to the English mental health dataset released by the shared-task organizers. Given the dataset’s small size and class imbalance, the generalizability of our findings should be interpreted with caution. The limited shared-task timeframe also restricted experimentation, preventing extensive ablation studies. Finally, the methods presented in this work were developed solely for research purposes within the shared-task setting and are not intended for clinical use.

8 Ethics

This work uses the annotated PSYDEFCONV dataset provided by the shared-task organizers. The conversational data was crowdsourced from consenting participants, and any future use of the dataset should continue to preserve participant anonymity. Because our approach relies on the PSYDEFCONV annotation scheme and the DMRS psychological framework, it may reproduce or amplify biases embedded in these theoretical and annotation frameworks. Therefore, any future application of this work should involve multidisciplinary collaboration with clinical psychologists and affected communities, particularly if the methods are considered for use beyond research settings.

Acknowledgments

Research reported in this publication was supported by the Natural Science and Engineering Research Council of Canada (Diana Inkpen). Juan Antonio Lossio-Ventura and Francisco Pereira were supported by the National Institute of Mental Health Intramural Research Program (ZICMH002968). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

- Callum Chan, Sunveer Khunkhun, Diana Inkpen, and Juan Antonio Lossio-Ventura. 2025. [Prompt engineering for capturing dynamic mental health self states from social media posts](#). In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 256–267, Albuquerque, New Mexico. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Mariagrazia Di Giuseppe and J Christopher Perry. 2021. The hierarchy of defense mechanisms: Assessing defensive functioning with the defense mechanisms rating scales q-sort. *Frontiers in psychology*, 12:718440.
- Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.
- Muskan Garg. 2023. [Mental health analysis in social media posts: A survey](#). *Archives of Computational Methods in Engineering*, 30.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *proceedings of the thirteenth language resources and evaluation conference*, pages 7184–7190.
- Juan Antonio Lossio-Ventura, Callum Chan, Arshitha Basavaraj, Hugo Alatrística-Salas, Francisco Pereira, and Diana Inkpen. 2025. [5cNLP at BioLay-Summ2025: Prompts, retrieval, and multimodal fusion](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 215–231, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.
- J. Christopher Perry and Melissa Henry. 2004. [Chapter 9 - studying defense mechanisms in psychotherapy using the defense mechanism rating scales](#). In Uwe Hentschel, Gudmund Smith, Juris G. Draguns, and Wolfram Ehlers, editors, *Defense Mechanisms*, volume 136 of *Advances in Psychology*, pages 165–192. North-Holland.
- Ruba Skaik and Diana Inkpen. 2020. [Using social media for mental health surveillance: A review](#). *ACM Comput. Surv.*, 53(6).
- Bin Wang and C-C Jay Kuo. 2020. Sbert-wk: A sentence embedding method by dissecting bert-based word models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2146–2157.

A Appendix

A.1 Additional Results for the Hierarchical Approach

We present additional results on the validation data held out from the training set. Table 5 shows the results of the hierarchical approach (2 steps), and compares them with the results of the 9-way classifier (Single step).

A.2 Aggregated Category Classification

In developing the hierarchical approach, we conducted a series of experiments on our validation set to optimize the initial categorization. We evaluated both MLR and Support Vector Machine (SVM) models across various feature representations, including multiple combinations of TF-IDF, standalone mental-bert, and SBERT-WK embeddings. Furthermore, we experimented with incorporating different context sizes (e.g., 0, 1, and 3 prior dialogue turns) alongside the target utterance.

This comparative evaluation was utilized to classify the three primary aggregated defense categories: Immature, Neurotic, and Mature, explicitly excluding the "No Defense" category. As detailed in Table 6, our results indicate that the target utterance only setting combined with an MLR classifier using concatenated SBERT-WK (mental-bert-base-uncased) and TF-IDF features yielded the best performance.

Features	Method	3 Turns		1 Turn		0 Turns	
		Acc.	F1	Acc.	F1	Acc.	F1
TF-IDF	MLR	0.4676	0.3772	0.5417	0.4588	0.5093	0.4362
	SVM	0.4722	0.3725	0.5139	0.4143	0.5185	0.4380
MentalBERT	MLR	0.5046	0.4381	0.5231	0.4632	0.5463	0.4884
	SVM	0.4537	0.3885	0.5324	0.4587	0.5463	0.4666
MentalBERT + TF-IDF	MLR	0.4907	0.4160	0.5370	0.4581	0.5694	0.4997
	SVM	0.4769	0.3840	0.5000	0.4021	0.5648	0.4899
SBERT-WK (mental-bert-base-uncased) + TF-IDF	MLR	0.5417	0.4610	0.5694	0.4977	0.5833	0.5189
	SVM	0.5185	0.4390	0.5046	0.4106	0.5139	0.4478

Table 6: Performance comparison for the first-step aggregated classifier in different model configurations.

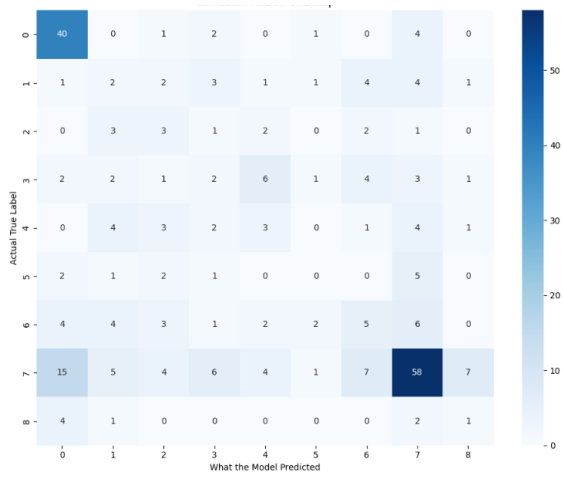
A comparison of the confusion matrices on Figure 2 reveals that, the single-step baseline struggled because it defaulted too often to the majority classes (Class 7 and Class 0). The hierarchical model improved this by routing data into aggregated categories first. This prevented mid-tier classes from being overshadowed by the majority.

Class (Defense Categ.)	Support (N)	Single-Step F1	Hier. F1	Δ F1	Key Error Dynamic / Impact of Hierarchical Routing
0 (No Defenses)	48	0.69	0.68	-0.01	Remains default prediction; reduced swamping of mid-tier classes.
1 (Action Defenses)	19	0.10	0.16	+0.06	Improved boundary resolution; less conflation with Class 6.
2 (Major Image-Distorting)	12	0.19	0.16	-0.03	Signal degradation; misclassified through sequential routing.
3 (Disavowal Defenses)	22	0.10	0.10	0.00	Persistent low discriminative power across both architectures.
4 (Minor Image-Distorting)	18	0.17	0.23	+0.06	Benefited from localized, secondary-level feature evaluation.
5 (Neurotic Defenses)	11	0.00	0.00	0.00	Fundamental data sparsity; unresolvable by structural changes.
6 (Obsessional Defenses)	27	0.20	0.32	+0.12	Significant Gain: Successfully shielded from Class 0 swamping.
7 (High-Adaptive Defenses)	107	0.60	0.61	+0.01	Marginal stabilization; remains the dominant statistical bias.
8 (Needs More Information)	8	0.11	0.00	-0.11	Total Attrition: Weak minority signal filtered out by two-step routing.
Global Metrics	272	Acc: 41.9%	Acc: 44.5%	+2.6%	Overall gains driven by localized resolution of mid-tier classes.

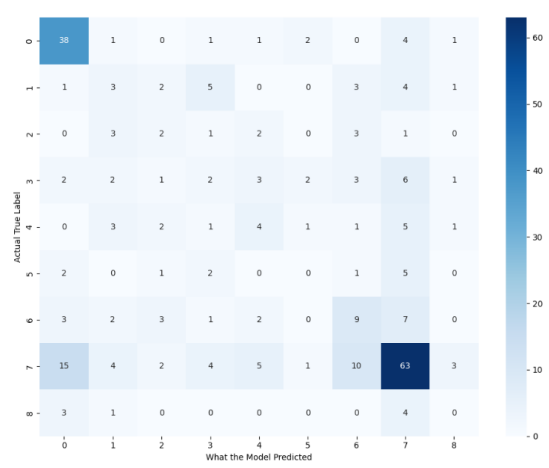
Note: Δ F1 denotes the absolute change in performance, calculated as (Hierarchical Macro F1 – Single-Step Macro F1). Positive values indicate an improvement utilizing the two-step approach.

Table 5: Comparative Performance and Error Dynamics (Single-Step vs. Two-Step Hierarchical Approach)

For example, true positives for Class 6 almost doubled (from 5 to 9) because the secondary classifier had a narrower, more focused task. However, the hierarchical approach has a noticeable drawback: if a mistake is made at the first step, it would propagate into the next step. This cascading error makes it difficult to correctly predict the lower frequency classes. For instance, the first-level classifier misclassified all Class 8 instances (mostly categorizing them into Classes 0 and 7), so the secondary classifier will not evaluate them. A similar case happened in Class 2, where true positives fell from 3 to 2. Finally, the confusion matrices show that some issues are strictly due to a lack of data. Class 5 has very few examples, and neither model could predict it correctly. In both approaches, nearly half of the Class 5 instances were mistakenly predicted as Class 7.



(a) Single-step Baseline



(b) Two-step Hierarchical Model

Figure 2: Confusion matrices comparing the final 9-way classification performance of (a) the Single-step baseline and (b) the two-step Hierarchical model on the validation set.

CanSA at MedExACT@ACL 2026: Zero-Shot, Fine-Tuned, and Retrieval-Augmented Extraction of Clinical Decisions with Corpus Boundary Diagnostics

Mohammed Alliheedi^{1*}, Robert E. Mercer^{2*}, Anemily Vincens Machina^{2*},
Sudipta Singha Roy^{2*}, Yetian Wang^{3*}, Xindi Wang^{4*}

¹Al-Baha University, Saudi Arabia

²University of Western Ontario, Canada

³University of Waterloo, Canada

⁴Shandong University, China

**These authors contributed equally to this work.*

maliheedi@bu.edu.sa, {rmercer, anemily.machina, ssinghar}@uwo.ca,
yetian.wang@uwaterloo.ca, xindi.wang@sdu.edu.cn

Abstract

We present the CanSA system for the MedExACT@ACL 2026 shared task, which requires extracting and classifying clinical decisions from ICU discharge summaries into nine DICTUM categories. We have developed three approaches: (1) a training-free system which consists of a preprocessing module that normalizes text and an inference engine combining zero shot LLMs with a RAG ensemble, (2) a supervised fine-tuning method which required training, and (3) a training-free retrieval-augmented pipeline employing TF-IDF-based lexical retrieval to surface in-context exemplars from the development corpus, combined with section-aware chunking and structured extraction calls to a large language model. Our team’s best submission achieved a Final Score of 0.41, ranking 34th out of 37 on the official test leaderboard.

1 Introduction

The MedExACT shared task at the BioNLP 2026 workshop (Elgaar et al., 2026) challenges systems to extract exact text spans representing medical decisions from Intensive Care Unit (ICU) discharge summaries and classify them according to the DICTUM taxonomy (Ofstad et al., 2016; Elgaar et al., 2024). Because the evaluation heavily penalizes boundary misalignment, systems must demonstrate both deep semantic classification and strict character level precision.

Team CanSA presents three methodologies. The first addressed this challenge by building a modular, training-free inference pipeline. To avoid the heavy costs of task specific fine tuning, we evaluated several leading models, specifically Llama 3, Qwen, and Kimi, by deploying them through

local Ollama instances on an NVIDIA A6000 and remote NVIDIA NIM API endpoints. The second provided a supervised fine-tuning of a bi-directional encoder model on a token labelling task. And lastly, employed a training-free, retrieval-augmented pipeline in which TF-IDF-based lexical retrieval surfaces in-context exemplars from the development corpus, and a large language model performs sequential span proposal and structured extraction over section-aware chunks of each discharge note, followed by schema-based validation and automated repair. We also contribute a rigorous error analysis of the MedDec corpus by auditing 104 same category overlapping annotation pairs providing actionable insights for future clinical dataset curation.

The remainder of this paper is organized as follows: Section 2 details the specifics of each of the three methodologies. Section 3 outlines our experimental setup. Section 3.2 presents the shared task results and post deadline experiments, followed by the corpus audit in Section 4. We conclude in Section 5.

2 System Architectures

2.1 Training-Free Inference Pipeline

The MedExACT task requires exact character offset matching. During initial pipeline development, we observed severe boundary extension errors caused by standard tokenizers splitting at MIMIC III de identification markers (e.g., [**. . .**]). To mitigate this, we implemented a strict length preserving normalization heuristic: replacing all new-line characters ($\backslash n$) with spaces. This single transformation resolved tens of thousands of offset mis-

matches across the raw clinical text.

Following normalization, documents are segmented using a section aware sliding window. To optimize inference efficiency and reduce API costs, the pipeline explicitly skips the *Social History* and *Family History* sections, as corpus analysis revealed zero gold standard annotations in these segments. Finally, a post processing filter enforces a minimum annotation length of five characters to eliminate single token model artifacts.

Our primary baseline module relies on zero shot extraction using frontier LLMs. The system prompts the model with a strict JSON schema definition encompassing the nine DICTUM categories.

A critical engineering challenge involved managing models equipped with implicit reasoning pathways (e.g., Qwen and DeepSeek). These models frequently output reasoning chains (often enclosed in <think> tags) that corrupt the required JSON structure. To enforce deterministic formatting, we implemented a dual suppression strategy: prepending a /no_think directive to the user prompt and explicitly setting enable_thinking: false in the API payload.

To evaluate the impact of in context learning without task specific fine tuning, we developed an alternative RAG enhanced pipeline. We indexed the MedDec training dataset using FAISS GPU (Johnson et al., 2019). For each sliding window chunk of the target document, the system retrieves the top k most semantically similar training examples and appends them to the prompt as few shot exemplars, attempting to ground the model’s predictions in gold standard phrasing.

Initial development was conducted locally on an NVIDIA A6000 GPU (48 GB VRAM) using Ollama to host a high density ensemble. However, hardware constraints necessitated a transition to a constrained local RTX 3060 (12 GB VRAM) setup, relying on NVIDIA NIM API endpoints to access models exceeding 70 billion parameters.

This remote transition introduced severe operational bottlenecks. We observed that the increased token payload of the RAG pipeline highly correlated with HTTP 504 (Gateway Timeout) and HTTP 429 (Too Many Requests) errors. These connection drops frequently caused truncated JSON responses mid transit—a degradation effect that forced zero span extractions for the affected chunks. Consequently, the final pipeline enforces a stability first configuration, utilizing a strict 15 second API delay and exponential backoff to ensure continuous

execution during the evaluation phase.

2.2 Supervised Fine-tuning

The supervised learning method fine tuned a bi-directional encoder model: BAAI/bge-m3 (Chen et al., 2023). The goal was to start with a single label classification task, and then use the learned classifiers to extend to multi-label classification. A bi-directional encoder was used, compared to an auto-regressive Language Model, as the extra right side context should be important for labelling tokens and spans. To create a single label for each token, any token with multiple labels was assigned the most frequent label as they appear in the training dataset. Unfortunately, time did not permit the multi-label classification, so only results for the single label task are reported.

The model was trained using Cross Entropy Loss weighted so that each class had equal contribution. The Huggingface training interface was used: it is relatively straightforward to use and hack as needed; it allows multiple GPUs with no extra effort. Parameters used: learning rate $2e-4$, per device batch size of 1 with gradient accumulation of 16 steps, 100 training epochs with best checkpoint per epoch retrained based on validation F1 score, 100 total epochs with a warmup of 110 steps. Three different seeds were used after which final token labels on the test set were based on the soft ensemble, summation followed by argmax, of all three best checkpoints.

Given more time, other foundation models would be used for fine-tuning on the single label task, and the classification vectors from those training runs would be used to seed a binary classifier for each label: with bias initialization tuned on the training set.

2.3 Training-Free Retrieval-Augmented Pipeline

Span-level adverse-event–style extraction from long discharge summaries is addressed through a training-free, retrieval-augmented language model pipeline. Each discharge note is segmented into clinical sections. Following task conventions, predicted spans overlapping allergy regions are removed in a post-hoc filtering step. The remaining text is split into section-aware chunks of at most 5,000 characters, ensuring that each language model call fits within context limits while coarse document structure is preserved.

A lexical index is constructed over the development corpus: for every document with paired raw text and reference annotations, a truncated prefix of the discharge (up to 80,000 characters per note) is indexed together with a compact serialization of up to 40 gold spans comprising shortened decision text and category labels. Documents are represented using TF-IDF features (unigrams and bigrams; maximum 50,000 features; document-frequency cap of 0.95). At inference time, each chunk’s text serves as the query; training documents are ranked by cosine similarity and the top three matches are retrieved, excluding the current discharge ID to prevent trivial self-retrieval. For each retrieved document, a truncated text excerpt (up to 3,500 characters) along with the corresponding compact gold annotations is incorporated into the prompt as in-context exemplars.

Claude Sonnet 4 is employed at temperature zero. For each chunk, two sequential calls are issued: (1) a span proposal step, conditioned on the chunk text, section headers, and the retrieval block; and (2) a structuring step, in which the proposals are converted into JSON with chunk-local character offsets, span text (referred to as *decision*), and a category label. Local offsets are subsequently shifted to global UTF-8 offsets within the full note. Where a predicted span text does not exactly match the underlying substring, local realignment is attempted within a window around the chunk position, including whitespace-tolerant matching. Annotations from all chunks are then merged, deduplicated by start position, end position, and category, sorted, and assigned stable identifiers; spans overlapping allergy regions are discarded.

The assembled document-level JSON is validated against a schema tied to the full raw text. In cases where validation fails, a single additional language model call is issued, conditioned on the entire discharge note and the current prediction; the model is instructed to rewrite the full annotation set so as to satisfy all constraints. Allergy-region filtering is subsequently re-applied where applicable. Finally, category labels are mapped to the numeric codes required by the shared-task evaluation format.

3 Experimental Setup and Results

3.1 Evaluation Metrics

System performance is evaluated using the official MedExACT Final Score, defined as the arithmetic

mean of the Base Score and the Worst Group Score (a fairness metric evaluated across patient demographic subgroups). The Base Score itself is the arithmetic mean of Span F1 (requiring exact character boundary matches) and Token F1 (measuring word level overlap). Throughout our diagnostics, we track the differential between Token F1 and Span F1 to isolate boundary tokenization artifacts from clinical reasoning errors.

3.2 Official Shared Task Results

For the official MedExACT test phase, Team CanSA submitted three independent runs to the Codabench leaderboard. Our training free zero shot pipeline using Qwen 3.5 achieved a Span F1 of 0.31 and a Token F1 of 0.44 (resulting in a Base Score of 0.38) alongside a Final Score of 0.34.

For the supervised fine-tuning method, the best checkpoints had training/eval token F1 scores 0.93/0.63, 0.87/0.61, and 0.91/0.62. Note: this is on the single label task. On the validation set, the ensemble labels achieved a 0.61 validation and span annotation F1 of 0.16 (task score 0.38) with sub group scores from 0.28 to 0.43. The final score on the validation set was 0.33. The leaderboard score for this method was 0.37.

The training-free retrieval-augmented pipeline achieves a span F1 of 0.35, token F1 of 0.51, and a final score of 0.38 on the validation set. On the official test set, a span F1 of 0.34, token F1 of 0.50, and a final leaderboard score of 0.41 are obtained.

Our highest performing submission achieved a Base Score of 0.42 and a Final Score of 0.41, placing 34th out of 37 on the official test set leaderboard.

3.3 Post Deadline Architectural Analysis

To properly evaluate the architectural trade offs of the first pipeline, we conducted post deadline experiments on the complete validation set. Table 1 summarizes the performance of frontier models, scored strictly via the official MedExACT evaluation script. An analysis of these results highlights two critical operational findings:

The RAG API Penalty. In a high latency remote environment, appending FAISS retrieved context dynamically increases payload size, correlating strongly with HTTP 504 timeouts and truncated JSON responses. While models like Kimi K2 proved highly resilient, other models suffered from mid transit connection drops degrading overall throughput.

Configuration	Span F1	Tok F1	Base	Final
Kimi K2 (Zero-Shot)	0.28	0.43	0.35	0.32
Kimi K2 (RAG Ens.)	0.29	0.44	0.36	0.34
Qwen 3.5 (Zero-Shot)	0.31	0.44	0.38	0.34
Qwen 3.5 (RAG Ens.)	0.28	0.43	0.35	0.30
Llama 3.3 (Zero-Shot)	0.29	0.42	0.36	0.31
Mistral (Zero-Shot)	0.27	0.43	0.35	0.30

Table 1: Validation set ($n = 53$) performance across NVIDIA NIM API endpoints.

The Token to Span Gap. Across all evaluated models, Token F1 consistently outperformed Span F1 by a margin of 0.13 to 0.16. For example, while Mistral achieved a peak Token F1 of 0.43, its strict Span F1 dropped to 0.27. This massive differential proves that while frontier LLMs possess the semantic capability to identify clinical decisions, they fail to predict the exact character boundaries required by the MedDec gold standard.

4 Corpus Audit: The Boundary Gap

Category	Tokens	Gap Size	Count
1	12.1 ± 25.7	178.5 ± 1959.3	6.5 ± 9.7
2	8.7 ± 10.9	276.1 ± 4400.4	1.1 ± 4.9
3	6.9 ± 69.3	23.2 ± 64.1	24.6 ± 175.4
4	6.4 ± 3.4	294.5 ± 5023.8	0.3 ± 3.2
5	8.1 ± 55.5	31.9 ± 124.7	31.2 ± 85.1
6	5.7 ± 39.9	79.6 ± 445.4	15.2 ± 20.1
7	12.8 ± 43.8	44.7 ± 184.9	16.6 ± 28.2
8	9.3 ± 14.8	358.1 ± 6756.6	0.24 ± 3.08
9	19.6 ± 27.5	36.5 ± 197.9	4.3 ± 7.2

Table 2: Annotation statistics at the document level: Tokens (annotation size); Gap Size (tokens between annotations); Count (annotations in document).

The persistent 0.13 to 0.16 differential between Token F1 and Span F1 highlighted in Section 3.3 indicates that while frontier LLMs successfully identify clinical concepts, they don’t always map these concepts to the exact character boundaries defined by the gold standard. To investigate the structural cause of this gap, we conducted a comprehensive annotation quality audit of the MedDec corpus and a token level analysis of annotation statistics (see Table 2).

We identified 104 same category overlapping annotation pairs distributed across 79 documents (representing roughly 17.5% of the validation corpus). A manual review of these pairs revealed three distinct fragmentation typologies:

- **Boundary Extension Errors (n=21):** These

are artificial splits caused by tokenizer artifacts, specifically where automated tools (e.g., Stanza) over extended annotation boundaries adjacent to MIMIC III de-identification markers (e.g., [****** . . . ******]).

- **Near Identical Duplicates (n=7):** Annotations covering the exact same semantic clinical concept with minor 1 to 2 character shifts.
- **Genuine Subspans (n=74):** Hierarchical or nested clinical decisions, of which 33 were fragment level annotations under 20 characters in length.

The presence of these structural artifacts, particularly the 21 boundary extension errors linked to de-identification masking, imposes a hard recall ceiling on exact match extraction systems. Because strict span evaluation harshly penalizes partial boundary misalignment, heuristic post processing tools cannot systematically recover these points without introducing false positives. Consequently, we argue that the Token to Span gap is largely a reflection of corpus level tokenization methodology rather than a deficit in LLM clinical reasoning.

5 Conclusion

In this paper, we described the CanSA system for the MedExACT@ACL 2026 shared task. Our team developed three approaches including a training-free extraction pipeline that combines length preserving text normalization, stable API management, and zero shot LLM inference, a supervised fine-tuning approach, and a training-free retrieval-augmented pipeline employing TF-IDF-based lexical retrieval to surface in-context exemplars from the development corpus combined with section-aware chunking, achieving a Final Score of 0.41 on the official test leaderboard (Rank 34/37). Through extensive post deadline evaluation on the first method, we demonstrated the “RAG API Penalty” where increased payload sizes trigger silent mid transit connection timeouts and highlighted a systemic Token to Span F1 gap. Our corpus audit links this boundary fragmentation directly to tokenizer artifacts at de-identification markers. We conclude that while frontier LLMs possess strong clinical semantic capabilities, their deployment in exact match extraction tasks requires stability-first engineering and necessitates tokenizer aware corpus curation for future clinical benchmarks.

Ethics and Data Statement

This work uses the MIMIC III critical care database (Johnson et al., 2016), which is a restricted dataset requiring credentialed access via PhysioNet. All analyses were conducted in compliance with the MIMIC III data use agreement. No patient re-identification was attempted. The MedDec annotations (Elgaar et al., 2024) are publicly available for research purposes.

References

- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2023. [Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). Preprint, arXiv:2309.07597.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024. [MedDec: A dataset for extracting medical decisions from discharge summaries](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16442–16455, Bangkok, Thailand. Association for Computational Linguistics.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Mehrnaz Sadrolashrafi, Mitra Mohtarami, Adrian Wong, Hadi Amiri, and Leo A. Celi. 2026. [Overview of medical decision extraction, analysis, and classification task \(medexact\) 2026](#). In *The 25th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, San Diego, California, USA. Association for Computational Linguistics.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.
- Eirik H Ofstad, Jan C Frich, Edvin Schei, Richard M Frankel, and Pål Gulbrandsen. 2016. Dictum: a taxonomy of medical decisions in clinical encounters. *BMJ open*, 6(2):e010078.

CASPAR: A Context-Aware Span Refinement Approach for Decision Support

Jing Tao
School of Computing
Queen’s University
Kingston, ON, Canada
jing.tao@queensu.ca

Amir Eskandari
School of Computing
Queen’s University
Kingston, ON, Canada
Amir.Eskandari@queensu.ca

Farhana Zulkernine
School of Computing
Queen’s University
Kingston, ON, Canada
farhana.zulkernine@queensu.ca

Abstract

Extracting structured clinical information by selecting a text span from long medical discharge summaries using Natural Language Processing (NLP) poses difficult challenges. Heterogeneity of the text, semantic overlap of the decision categories, and uncertainty in identifying boundaries of text span aligned with highly unbalanced labels. Contribute to these challenges, the MedExACT competition in BioNLP @ ACL 2026 formulates the medical concept and text span extraction problem from unstructured text as a classification task over nine DICTUM decision categories. Unstructured ICU discharge summaries are used from the MedDec dataset, for solution approach across eight demographic subgroups. We propose a two-stage data processing approach for the text span extraction and classification task where Stage 1 performs document-level sequence labeling with a pipeline composed of a RoBERTa-base, a Bidirectional Gated Recurrent Unit (Bi-GRU), and a Conditional Random Field (CRF). Stage 2 applies a local span refinement module which revisits each candidate span to adjust both category assignments and boundary offsets. Our framework achieves a final score of 0.5668 which substantially outperforms the provided baseline. Ablation results further suggest that BiGRU-based contextual consolidation improves CRF-based decoding in text span extraction and classification.

1 Introduction

Discharge summaries in electronic health records provide a rich source of clinical information for decision-making. However, the heterogeneity of the text, semantic overlap of the decision categories, and uncertainty in identifying text boundaries corresponding to highly unbalanced labels make automatic extraction of text for decision support extremely challenging. The MedExACT (Elgaard et al., 2026) competition in BioNLP @ ACL

2026 formulates medical concept and text span extraction as a classification task over the MedDec dataset (Elgaard et al., 2024), the task requires systems to jointly detect decision spans and assign each span to one of nine DICTUM categories (Ofstad et al., 2016), while evaluating robustness of the approach across multiple demographic subgroups. This setting poses three compounding challenges as discharge summaries often exceed encoder context windows, DICTUM categories overlap semantically, and the stringent exact-match criteria render minor boundary offsets as complete span-level failures.

To address these challenges, we propose CASPAR, a two-stage approach to here. Stage 1 implements a pipeline with the deep learning backbone of a RoBERTa-base, a Bidirectional Gated Recurrent Unit (Bi-GRU), and a Conditional Random Field (CRF), to improve contextual representation learning for document-level span detection. To further improve the precision of span boundary detection, we introduce a lightweight span refinement module in Stage 2, which revisits each candidate span to correct both the category assignment and boundary offsets. Our system ranks 10th among the submissions from 37 competing teams (Final Score: 0.5668), substantially outperforming the organizer baseline. The source code for CASPAR is available at <https://github.com/schorm/caspar-medexact>.

2 Related Work

Recent NER research largely builds on token-level sequence labeling, where each token is assigned a BIO-, BILOU-, or related tag (Ramshaw and Marcus, 1995; Ratnov and Roth, 2009). In biomedical NER, this remains a central baseline formulation. Verma et al. (2023) organize mainstream biomedical NER approaches into simple token classifiers, CRF-based sequence labeling, and span

prediction. CRF-based decoders remain widely used to impose local label-transition constraints on top of contextual encoders (Huang et al., 2015; Jonker et al., 2024). In biomedical settings, token-level formulations also underpin strong practical systems such as AIONER (Luo et al., 2023), and recent cross-corpus tools such as HunFlair2 (Sanger et al., 2024). More recent works address dataset inconsistency and generalization issues within the sequence-labeling paradigm (Ruano et al., 2025).

Beyond token-level tagging, alternative formulations have been proposed to better model entity boundaries and span-level structure. Query-based and span-oriented methods include PIQN, which uses parallel instance queries for entity extraction (Shen et al., 2022), and BINDER, which models span-type matching through a contrastive bi-encoder formulation (Zhang et al., 2023). Two-stage and span-aware variants further emphasize boundary recovery including Locate and Label (Shen et al., 2021), document-level span fusion in ScdNER (Wei and Li, 2023), and boundary-aware generation in DiffusionNER (Shen et al., 2023). Tang et al. (2023) note that span-based methods can suffer from imbalanced span candidate spaces and difficulties in accurate boundary modeling.

Biomedical and clinical NER remains particularly challenging because texts are terminology-dense, annotation schemes often differ across datasets, and label spaces can be fine-grained and semantically overlapping. Recent work highlights these challenges from multiple angles, including annotation inconsistency across corpora (Ruano et al., 2025), large fine-grained label spaces (Yang et al., 2023), and the need to generalize beyond fixed entity inventories (Cocchieri et al., 2025). In medical decision extraction, these issues become even more pronounced. Elgaar et al. (2024) show that strong baselines still struggle under strict span-level evaluation; stronger token-level predictions do not necessarily translate into better exact-match span recovery, and adding CRF does not improve the baseline. These observations motivate our focus on contextual consolidation before CRF decoding and explicit second-stage span refinement.

3 Methodology

Figure 1 presents an overview of our proposed two-stage framework. Stage 1 performs document-level sequence labeling to detect candidate deci-

sion spans and assigns preliminary category labels. Stage 2 subsequently revisits each candidate span independently and refines its prediction using local context, correcting both category assignments and span boundaries. The two stages are trained separately, with Stage 2 taking the outputs of Stage 1 as its input.

The text span extraction and classification task is formulated as follows. Given a discharge summary D , the model predicts a set of labeled decision spans $S = \{(s_i, e_i, c_i)\}$, where s_i and e_i denote the start and end token indices of span i , and $c_i \in C$ denotes one of the nine following DICTUM categories: Contact related, Gathering information, Defining problem, Treatment goal, Drug related, Therapeutic procedure, Evaluating test result, Deferment, and Advice/precaution (Ofstad et al., 2016). The task therefore requires the model to jointly localize decision evidence in free text and assign each extracted span to the correct category. Following common practice in biomedical NER, we cast the problem as token-level sequence labeling with BIO tags over the label space

$$\mathcal{Y} = \{B-c, I-c, O \mid c \in C\}$$

where $B-c$ marks the beginning of a span of category c , $I-c$ marks its continuation, and O denotes tokens outside any decision span.

3.1 Stage 1: Document-Level Sequence Labeling

Discharge summaries are typically unstructured and often exceed the maximum input length supported by pretrained encoders. To preserve as much surrounding context as possible, we adopt a sliding-window strategy that partitions each document into overlapping chunks, which are encoded independently.

Given a discharge summary tokenized into a sequence of n tokens, $D = \{w_1, \dots, w_n\}$, where w_i denotes the i -th token, we partition the document into M overlapping windows, denoted by $\{X^{(j)}\}_{j=1}^M$, where each chunk $X^{(j)}$ contains at most m tokens based on a fixed maximum window length shared across all windows. For each chunk, which is padded to length m when necessary, the encoder first produces contextualized token representations:

$$H^{(0,j)} = \text{Encoder}(X^{(j)}) \in R^{m \times d} \quad (1)$$

where d denotes the hidden dimensionality of the encoder.

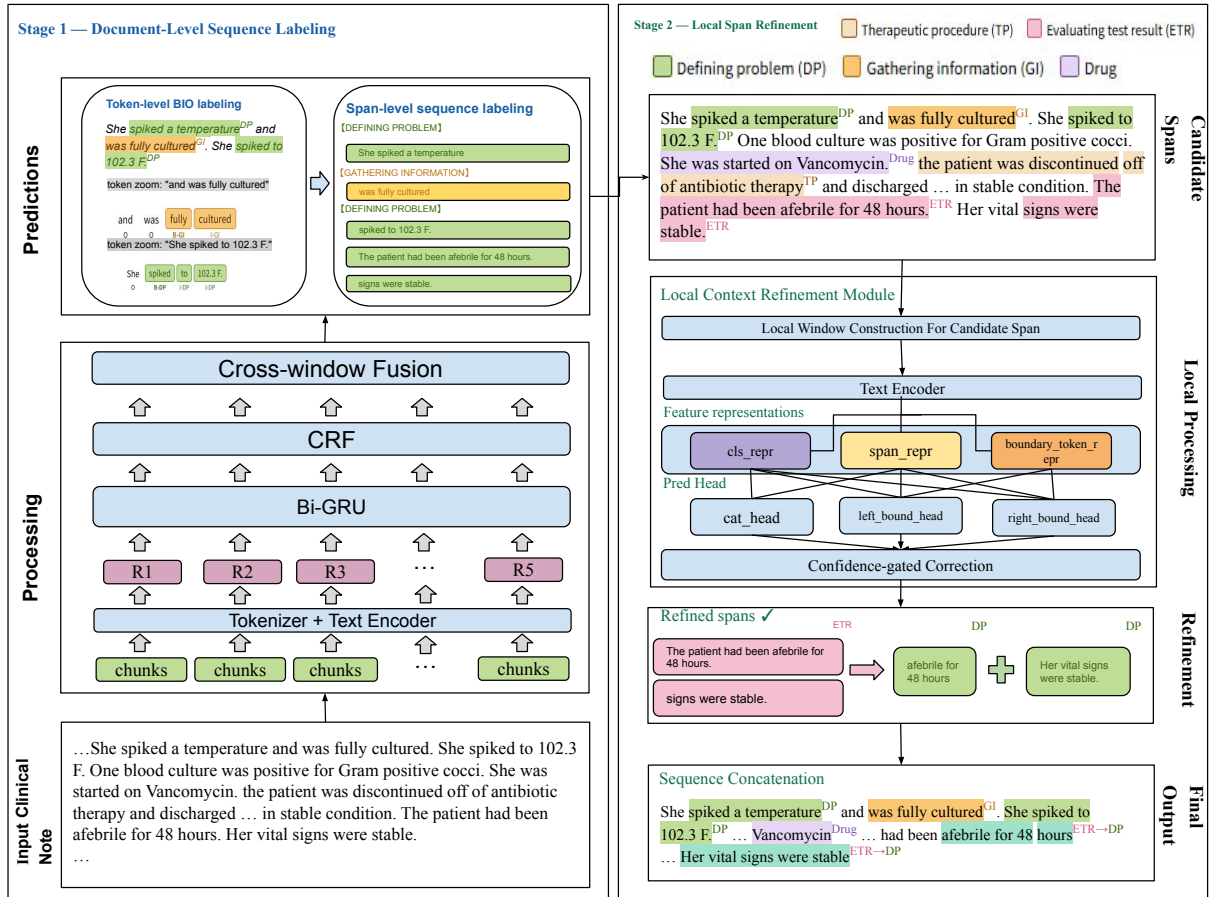


Figure 1: Architecture of CASPAR, our two-stage framework for medical decision span extraction and classification. Stage 1 performs document-level sequence labeling to produce candidate spans. Stage 2 applies local span refinement to correct category assignments and span boundary detection.

To further consolidate sequential dependencies before structured decoding, we pass these representations through a single-layer bidirectional GRU:

$$H^{(j)} = \text{BiGRU}(H^{(0,j)}) \in R^{m \times d} \quad (2)$$

The resulting hidden states are then projected to token-level logits over the BIO label set \mathcal{Y} :

$$E^{(j)} = H^{(j)}W_e + b_e, \quad E^{(j)} \in R^{m \times |\mathcal{Y}|} \quad (3)$$

where $W_e \in R^{d \times |\mathcal{Y}|}$ and $b_e \in R^{|\mathcal{Y}|}$ are learnable projection parameters.

Since each document token may appear in multiple overlapping windows, we reconstruct document-level predictions by fusing the emission logits across windows. Let w_t denote the t -th token in the original discharge summary D . Each window $X^{(j)}$ corresponds to a contiguous segment of the original document, covering token positions from a_j to b_j . Let $\mathcal{W}(t)$ denote the set of all windows that contain token w_t , i.e.

$$\mathcal{W}(t) = \{j \mid a_j \leq t \leq b_j\}. \quad (4)$$

For each such window $j \in \mathcal{W}(t)$, the local position of token w_t within window $X^{(j)}$ is defined as

$$p_j(t) = t - a_j + 1. \quad (5)$$

We then compute the fused emission logits for token w_t as

$$\tilde{E}_t = \frac{\sum_{j \in \mathcal{W}(t)} \alpha_{j,t} E_{p_j(t)}^{(j)}}{\sum_{j \in \mathcal{W}(t)} \alpha_{j,t}}. \quad (6)$$

Here, $\alpha_{j,t}$ is a triangular weight that attenuates predictions made near window boundaries. For a window of length m_j , with $\epsilon = 0.1$, center $c_j = (m_j + 1)/2$, and radius $r_j = \max(c_j - 1, m_j - c_j)$, we define

$$\alpha_{j,t} = \max\left(\epsilon, 1 - \frac{|p_j(t) - c_j|}{r_j}\right). \quad (7)$$

Thus, predictions near chunk boundaries contribute less to the final document-level representation.

Finally, we apply a global CRF decoder over the fused logits to obtain the document-level BIO

tag sequence, where \mathcal{Y} is the token-level BIO label space defined above, and \mathcal{Y}^n denotes the set of all possible length- n BIO tag sequences

$$\hat{y} = \arg \max_{y \in \mathcal{Y}^n} \text{Score}_{\text{CRF}}(\tilde{E}, y). \quad (8)$$

Here, the CRF combines token-level emission scores with learned transition scores between adjacent labels, enabling globally consistent BIO decoding.

3.2 Stage 2: Local Span Refinement

Each Stage 1 candidate span $\hat{S}_i = (\hat{s}_i, \hat{e}_i, \hat{c}_i)$ is further processed by a lightweight local refinement module. Rather than re-encoding the full document, Stage 2 revisits each candidate in a local window. Given a document of length n tokens, we expand the predicted span by k context tokens on both left and right sides:

$$l_i = \max(1, \hat{s}_i - k), \quad r_i = \min(n, \hat{e}_i + k) \quad (9)$$

The resulting token interval $[l_i, r_i]$ is mapped back to the original character offsets to extract the corresponding local text, which is then re-tokenized for the encoder model. The predicted span boundaries are aligned to the re-tokenized local sequence via the corresponding character offsets, yielding the local boundary indices p_i^L and p_i^R . This yields contextual token representations $H_i^{\text{loc}} \in R^{L_i \times d}$, where L_i is the local sequence length after tokenization.

Followed the local processing module As shown in Fig. 1, from H_i^{loc} , we derive three feature representations as following a global context representation $\mathbf{h}_i^{\text{cls}} = H_i^{\text{loc}}[0]$ from [CLS] pooling; a span representation

$$\mathbf{h}_i^{\text{span}} = \frac{1}{|M_i|} \sum_{t \in M_i} H_i^{\text{loc}}[t], \quad (10)$$

obtained by mean-pooling over the predicted span tokens within the local window, where p_i^L and p_i^R denote the local token indices of the predicted left and right span boundaries, respectively, and

$$M_i = \{p_i^L, p_i^L + 1, \dots, p_i^R\}$$

is the set of local token indices covered by the i -th predicted span; and a boundary representation

$$\mathbf{h}_i^{\text{bdry}} = [H_i^{\text{loc}}[p_i^L]; H_i^{\text{loc}}[p_i^R]], \quad (11)$$

formed by concatenating the hidden states at the predicted left and right span boundaries.

These representations are consumed by a category head, two boundary heads, and two binary gate heads, where the gate heads predict whether a boundary correction should be applied on the left and right sides, respectively.

$$\mathbf{o}_i^{\text{cat}} = W_{\text{cat}}[\mathbf{h}_i^{\text{cls}}; \mathbf{h}_i^{\text{span}}] + b_{\text{cat}} \quad (12)$$

where $\mathbf{o}_i^{\text{cat}} \in R^{|C|}$, $\mathbf{h}_i^{\text{cls}} \in R^d$, $\mathbf{h}_i^{\text{span}} \in R^d$, and $|C|$ is the number of decision categories.

The left and right boundary heads operate on $\mathbf{h}_i^{\text{bdry}}$ and predict discrete offset classes over a label space of size $2\delta_{\text{max}} + 1$:

$$\mathbf{o}_i^L = W_L \mathbf{h}_i^{\text{bdry}} + b_L, \quad \mathbf{o}_i^R = W_R \mathbf{h}_i^{\text{bdry}} + b_R \quad (13)$$

where $\mathbf{h}_i^{\text{bdry}} \in R^{2d}$, $W_L, W_R \in R^{(2\delta_{\text{max}}+1) \times 2d}$, and δ_{max} is the maximum allowable offset for boundary predictions.

To construct Stage 2 training instances, we run Stage 1 on the training split and retain each predicted span that has a token-level IoU > 0 with at least one gold span. For each retained candidate, the matched gold span (s_i, e_i, c_i) defines the boundary offsets

$$\Delta_i^L = s_i - \hat{s}_i, \quad \Delta_i^R = e_i - \hat{e}_i \quad (14)$$

which are clipped to $[-\delta_{\text{max}}, \delta_{\text{max}}]$ to keep the prediction space tractable. The category head is optimized with label-distribution-aware margin loss (LDAM) (Cao et al., 2019) to mitigate class imbalance, while boundary and gate heads are trained with standard cross-entropy. The gate loss $\mathcal{L}_{\text{gate}}$ is defined as the average of the left and right gate classification losses. The joint objective is

$$\mathcal{L} = \mathcal{L}_{\text{cat}} + \lambda_b \cdot \frac{1}{2} (\mathcal{L}_{\text{left}} + \mathcal{L}_{\text{right}}) + \lambda_g \cdot \mathcal{L}_{\text{gate}}. \quad (15)$$

At inference time, category refinement is always applied, whereas boundary correction is applied only when both the boundary-head confidence and the gate probability exceed predefined thresholds τ_b and τ_g , respectively.

Denoting the accepted offsets as $\tilde{\Delta}_i^L$ and $\tilde{\Delta}_i^R$, the refined span is

$$s_i^* = \hat{s}_i + \tilde{\Delta}_i^L, \quad e_i^* = \hat{e}_i + \tilde{\Delta}_i^R \quad (16)$$

Finally, refined spans are aggregated at the document level. Rather than score-based suppression, we merge spans of the same category that overlap or are directly adjacent, sorting by (category, start, end) and merging consecutive same-category spans whenever the next span starts no later than the end of the previous.

Race Gap Heatmap for (Error Type, Category)

Cell value: subgroup error rate (num/den). Right side: high-low gap in percentage points.



Figure 2: Error-type disparity heatmap across racial subgroups.

3.3 Implementation Details

We follow the official MedExACT shared-task data split (Elgaar et al., 2026) and use the provided training and validation partitions throughout the experiments. As gold labels for the test set remain proprietary to the task organizers, local evaluation is restricted to the validation split. Consequently, all reported test results are derived exclusively from the official leaderboard following system submission.

All experiments use ROBERTA-base (Liu et al., 2019) with $m=512$ maximum length window as the text encoder. Stage 1 is trained for 10 epochs with an encoder learning rate of 1×10^{-5} , a task-head learning rate of 5×10^{-5} , and a batch size of 8. Stage 2 uses an independently initialized ROBERTA-base encoder which is trained for 10 epochs using the same learning rates with a batch size of 32. The local window expands each candidate span by 32 tokens on both sides before re-tokenization, and the local encoder input is truncated to a maximum length of 128 tokens. Boundary offsets are discretized with $\delta_{\max} = 3$, yielding 7 offset classes per boundary head. On training-set matched candidates, ± 3 tokens cover 92.6%/93.2% of left/right boundary errors (86.8% jointly). The boundary loss weight λ_b is set to 1.0, the gate loss weight

λ_g to 0.2, the boundary confidence threshold τ_b to 0.6, the gate probability threshold τ_g to 0.5, and the merge gap to 0.

3.4 Main Results

Table 1 reports the official MedExACT test results. CASPAR achieves a Final Score of 0.5668 on the shared-task leaderboard. For reference, we also report the mean and median scores across the 37 visible leaderboard submissions. Compared with the organizer baseline, CASPAR substantially improves Span F1 (0.5060 vs. 0.348).

In addition to the results from official leaderboard, Table 2 reports the repeated-run validation performance of CASPAR as mean \pm standard deviation together with 95% confidence intervals which confirms the reliability of CASPAR and demonstrates that its gains are not artifacts of random initialization.

3.5 Ablation Study

Table 3 presents validation results across system variants. Adding BiGRU before CRF yields a substantial gain in Final Score (+8.0pp), suggesting that contextual consolidation makes CRF decoding more effective for boundary-sensitive prediction. This finding also helps explain why CRF alone provides only limited gains, consistent with the

Dominant Category by Error Type

Bar height: error count in dominant category; color: category id

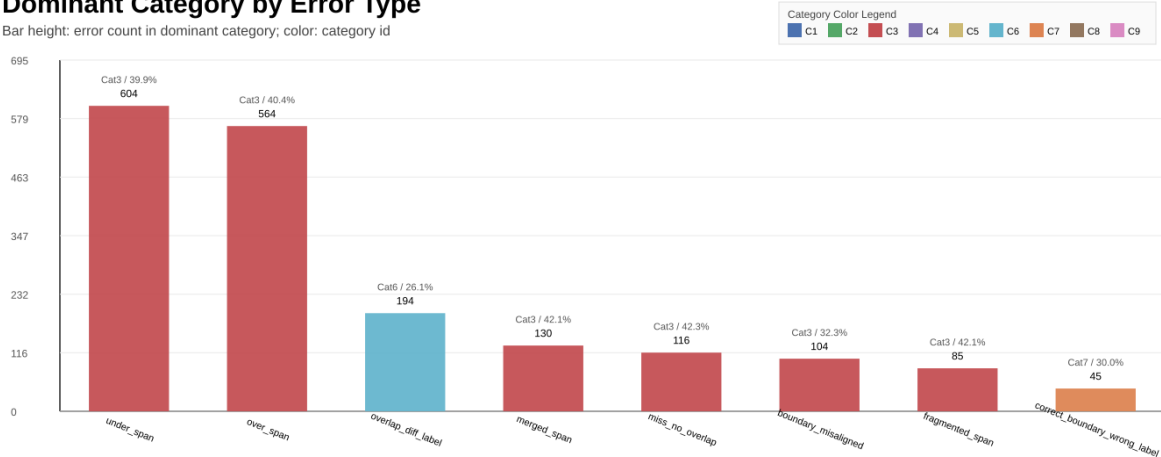


Figure 3: Dominant decision category by error type. Bar height shows the error count in the dominant category, and color indicates the category. Categories correspond to the nine DICTUM labels: Contact related (C1), Gathering information (C2), Defining problem (C3), Treatment goal (C4), Drug related (C5), Therapeutic procedure (C6), Evaluating test result (C7), Deferment (C8), and Advice/precaution (C9).

System	Final	WG	Span F1	Token F1
1st place	0.5965	0.5886	0.5419	0.6667
LB median	0.5331	0.5210	0.4605	0.6487
LB mean	0.5118	0.4983	0.4320	0.6193
Baseline [†]	–	–	0.348	–
CASPAR	0.5668	0.5602	0.5060	0.6409

Table 1: Official MedExACT test results. LB median = the Median of the results from MedExact Leaderboard, LB mean = the mean of the results from MedExact Leaderboard, Final = Final Score, WG = Worst-Group F1 which following the MedExACT evaluation metric.

Metric	Mean \pm Std	95% CI
Final Score	0.5007 \pm 0.0073	\pm 0.0064
WG Score (Hispanic)	0.4443 \pm 0.0129	\pm 0.0113
BG Score (Female)	0.5759 \pm 0.0010	\pm 0.0009
Span F1	0.4872 \pm 0.0008	\pm 0.0007
Token F1	0.6271 \pm 0.0040	\pm 0.0035

Table 2: Run-to-run stability of CASPAR measured over five validation runs using different random seeds (42, 13, 21, 84, and 100), and where WG = Worst-Group and BG = Base-Group.

observations of Elgaar et al. (2024). Stage 2 refinement further improves prediction performance, with a more modest gain in the Final Score.

4 Discussion

Figure 2 shows that racial subgroup disparities are associated with different error patterns. Hispanic patients exhibit the highest label confusion rate (31.69%), mainly between semantically over-

System	Final	Span F1	Token F1
RoBERTa (only)	0.4359	0.3939	0.5601
RoBERTa + CRF	0.4378	0.4065	0.5757
+ BiGRU (Stage 1)	0.4649	0.4396	0.5961
+ Stage 2 (full)	0.4998	0.4873	0.6249

Table 3: Ablation results based on the validation set of MedExACT competition benchmark.

lapping categories such as *Therapeutic Procedure* (cat 6) and *Defining Problem* (cat 3). For example, “*NG lavage with coffee ground and brown clot*” is repeatedly predicted as cat 3 despite being annotated as cat 6. African American patients show the highest under-span rate in cat 3 (27.8%), where clinically meaningful qualifiers are truncated, as in “*AFib on coumadin*” being reduced to “*AFib*”. By contrast, the *Other* group has the highest miss rate (6.35%), indicating a stronger recall problem. Overall, these results suggest that subgroup disparities in this task arise from different combinations of category ambiguity, boundary truncation, and recall loss.

Additionally, From Table. 1, the gap between token F1 score (0.64) and span F1 score (0.51) highlights a structural mismatch between token-level BIO modeling and exact-match span evaluation. As shown in Figure 3, under-span and over-span errors dominate (24.21% and 19.88%, respectively), while label confusion (overlap_diff_label) is concentrated in cat 6 (26.1%). As a result, even small boundary deviations such as retaining a trail-

Subgroup Performance

Blue: Span F1 | Green: Token F1 | Orange: Base Score

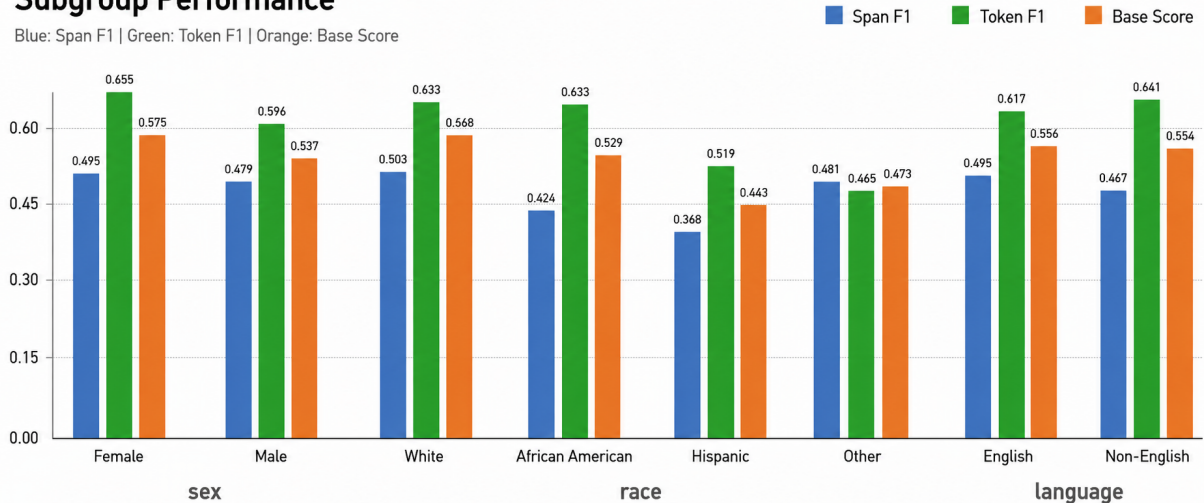


Figure 4: Subgroup-level performance across demographic groups. Token F1 consistently exceeds Span F1 for all subgroups, with the largest gap observed in Hispanic patients (Token F1: 0.519 vs. Span F1: 0.368), reflecting the challenge of exact span boundary recovery under demographic distribution shift.

ing punctuation mark or dropping a leading determiner lead to complete span-level errors. This is partly due to the BIO formulation itself, predicting $2|\mathcal{C}| + 1$ labels over a multi-class tag space increases task complexity and offers only indirect supervision for recovering span boundaries as complete units. The reasons for this observation is that the model frequently suffers from boundary expansion and category drifting. For instance, in detecting cat 6 co-occurring are often merged medications (e.g., swallowing “combivent neb” into a preceding drug phrase) or procedures are not properly isolated from their narrative context (e.g., expanding “CTA” into the entire event description). These cases demonstrate that while the model effectively locates the general vicinity of *Therapeutic Procedure* (cat 6) entities, it fails to filter out redundant contextual noise. Consequently, the model frequently conflates core denoting medical tokens procedure with adjacent narrative descriptions or co-occurring phrases, leading to imprecise boundaries and degraded span-level performance.

The CASPAR pipeline relies on BIO-based token-level prediction for span detection, which introduces an inherent mismatch with strict span-level evaluation. As shown in Figure 4, strong performance at the token level does not necessarily translate to precision in span recovery based on exact-match (EM) criteria. The *Non-English* subgroup achieves the highest Token F1 (0.641) but only a moderate Span F1 (0.467), while the *Other* subgroup attains comparable span-level per-

formance despite much weaker token-level predictions. It suggests that token-level supervision alone is insufficient for fully capturing precise span boundaries. This discrepancy arises primarily because most failed recoveries are not due to missed detections, but rather flawed span boundaries. As illustrated by the *Defining problem (C3)* category in Figure 3, which exhibits the highest rate of under-span errors, the model tends to prioritize high-confidence keywords that define the problem while neglecting the complete boundary of the span.

A complementary limitation arises from the design of Stage 2. Because training instances are constructed from Stage 1 predictions that have token-level IoU greater than 0 with a gold span, the refinement module is trained exclusively on partially correct candidates. As a result, Stage 2 cannot explicitly suppress false-positive spans produced by Stage 1, nor can it recover spans that Stage 1 missed entirely. This structural constraint provides a principled explanation for why Stage 2 yields meaningful but modest gains in the ablation study (Table 3). It improves boundary precision and category assignment within the set of retrieved candidates, but leaves recall errors inherited from Stage 1 unaddressed. A natural direction for future work is to incorporate false-positive rejection and span proposal generation directly into the refinement stage, so that Stage 2 can correct a broader range of Stage 1 errors.

5 Conclusion

We present a two-stage pipeline for medical decision span extraction and classification. Stage 1 combines a sliding-window text encoder with BiGRU and CRF to produce document-level candidate spans, while Stage 2 applies a lightweight local refinement module to correct category assignments and span boundaries per candidate. Our system achieved a final score of 0.5668. Results from the ablation study confirm that BiGRU consolidation is critical for effective CRF decoding. The error analysis shows that demographic subgroup disparities are driven by different combinations of category confusion and boundary truncation. Moreover, the persistent gap between the token-level and span-level performance highlights an inherent limitation of BIO-based formulations under exact-match evaluation. A natural direction for future work is to incorporate span-aware objectives directly into training, so that the model can optimize token classification while simultaneously learning to recover exact span boundaries more reliably. Beyond the Stage 1 formulation, extending Stage 2 to handle a broader range of candidate types, including false-positive rejection and recovery of spans missed by Stage 1 would address the structural recall ceiling identified in our analysis and potentially yield more substantial gains from the refinement stage.

References

- Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.
- Alessandro Cocchieri, Viet Dac Lai, Sukananya Purkayastha Boro, and et al. 2025. [Openbioner: Lightweight open-domain biomedical named entity recognition through entity type description](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 847–860. Association for Computational Linguistics.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024. [MedDec: A dataset for extracting medical decisions from discharge summaries](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16442–16455, Bangkok, Thailand. Association for Computational Linguistics.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Mehrnaz Sadrolashrafi, Mitra Mohtarami, Adrian Wong, Hadi Amiri, and Leo A. Celi. 2026. [Overview of medical decision extraction, analysis, and classification task \(medexact\) 2026](#). In *The 25th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, San Diego, California, USA. Association for Computational Linguistics.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence labeling. *arXiv preprint arXiv:1508.01991*.
- Richard A. A. Jonker, Tiago Almeida, Rui Antunes, João R. Almeida, and Sérgio Matos. 2024. [Multi-head crf classifier for biomedical multi-class named entity recognition on spanish clinical notes](#). *Database*, 2024:baae068.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ling Luo, Chih-Hsuan Wei, Po-Ting Lai, Robert Leaman, Qingyu Chen, and Zhiyong Lu. 2023. [Aioner: all-in-one scheme-based biomedical named entity recognition using deep learning](#). *Bioinformatics*, 39(5):btad310.
- Eirik H. Ofstad, Jan C. Frich, Edvin Schei, Richard M. Frankel, and Pål Gulbrandsen. 2016. What is a medical decision? A taxonomy based on physician statements in hospital encounters: a qualitative study. *BMJ Open*, 6(2):e010098.
- Lance Ramshaw and Mitch Marcus. 1995. [Text chunking using transformation-based learning](#). In *Third Workshop on Very Large Corpora*.
- Lev Ratnov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In

- Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155. Association for Computational Linguistics.
- João Ruano, Gonçalo Correia, Leonor Barreiros, and Afonso Mendes. 2025. [Effective multi-task learning for biomedical named entity recognition](#). In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 225–239, Vienna, Austria. Association for Computational Linguistics.
- Mario Sanger, Alan Akbik, Amir Pouran Ben Veyseh, and et al. 2024. [Hunflair2 in a cross-corpus evaluation of biomedical named entity recognition and normalization tools](#). *Bioinformatics*, 40(10):btac564.
- Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. [Locate and label: A two-stage identifier for nested named entity recognition](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2782–2794. Association for Computational Linguistics.
- Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. [Diffusioner: Boundary diffusion for named entity recognition](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7186–7201. Association for Computational Linguistics.
- Yongliang Shen, Xiaobin Wang, Zeqi Tan, Guangwei Xu, Pengjun Xie, Fei Huang, Weiming Lu, and Yueting Zhuang. 2022. [Parallel instance query network for named entity recognition](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 947–961, Dublin, Ireland. Association for Computational Linguistics.
- Minghao Tang, Yongquan He, Yongxiu Xu, Hongbo Xu, Wenyuan Zhang, and Yang Lin. 2023. [A boundary offset prediction network for named entity recognition](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14834–14846, Singapore. Association for Computational Linguistics.
- Harsh Verma, Sabine Bergler, and Narjesossadat Tahaei. 2023. [Comparing and combining some popular NER approaches on biomedical tasks](#). In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 273–279, Toronto, Canada. Association for Computational Linguistics.
- Ying Wei and Qi Li. 2023. [Scdner: Span-based consistency-aware document-level named entity recognition](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15677–15685. Association for Computational Linguistics.
- Jiuding Yang, Jinwen Luo, Weidong Guo, Di Niu, and Yu Xu. 2023. [Exploiting hierarchically structured categories in fine-grained chinese named entity recognition](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3407–3421, Toronto, Canada. Association for Computational Linguistics.
- Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2023. [Optimizing bi-encoder for named entity recognition via contrastive learning](#). In *The Eleventh International Conference on Learning Representations*.

KCL-Cogstack at PsyDefDetect: A Hierarchical Approach to Detecting Defense Mechanisms in Supportive Dialogue

Shubham Agarwal¹, Thomas Searle^{1,3}, Richard Dobson^{1,2,3}

¹Department of Biostatistics & Health Informatics, King’s College London, London, U.K.

²Institute of Health Informatics, University College London, London, U.K.

³CogStack Limited, London, U.K.

Correspondence: shubham.agarwal@kcl.ac.uk

Abstract

We present our system for the PsyDefDetect shared task, which focuses on detecting and classifying psychological defense mechanisms in peer emotional support conversations. Our core contribution is a hierarchical classification framework that structures prediction as a coarse-to-fine pipeline over a clinically validated label hierarchy, grounded in the Defense Mechanism Rating Scales (DMRS). Through systematic experimentation with flat fine-tuning, few-shot prompting, and hierarchical classification, we demonstrate that explicitly modelling the structured relationships among defense levels offers a more effective alternative to flat classification, achieving a macro F1 of 0.23 on the official test set.

1 Introduction

As digital and transcription-based health records capture increasing volumes of patient–clinician interactions, there is a growing opportunity to analyze how mental states and psychological processes manifest in language (Zimmerman et al., 2022; Drew et al., 2001). Transcripts of therapy sessions, clinical interviews, and patient communications provide rich evidence of cognitive and emotional patterns that can inform mental health assessment and care (Ewbank et al., 2021). This offers substantial opportunities to better understand patient experiences and mental states, and to facilitate timely, accurate, and psychologically informed interpretation of patient communications — essential for ensuring that individuals feel heard, supported, and engaged in their care (Drew et al., 2001). Such data can further support a wide range of mental health applications, from screening and monitoring to personalized assistance (Luxton, 2014).

Among the many phenomena expressed in clinical dialogue, psychological defense mechanisms are of particular importance. Psychological defense mechanisms are unconscious or automatic

strategies that individuals use to manage distress, reduce emotional conflict, and protect themselves from anxiety or perceived threats (Freud, 1936; Vaillant, 1992). Because these mechanisms are often expressed indirectly in language, they provide a clinically meaningful target for automatic analysis of patient text and dialogue (Cramer, 2000).

Recent work on automatically identifying defense mechanisms from text has begun to show that language models can capture clinically meaningful defensive cues in dialogue. A notable study applied RoBERTa-based models (Liu et al., 2019) to interviewer–participant talk–turn pairs from adult attachment interview transcripts and demonstrated that language models can distinguish the presence of defenses and detect several common mechanisms, including repression, intellectualization, reaction formation, and undoing (Tasca et al., 2023). This provides an important proof of concept that defensive functioning can be modeled computationally from conversational data.

The broader psychodynamic literature offers a strong theoretical foundation for this task. The Defense Mechanism Rating Scales (DMRS) (Di Giuseppe and Perry, 2021) are a clinically validated, hierarchically organised system for rating defensive maturity in clinical transcripts, ranging from immature action-based defenses to high-adaptive coping strategies. Originally developed for use by trained clinicians, the DMRS have long been used to assess defensive functioning, and recent work continues to support their relevance in contemporary assessment and computational annotation settings.

However, despite these advances, automatic recognition of defense mechanisms from text remains an underexplored problem, particularly in the setting of supportive conversations and multi-turn dialogue (Na et al., 2025). The PsyDefDetect shared task (Na et al., 2026a) addresses this gap by introducing a benchmark for detecting seven de-

fense mechanisms plus two auxiliary labels in text, thereby providing a clinically grounded testbed for future NLP methods.

2 Background

2.1 PsyDefDetect shared task

The *PsyDefDetect* shared task (Na et al., 2026a) introduces a benchmark for the detection and classification of psychological defense mechanisms in peer emotional support conversations. The task is motivated by recent work on computational modeling of defensive functioning in dialogue, which shows that transformer-based language models can distinguish the presence of defenses and identify common mechanisms from conversational data. *PsyDefDetect* extends this line of research by providing an annotated dataset of conversations explicitly labeled for defense levels, linked to a clinically grounded taxonomy based on the DMRS framework.

2.2 Dataset and setup

The dataset, PSYDEFCONV (Na et al., 2026b), is derived from a stratified subset of the ESConv corpus (Liu et al., 2021) and contains 200 dialogues and 4709 utterances, of which 2336 are help-seeker turns annotated for defensive functioning. Each instance consists of a multi-turn dialogue and a target utterance; systems must classify the target into one of nine categories: seven hierarchical DMRS defense levels, a *No Defenses* label for phatic or functional turns, and a *Needs More Information* label for utterances lacking sufficient context. By providing a structured, publicly available benchmark, *PsyDefDetect* aims to stimulate work on interpretable and clinically meaningful methods for understanding how individuals express psychological distress in language.

3 Methodology

We explore several modeling strategies for detecting psychological defense mechanisms from supportive conversations. First, we investigate standard supervised learning, where a BERT model (Devlin et al., 2019) is fine-tuned to classify text directly into the nine defense-mechanism classes. This *flat* classification approach treats all mechanisms as independent labels and does not explicitly model higher-level groupings or developmental levels.

Second, we consider few-shot prompting of Large Language Models (LLMs), where the task is cast as in-context learning rather than fine-tuning. This approach avoids the need for large annotated training sets and can capture nuanced patterns via the model’s pre-existing knowledge of language and psychology.

Finally, we propose a hierarchical classification framework grounded in clustering. Rather than treating the nine defense mechanisms as independent labels, we first group them into higher-level clusters based on semantic and developmental similarity. These clusters then structure a hierarchical classifier that first predicts the level or cluster of defensive functioning before refining the prediction to the specific mechanism. This approach reflects the theoretical structure of the DMRS taxonomy and enables more structured and interpretable modeling of the task.

Our experimentation utilized a shared university resource machine with 2 Nvidia A100 GPUs via KCL CREATE (King’s College London e-Research team, 2025).

3.1 Approaches

3.1.1 Fine-tuning BERT

We fine-tune a BERT-based encoder for defense-mechanism classification. Specifically, we use `bert-base-uncased` as the backbone model and apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to efficiently adapt the encoder to the target task. Given the limited size of the annotated dataset and the clinically nuanced nature of the labels, parameter-efficient fine-tuning is a natural choice, as LoRA introduces a small number of trainable parameters while keeping the original model weights largely frozen, reducing the risk of overfitting. A key challenge is that full dialogue context often exceeds BERT’s 512-token limit and also contains noise and less relevant turns. We therefore explore two context representation strategies:

- *last-3 turns*, where only the three most recent dialogue turns along with the target utterance are included as context
- *summarised context*, where all preceding turns except the final one are condensed into a summary using an LLM (Gemma 4B (Team, 2025)*), which is then concatenated with the final

*<https://huggingface.co/google/gemma-3-4b-it>

turn and the target utterance

Both representations are kept within the 512-token limit. For training, we use the Adam optimizer with a learning rate of $1e-4$, a dropout rate of 0.25, and gradient clipping with a maximum norm of 1.0, with a batch size of 64. The model is trained on an 80–20 train–validation split of the official training data.

3.1.2 Prompting LLM

We leverage in-context learning via few-shot prompting, providing a structured prompt that describes the task, introduces the nine labels with brief definitions for each class, and includes a small set of example instances, each containing a dialogue segment and its corresponding ground-truth label. We implement this using MedGemma-4B (Sellersgren et al., 2025), a model optimized for clinical and health-related language, which we hypothesize may better capture the nuanced psychological language present in emotional support conversations. Prompt design was iteratively refined based on empirical performance. Full prompt templates are provided in Appendix A.4.

3.1.3 Hierarchical classification

In addition to flat fine-tuning and few-shot prompting, we explore a hierarchical classification strategy designed to reflect the structured nature of the defense-mechanism space. Rather than treating all nine labels as independent classes, we first group them into higher-level clusters before performing final classification.

To construct these groupings, we apply spectral clustering (Von Luxburg, 2007) over sentence embeddings of the training utterances using an RBF kernel over the normalised graph Laplacian to obtain label-level clusters. The resulting groupings were validated by a psychiatrist at South London and Maudsley NHS Foundation Trust (SLaM) for clinical plausibility.

The resulting hierarchy, summarized in Appendix A.2 (Table 2), organizes the nine labels into three levels of increasing specificity. This coarse-to-fine pipeline exploits both the natural hierarchy of defensive functioning and the clinically meaningful relationships among the labels, providing a more structured alternative to flat classification. Here as well we use bert-base-uncased as the classification model. The pipeline consists of a separate BERT models trained for each node in the hierarchy, with predictions at each level routing the

input to the appropriate child classifier. Errors at higher levels propagate to subsequent stages, as no recovery mechanism is employed.

3.2 Evaluation

We evaluate all systems using F1-score as the primary metric, following the official PsyDefDetect evaluation protocol. We additionally report accuracy and weighted F1-score to provide a more complete picture of model behaviour across the nine categories. We report results on both the validation set (20% of the dev set), and the official test set evaluated through the CodaBench platform. Validation results are used to guide model selection and hyperparameter tuning, while test set results reflect final system performance under the shared task evaluation conditions.

4 Results

4.1 Fine-tuned BERT

The fine-tuned BERT model achieves moderate performance on both the validation and test sets. On the validation split, it obtains a macro F1 of 0.19. On the official test set, performance drops to a macro F1 of 0.15, suggesting some degree of overfitting to the development split. These results suggest that while standard supervised fine-tuning provides a reasonable starting point, it struggles to fully capture the nuanced distinctions between defense mechanisms.

4.2 LLM Prompting

The prompting-based approach performs worse than fine-tuned BERT on both splits. On the entire development set (no training required), it reaches a macro F1 of 0.12, and on the official test set, performance is comparable at 0.13. This indicates that few-shot prompting alone is not sufficient for reliably identifying defense mechanisms in this task, likely because the label space is clinically subtle and the examples are highly heterogeneous.

4.3 Hierarchical BERT

Hierarchical BERT is the strongest-performing approach across both validation and test evaluations. On the validation split, it achieves a macro F1 of 0.24, achieving the rank of 17/21 registered teams, outperforming both the fine-tuned BERT and LLM prompting approaches. On the test set, it again ranks highest with an F1 of 0.23. These results suggest that explicitly modelling the structure among

Table 1: Overall performance across approaches on the validation and test sets

Model	<i>Micro</i>			<i>Macro</i>			<i>Weighted</i>		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
<i>Validation set performance</i>									
Fine-tuned BERT	0.38	0.38	0.38	0.18	0.19	0.19	0.33	0.38	0.35
LLM *	0.31	0.31	0.31	0.14	0.13	0.12	0.31	0.35	0.32
Hierarchical BERT	0.47	0.48	0.47	0.26	0.24	0.24	0.48	0.61	0.52
<i>Test set performance</i>									
Model	Recall			Precision			F1		
Fine-tuned BERT	0.17			0.26			0.15		
LLM	0.16			0.17			0.13		
Hierarchical BERT	0.25			0.29			0.23		

* Evaluation results on entire dev set

defense mechanisms is beneficial and that hierarchical classification offers a more effective way to represent the label space than flat prediction.

Level-wise evaluation, provided in Appendix A.3 (Table 3), reveals that the hierarchy performs strongly at coarse discrimination but degrades at finer granularities, suggesting that distinguishing between specific mechanisms within a cluster remains the key challenge. Additionally, the confusion matrices for all approaches are also provided in Appendix A.1.

5 Discussion

Our approach to the PsyDefDetect shared task evolved through several iterations, each motivated by the limitations of the previous. We began with a flat fine-tuned BERT classifier, which provided a reasonable starting point but exposed a key weakness: the model heavily biased toward frequent classes, particularly *No Defenses* and *High-Adaptive Defenses*, while failing entirely to predict the *Needs More Information* label. This suggests that standard supervised fine-tuning struggles to handle the class imbalance inherent in the PSYDEFCONV dataset.

We then evaluated a few-shot prompting approach using MedGemma-4B, and while this avoided the need for fine-tuning, it proved even more prone to prediction collapse, effectively reducing the label space to a handful of classes and ignoring clinically distinct mechanisms such as *Minor Image-Distorting* and *Neurotic* defenses entirely. This indicates that few-shot prompting alone is insufficient for a task requiring fine-grained,

context-dependent clinical judgment.

These observations motivated our hierarchical classification framework, which explicitly models the structured relationships among defense levels rather than treating all labels as independent. By introducing coarse-to-fine prediction grounded in a clinically validated label hierarchy, this approach achieved the strongest performance across both development and test evaluations. Notably, the hierarchical approach produced more balanced predictions across the label space compared to flat classification, though rare and clinically subtle classes remained challenging across all systems.

Compared to the best results reported in the dataset paper (Na et al., 2026b), where a macro F1 of 0.31 is achieved, our results highlight the difficulty of this task and suggest that further gains may require richer contextual representations, larger models, or more sophisticated handling of class imbalance.

6 Conclusion & Future Work

This paper presents a hierarchical classification framework for detecting psychological defense mechanisms in the PsyDefDetect shared task, incorporating LLM-based dialogue summarisation to handle long conversational contexts.

By structuring prediction as a coarse-to-fine pipeline over a psychiatrist-validated label hierarchy, our approach achieves a F1 of 0.23 on the official PsyDefDetect test set, outperforming both flat fine-tuning and few-shot prompting baselines. The results demonstrate that explicitly modelling the structured relationships among defense levels

offers a more effective and interpretable alternative to flat classification for this clinically nuanced task.

Our future work will focus on addressing class imbalance and exploring domain-adapted mental health language models to better capture subtle defensive cues.

Limitations

Our work presents a hierarchical classification framework for the PsyDefDetect shared task, evaluated on both a development split and the official unseen test set. However, the proposed system is limited in a number of ways.

First, all three approaches struggle with class imbalance in the PSYDEFCONV dataset, where High-Adaptive Defenses account for over 50% of annotated turns. This leads to prediction bias toward frequent classes and poor performance on rare but clinically important categories such as Major Image-Distorting and Neurotic defenses.

Second, the dataset is relatively small, comprising 200 dialogues drawn from a single corpus of peer emotional support conversations. This limits the generalisability of our findings to other clinical settings, such as formal psychotherapy or psychiatric interviews, where defensive functioning may manifest differently.

Third, the hierarchical pipeline chains independent classifiers with hard decisions at each level, meaning that misclassifications at coarser levels propagate to finer levels with no recovery mechanism. Future work could explore soft or joint training approaches to mitigate this.

Finally, deployment of such systems in real clinical environments would require careful attention to data privacy, regulatory compliance (e.g. GDPR), and the ethical implications of automated psychological assessment, particularly given the sensitivity of mental health data.

Acknowledgments

This work was supported by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. SA, TS, RD are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RD is also supported by The National Institute for Health

Research University College London Hospitals Biomedical Research Centre. This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- Phebe Cramer. 2000. Defense mechanisms in psychology today: Further processes for adaptation. *American psychologist*, 55(6):637.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Mariagrazia Di Giuseppe and J Christopher Perry. 2021. The hierarchy of defense mechanisms: Assessing defensive functioning with the defense mechanisms rating scales q-sort. *Frontiers in psychology*, 12:718440.
- Paul Drew, John Chatwin, and Sarah Collins. 2001. Conversation analysis: a method for research into interactions between patients and health-care professionals. *Health Expectations*, 4(1):58–70.
- MP Ewbank, R Cummins, V Tablan, A Catarino, S Buchholz, and AD Blackwell. 2021. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, 31(3):300–312.
- Sigmund Freud. 1936. Inhibitions, symptoms and anxiety. *The Psychoanalytic Quarterly*, 5(1):1–28.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.
- King's College London e-Research team. 2025. [King's computational research, engineering and technology environment \(CREATE\)](#).
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting*

- of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pages 3469–3483.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- David D Luxton. 2014. Artificial intelligence in psychological practice: Current and future applications and implications. *Professional Psychology: Research and Practice*, 45(5):332.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cian Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.
- Anthony N Tasca, Samantha Carlucci, James C Wiley, Matthew Holden, Ahmed El-Roby, and Giorgio A Tasca. 2023. Detecting defense mechanisms from adult attachment interview (aai) transcripts using machine learning. *Psychotherapy Research*, 33(6):757–767.
- Gemma Team. 2025. [Gemma 3](#).
- George E Vaillant. 1992. *Ego mechanisms of defense: a guide for clinicians and researchers*. American Psychiatric Pub.
- Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.
- Marc Zimmerman and 1 others. 2022. [Linguistic measures of psychological distance track symptom levels and treatment progress in psychotherapy](#). *Proceedings of the National Academy of Sciences*.

A Appendix

A.1 Confusion Matrices

A.1.1 Fine-tuned BERT

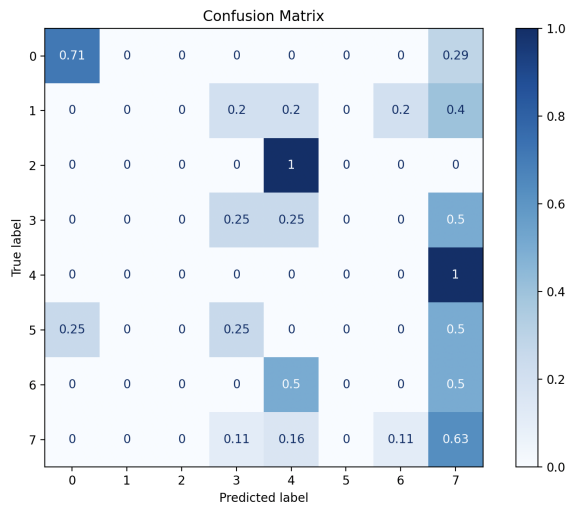


Figure 1: Confusion matrix for fine-tuned BERT

*Since class 8 wasn't predicted, it's not shown in the confusion matrix

A.1.2 LLM

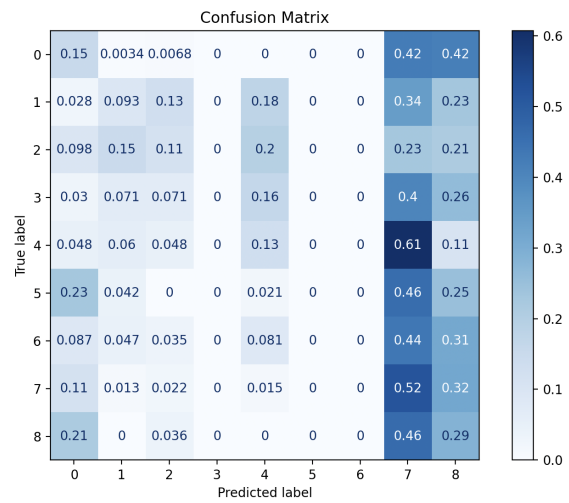


Figure 2: Confusion matrix for LLM

A.1.3 Hierarchical classifier

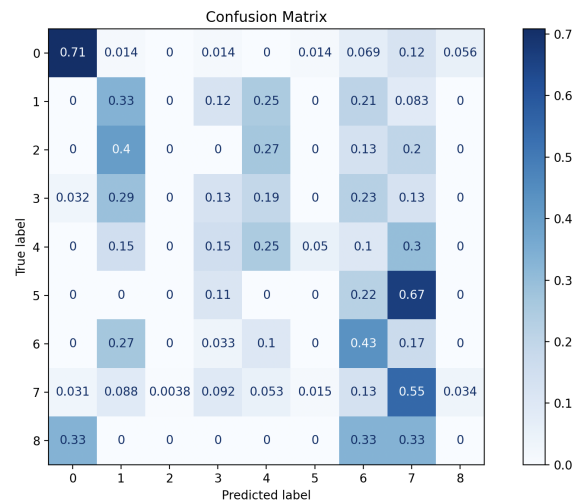


Figure 3: Confusion matrix for hierarchical classifier

A.2 Hierarchical label clustering

Table 2: Hierarchical label clustering used in the three-level classification pipeline

Level	Cluster	Classes
Level 1	Auxiliary	No Defenses, Needs More Info
	High-Adaptive	High-Adaptive Defenses
	Maladaptive	Action, Major Image Distorting, Disavowal, Minor Image Distorting, Neurotic, Obsessional
Level 2	Auxiliary → N/A (terminal)	No Defenses, Needs More Info
	Maladaptive → Externalizing / Image-Distorting	Action, Major Image Distorting, Minor Image Distorting
	Maladaptive → Avoidant / Internal-Regulating	Disavowal, Neurotic, Obsessional
Level 3	Externalizing / Image-Distorting → N/A (terminal)	Action, Major Image Distorting, Minor Image Distorting
	Avoidant / Internal-Regulating → N/A (terminal)	Disavowal, Neurotic, Obsessional

A.3 Hierarchical classification level-wise results

Table 3: Level-wise results for hierarchical classifier

Level	Recall	Precision	F1
<i>Validation set performance</i>			
Level 1	0.56	0.6	0.56
Level 2 - Auxiliary	0.71	0.97	0.79
Level 2 - Maladaptive	0.71	0.7	0.71
Level 3 - Externalizing	0.56	0.6	0.57
Level 3 - Avoidant	0.53	0.53	0.5

A.4 Prompt for LLM

```

"""
<bos><start_of_turn>user You are a emotional conversation expert. Your task is to classify the given
conversation into one of the below given classes.

---
# Instructions:
1) The input will include a transcript of what the seeker is saying to the supporter.
2) Classify these into the given classes by taking into account information from the text, you
will have to infer it, the information won't be directly present.
3) Give reasoning on why you have chosen the class, explaining why it makes sense and pointing
out key evidence.
4) The output must be 'label' and 'reasoning' as the two keys in a json format.
5) Respond only with the label and reasoning, do not include additional commentary.

---
# Classes:
0: No Defenses; Mechanism: N/A; Definition: Functional utterances that maintain conversational
flow without engaging conflict.
1: Action Defenses; Mechanism: Passive Aggression, Help-Rejecting Complaining, Acting Out;
Definition: Distress is released by acting on the environment instead of reflecting.
2: Major Image-Distorting; Mechanism: Splitting (self/other), Projective Identification;
Definition: Reduces anxiety via all-good/all-bad distortions of self or other.

```

- 3: Disavowal Defenses; Mechanism: Denial, Rationalization, Projection, Autistic Fantasy;
Definition: Rejects threatening reality by denying, excusing, blaming, or fantasizing.
- 4: Minor Image-Distorting; Mechanism: Devaluation/Idealization (self or other), Omnipotence;
Definition: Softer distortions temporarily inflate or deflate self-esteem.
- 5: Neurotic Defenses; Mechanism: Repression, Dissociation, Reaction Formation, Displacement;
Definition: Keeps unacceptable motives out of awareness; feelings surface indirectly.
- 6: Obsessional Defenses; Mechanism: Isolation of Affect, Intellectualization, Undoing; Definition:
Uses excessive logic or symbolic acts to separate feelings from events.
- 7: High-Adaptive Defenses; Mechanism: Affiliation, Altruism, Anticipation, Humor, Self-Assertion,
Self-Observation, Sublimation, Suppression; Definition: Mature coping that integrates
emotion and thought to channel affect constructively.
- 8: Needs More Information; Mechanism: N/A; Definition: Label used when an utterance is too
ambiguous or lacks context.

Question: Classify this text for me?

Input: {text}

<end_of_turn> <start_of_turn>model

"""

DAL Team at PsyDefDetect: From Supervised Encoders to Hierarchical LLM-RAG for Psychological Defense Detection

Duc-Luong Tran¹, Phuong-Anh Chu¹, Hoang-Dat Do¹, Tu-Phuong Mai¹,
Duy-Cat Can^{1,2,3}, Hoang-Quynh Le^{1*}

¹Faculty of Information Technology, VNU University of Engineering and Technology, Vietnam

²Platform of Bioinformatics, Lausanne University Hospital, Switzerland

³Faculty of Biology and Medicine, University of Lausanne, Switzerland

{22021148, 23020324, 24020060, 21020552, 1hquynh}@vnu.edu.vn
duy-cat.can@chuv.ch

Abstract

The PsyDefDetect shared task focuses on classifying nine psychological defense mechanisms in multi-turn dialogues, a problem complicated by severe label imbalance and the implicit, context-dependent nature of defenses. In this work, we investigate several approaches for dialogue-level defense detection, including supervised baselines and large language model (LLM)-based pipelines. Our primary system is a retrieval-augmented LLM framework with hierarchical prediction and lightweight heuristics for decision calibration. Experiments on the PSYDEFCONV dataset show that LLM-based methods improve overall performance compared to supervised baselines, but still struggle with fine-grained distinctions, especially for minority labels. These findings highlight the challenges of modeling implicit psychological constructs in dialogue.

1 Introduction

Psychological defense mechanisms are automatic, unconscious processes that regulate responses to emotional conflict (Perry, 2014). They influence both clinical outcomes (Mrozowicz-Wrońska, 2023) and user behaviors in emotional support conversations (ESC) (Na et al., 2026b; Di Giuseppe et al., 2024). While NLP has successfully modeled explicit constructs like empathy and sentiment (Shetty et al., 2024; Liu et al., 2021), psychological defenses remain challenging because they are highly implicit, pragmatic, and context-dependent (Na et al., 2025). Unlike emotions with distinct lexical triggers, defensive behaviors evolve dynamically across multi-turn dialogues, rendering conventional single-utterance classification insufficient.

To bridge the gap between clinical theory and NLP, the PsyDefDetect shared task (Na et al.,

2026a) introduces a framework for classifying defensive functioning in multi-turn dialogues using the DMRS taxonomy (Di Giuseppe and Perry, 2021). Beyond the latent nature of the task, automated detection on the PSYDEFCONV corpus faces two major structural challenges: (i) severe label imbalance driven by the natural prevalence of mature defenses, and (ii) long, noisy conversational histories that make contextual modeling difficult, as dense and extensive contexts often trigger information degradation in language models (Liu et al., 2024).

These challenges expose limitations of conventional supervised models, whose reliance on limited and imbalanced training data prevents them from capturing the subtle, context-dependent pragmatics of implicit defenses. More importantly, they highlight a deeper mismatch between the nature of psychological defenses and the assumptions of flat classification models. To address this, we leverage the external knowledge encoded in Large Language Models (LLMs) through a retrieval-augmented framework for dialogue-level inference.

Our primary approach is a retrieval-augmented LLM pipeline that operates at the dialogue level. It decomposes predictions into a coarse-to-fine hierarchy and integrates lightweight heuristics to improve robustness. In addition, we incorporate hybrid supervised-LLM filtering to handle frequent labels and a summary-based distillation module to compress noisy conversational contexts.

2 Related Work

Prior NLP work has mainly focused on explicit affective signals such as sentiment and empathy, which are often modeled with sequence classification (Rashkin et al., 2019; Sharma et al., 2020; Liu et al., 2021). More recent studies have moved toward more latent cognitive constructs, such as cognitive distortions (Maddela et al., 2023; Chen

* Corresponding author.

et al., 2023). Psychological defense mechanisms are even more implicit, pragmatic, and context-dependent (Na et al., 2025; Perry, 2014), especially in multi-turn dialogue where meaning unfolds across turns. This makes single-utterance classification insufficient and increases the need for dialogue-level modeling. In such settings, supervised models trained on limited and imbalanced data often struggle to capture the functional role of an utterance in context.

Large language models (LLMs) have shown strong potential in mental health applications (Yang et al., 2023; Xu et al., 2024) and related clinical language tasks (Galatzer-Levy et al., 2023). Their broad linguistic and world knowledge makes them attractive for low-data, context-dependent problems. However, prior work in psychotherapy-related NLP often formulates prediction as static label assignment over isolated instances (Tu et al., 2024; Bao et al., 2024), which is not well aligned with the dynamic and hierarchical nature of psychological processes (Na et al., 2025; Di Giuseppe and Perry, 2021). These limitations motivate more structured, dialogue-aware LLM pipelines for defense detection.

Retrieval-Augmented Generation (RAG) is widely used to ground LLM predictions by providing relevant examples or contextual evidence (Gao et al., 2024). In dialogue understanding, retrieval can improve prediction stability, but semantic similarity alone may be insufficient for latent psychological constructs, where similar utterances can serve different pragmatic functions depending on context (Bender and Koller, 2020). Likewise, heuristic cues may offer high-precision signals for explicit patterns, yet often generalize poorly to subtle or evolving behaviors (Chancellor and De Choudhury, 2020), motivating retrieval designs that incorporate richer discourse-level context and lightweight calibration.

3 Methodology

We tackle PsyDefDetect under two key challenges: severe label imbalance and the implicit, context-dependent nature of psychological defenses. In addition, the task involves multi-turn dialogues, where defensive behaviors are often expressed through subtle pragmatic cues and depend heavily on preceding context, making single-utterance classification insufficient.

To address these challenges, we investigate

four methodological approaches. *Approach 1* establishes supervised dialogue-level baselines. *Approach 2*, our core contribution, develops a retrieval-augmented LLM pipeline to utilize the internal clinical knowledge of LLMs for identifying latent psychological patterns. Building upon this, *Approach 3* extends the framework into a hybrid supervised-LLM pipeline. Finally, *Approach 4* introduces an auxiliary summary-based distillation strategy.

3.1 Approach 1: Supervised Encoders (Baseline)

We establish a supervised baseline using a RoBERTa-LSTM dialogue encoder, where each utterance from both seeker and helper is first encoded via RoBERTa, and an LSTM aggregates these sequential representations into a dialogue-level embedding for classification. This hierarchical encoding is particularly suited to PsyDefDetect, as defense mechanisms can dynamically shift across turns and thus require interpreting each utterance in relation to its preceding context rather than in isolation.

On top of this encoder, we explore three prediction strategies:

- *Single-level classifier*: Directly predicts the final label from the full label set without any hierarchical decomposition.
- *2-level hierarchical classifier*: First predicts whether the instance belongs to label 7 or not, then classifies the remaining labels.
- *3-level hierarchical classifier*: Sequentially predicts label 0, then label 7, and finally classifies the remaining labels.

Motivated by extreme label imbalance, specifically the dominance of easily detectable labels 7 (51.93%) and 0 (15.88%), these hierarchical variants employ a sequence of strictly independent models. During inference, these models are applied sequentially: if an early stage predicts a frequent label, the process halts and outputs it; otherwise, the instance is passed to the next classifier. This early-exit strategy effectively isolates majority classes early on, allowing subsequent stages to focus entirely on more challenging and underrepresented minority classes.

3.2 Approach 2: Retrieval-Augmented LLM Inference

This is our main approach. The task is inherently difficult: labels are often implicit and expressed through complex discourse functions, while the dataset is small and highly imbalanced. Consequently, standard supervised models struggle to generalize, especially on minority classes. To address these challenges, we propose two retrieval-augmented inference strategies using the Gemini 3.1 Flash Lite model: a Direct LLM-RAG that predicts the final label in a single step, and a Hierarchical LLM-RAG that decomposes the task into coarse-grained grouping followed by fine-grained classification. Between these, we identify the hierarchical approach as our primary and most effective strategy, the complete architecture of which is illustrated in Figure 1.

Prompt construction and retrieval. We enhance the target dialogue with task-specific knowledge and contextual evidence through the following three-step pipeline:

- *Dialogue reconstruction:* We group samples by `dialogue_id` to recover their original conversation context. In PsyDefDetect, both training and test instances are truncated histories of larger dialogues, where labels may vary across turns. Reconstructing these partial dialogues enables the model to capture evolving context rather than treating samples as independent.
- *Generating explanations:* We use the LLM (Gemini 3.1 Flash Lite) to generate short explanations for labeled turns, turning each fragment into a (label + rationale) retrieval unit.
- *Retrieval:* For each target turn, we retrieve $k = 3$ dynamic dialogue fragments. Rather than using a neural embedding model, we utilize a domain-grounded symbolic matching approach. Local contextual similarity and lexical overlap are measured using Jaccard Similarity over preprocessed content tokens. The overall similarity score is a weighted linear combination of overlaps from the target utterance, preceding supporter/seeker turns, broader context, and a defined set of 17 rule-extracted discourse cues. To ensure context diversity and avoid redundancy, we employ a Maximal Marginal Relevance (MMR) selection procedure (Carbonell and Goldstein,

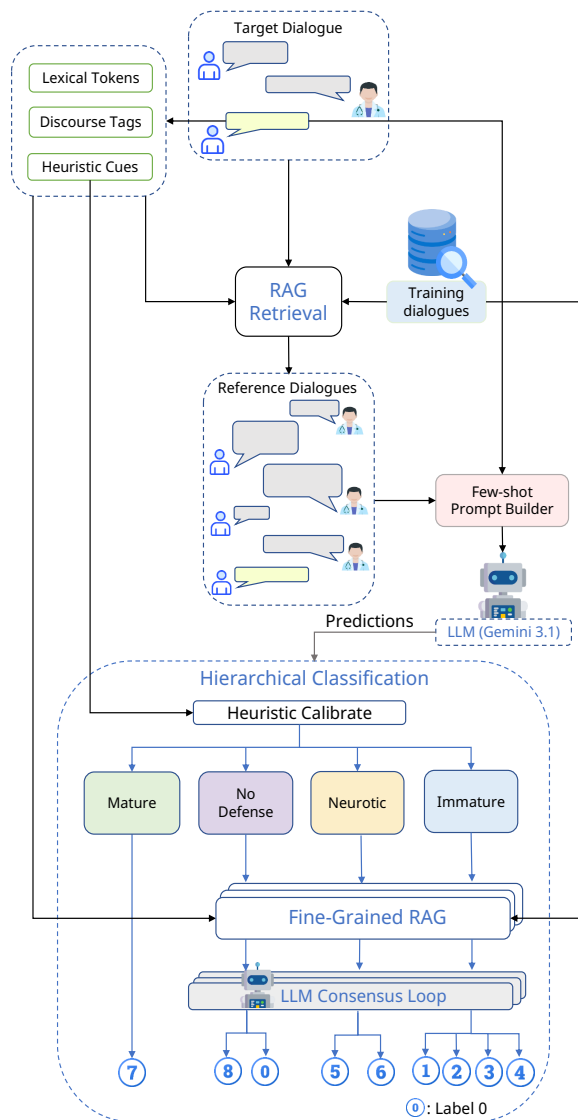


Figure 1: Architecture of the Hierarchical Retrieval-Augmented LLM Inference pipeline.

1998) with a diversity factor $\lambda = 0.20$. These dynamic examples are then merged with static class-representative anchor examples, capped at a maximum of 6 examples per prompt.

The final prompt contains reconstructed context, retrieved examples, and their explanations.

Inference strategies. On top of this pipeline, we explore two strategies using the same Gemini 3.1 Flash Lite model:

- *Direct LLM-RAG:* Directly predict the final label from the constructed prompt.
- *Hierarchical LLM-RAG:* Motivated by the severe label imbalance and the principle that decomposing complex tasks better elicits the

knowledge of LLMs (Wei et al., 2022), we decompose prediction into two sequential steps (illustrated in Figure 1): (i) predict a coarse group, and (ii) predict the final label within that group. We define four coarse-grained categories based on the DMRS taxonomy—*No Defense* (labels 0, 8), *Mature* (label 7), *Neurotic* (labels 5–6), and *Immature* (labels 1–4)—reflecting the skewed distribution in PSYDEFCONV (Na et al., 2026b). This reduces competition across semantically distant labels and focuses fine-grained prediction on a smaller, more coherent subset of classes.

Heuristic calibration. While LLMs excel at implicit semantics, they can over-analyze simple utterances or hallucinate. To ground the model, hard-coded heuristics based on 17 discourse cues are used to calibrate or override its predictions:

- *Rule 1 (Phatic/Logistical → Labels 0 or 8):* If the utterance is dominated by *phatic*, *gratitude*, or *closing* cues (e.g., "thank you", "good-bye") without substantive defense signals, it is strictly routed to *No Defense* (Label 0) or *Needs More Information* (Label 8) if exceptionally short (≤ 2 tokens).
- *Rule 2 (Action Defenses → Label 1):* If the utterance strongly triggers *barrier* or *rejecting_suggestion* cues immediately following a supporter’s *suggestion* turn, predictions are calibrated toward *Action Defenses* (Label 1).
- *Rule 3 (Neurotic Overrides → Labels 5 or 6):* If the LLM predicts *No Defense*, but the text contains *abstract_control* or a mix of *fact_logistic* and *affect* tokens, the coarse group is explicitly overridden to *Neurotic Defenses*.
- *Rule 4 (Consensus Fallback):* If the fine-grained LLM fails to output a valid label within the constrained taxonomy after maximum retries, the system completely overrides the LLM and defaults to the highest-weighted heuristic label.

On the development set, we observed that these overrides are most frequently triggered under two conditions: when the LLM over-analyzes short conversational turns by assigning complex defensive mechanisms (resolved by Rule 1), and when the model fails to produce a valid output within the allowed taxonomy (resolved by Rule 4).

3.3 Approach 3: Hybrid Supervised-LLM Pipeline

Building on Approach 2, we further combine the LLM-based method with supervised filtering from Approach 1. Our design is based on two observations:

- *Frequent, easy labels (0 and 7):* These labels are abundant and relatively simple, so supervised models trained on gold labels can learn them reliably. Thus, we use the model from Approach 1 to filter them first.
- *Rare, hard labels:* The remaining labels are fewer and more ambiguous, making them difficult to learn from limited data. Hence, we apply the LLM-RAG pipeline to leverage external knowledge and avoid bias toward dominant labels.

Based on these insights, we design a hybrid pipeline: a supervised model trained on gold labels first filters samples predicted as 0 or 7, and the LLM-RAG pipeline is applied only to the remaining cases.

Although this hybrid design is conceptually appealing and better aligns model capacity with label difficulty, our experiments show that the pure LLM-RAG pipeline (Approach 2) remains more robust overall.

3.4 Approach 4: Summary Distillation (Auxiliary)

As an auxiliary direction, we explore whether LLMs can improve supervised classification indirectly via representation learning.

Motivated by the long and noisy nature of dialogue context, and the known tendency of language models to lose critical information within the intermediate parts of extended sequences (Liu et al., 2024), we use the *Phi-4 14B* model to generate a short, task-oriented summary of the preceding context for each target turn.

- Focus on the *target utterance* rather than the full dialogue.
- Remain *grounded in the dialogue context*.
- Describe the *functional role* of the utterance.

The resulting summary captures the main stressor, the role of the utterance, and its relation to prior

Method	Acc.	Prec.	Rec.	F1
Supervised Encoders				
Single-level	58.69	18.27	14.08	13.27
2-level hierarchical	61.23	26.63	22.43	22.92
3-level hierarchical	59.11	27.53	26.51	26.49
LLM-RAG				
Direct	58.69	31.58	24.44	26.55
Hierarchical	48.31	41.87	27.73	31.13
Hybrid Pipeline				
Filter-0,7 + RAG	58.26	28.59	30.19	26.83
Summary Distillation				
Phi-4 summary	52.14	25.13	23.65	23.52

Table 1: Experimental results on the test set (%). F1 is macro-averaged over all positive classes (all classes except label 0).

context in a compact form. A RoBERTa classifier is then trained on these summaries.

This direction treats the LLM as a task-aware compressor, producing a concise and discriminative representation for downstream classification.

4 Experimental Results

Our official submission to the PsyDefDetect shared task corresponds to *Approach 2 (Hierarchical LLM-RAG)*, which we identify as our primary system based on development set performance. With this configuration, our system achieved an official rank of 10 out of 21 participating teams on the shared task leaderboard.

Table 1 shows that the *hierarchical LLM-RAG* is the strongest overall approach, achieving the best Macro-F1 (31.13%) and Precision (41.87%). This confirms that coarse-to-fine prediction is more effective than direct classification for PsyDefDetect, where labels are highly imbalanced and often implicitly expressed.

Among the supervised models, the single-level classifier performs worst, while hierarchical filtering consistently improves results. In particular, the 2-level model achieves the best Accuracy (61.23%), and the 3-level model further improves Macro-F1 to 26.49%, showing that separating frequent labels before final prediction is beneficial.

A similar pattern appears for LLM-based methods. The dialogue-level RAG model already surpasses all supervised baselines in Macro-F1, and its hierarchical extension yields a further clear gain. Although the hybrid pipeline achieves the best Recall (30.19%), it remains below the pure hierarchi-

cal LLM-RAG in Macro-F1. This suggests that early supervised filtering may improve coverage but also propagates errors. Finally, the summary distillation approach improves over the weakest supervised baseline but remains less effective than direct LLM-RAG inference.

As observed in Table 1, there is a clear trade-off between overall accuracy and balanced class performance. Although the hierarchical LLM-RAG framework improves Macro-F1 from 26.55% to 31.13%, its Accuracy decreases from 58.69% to 48.31%. This improvement is primarily driven by better recognition of minority classes, especially Class 8, whose F1 score increases from 0.00% to approximately 60% after the hierarchical reformulation. Additional gains are also observed for Class 5 (+7 points) and Class 2 (+6 points). In contrast, the reduction in Accuracy is largely caused by lower performance on Class 7 (-12 points), which constitutes the majority of the test set (243 out of 397 samples, approximately 61%). Overall, these findings suggest that the hierarchical formulation shifts the model toward more balanced predictions across classes, sacrificing some performance on the dominant class in exchange for substantially improved detection of rare defense mechanisms under severe class imbalance.

5 Conclusion

In this work, we investigated four directions for psychological defense detection under severe label imbalance and the implicit, context-dependent nature of defensive behaviors. Our core method is a retrieval-augmented LLM pipeline that operates at the dialogue level, with additional hybrid and summary-distillation extensions.

Experiments show that retrieval-augmented LLM methods outperform supervised baselines, but fine-grained defense detection remains challenging, especially for less frequent labels and subtle discourse-level distinctions. Although the hybrid design is conceptually appealing, the LLM-RAG pipeline remains the most robust overall.

These findings highlight both the promise and the current limitations of LLM-based methods for modeling latent psychological constructs in dialogue. Future work should focus on stronger pragmatic grounding and richer context modeling for fine-grained clinical language understanding.

Limitations

Our framework has three notable limitations. First, the hierarchical routing strategy improves minority-class performance but degrades accuracy on the dominant class (Class 7, ~60% of samples). This occurs because errors made at the coarse-grained grouping stage misroute majority-class instances into wrong categories, which heavily damages overall accuracy due to the severe class imbalance. Second, the heuristic calibration layer — relying on 17 discourse cues — achieves high precision on surface-level patterns but lacks semantic flexibility, failing to catch implicit defense mechanisms without clear lexical markers. Finally, our reliance on a closed-source commercial LLM limits data privacy and long-term reproducibility, restricting its immediate use in strict clinical environments where open-weight, local alternatives are preferred.

References

- Eliseo Bao, Anxo Pérez, and Javier Parapar. 2024. Explainable depression symptom detection in social media. *Health Information Science and Systems*, 12(1):47.
- Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.
- Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.
- Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.
- Zhiyu Chen, Yujie Lu, and William Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304.
- Mariagrazia Di Giuseppe, Katie Aafjes-van Doorn, Vera Békés, Bernard S Gorman, Karl Stukenberg, and Sherwood Waldron. 2024. Therapists’ defense use impacts their patients’ defensive functioning: a systematic case study. *Research in Psychotherapy: Psychopathology, Process, and Outcome*, 27(2):797.
- Mariagrazia Di Giuseppe and J. Christopher Perry. 2021. [The hierarchy of defense mechanisms: Assessing defensive functioning with the defense mechanisms rating scales q-sort](#). *Frontiers in Psychology*, 12.
- Isaac R. Galatzer-Levy, Daniel McDuff, Vivek Nataraajan, Alan Karthikesalingam, and Matteo Malgaroli. 2023. [The capability of large language models to measure psychiatric functioning](#). *arXiv preprint arXiv:2308.01834*.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. [Retrieval-augmented generation for large language models: A survey](#). *Preprint*, arXiv:2312.10997.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pages 3469–3483.
- Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13641–13660.
- Marta Mrozowicz-Wrońska. 2023. Defense mechanisms in affective disorders—the state of the art. *Psychiatria Polska*, 57(1):197–206.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you

- only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.
- J Christopher Perry. 2014. Anomalies and specific functions in the clinical identification of defense mechanisms. *Journal of clinical psychology*, 70(5):406–418.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.
- Vishal Anand Shetty, Shauna Durbin, Meghan S Weyrich, Airín Denise Martínez, Jing Qian, and David L Chin. 2024. A scoping review of empathy recognition in text using natural language processing. *Journal of the American Medical Informatics Association*, 31(3):762–775.
- Sichang Tu, Abigail Powers, Natalie Merrill, Negar Fani, Sierra Carter, Stephen Doogan, and Jinho D Choi. 2024. Automating ptsd diagnostics in clinical interviews: Leveraging large language models for trauma assessments. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 644–663.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.
- Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 8(1):1–32.
- Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyang Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077.

CUAMC @ MedExACT 2026: Robust Ensemble Voting for Fair Medical Decision Extraction

William A Baumgartner Jr and Lisa M Schilling

Department of Medicine | Division of General Internal Medicine

University of Colorado Anschutz Medical Campus

Correspondence: william.baumgartner@cuanschutz.edu

Abstract

Automated extraction of medical decisions from clinical notes is a critical step to constructing more granular patient health trajectories than what is currently obtainable from structured healthcare data. Here we present a system designed for the MedExACT shared task that employs an ensemble of BERT-based classifiers to account for demographic diversity when extracting mentions of medical decisions from MIMIC-III discharge summaries. A simple voting strategy combined with architectural diversity is demonstrated to work best when training data is limited.

1 Introduction

Structured clinical data detail the outcome of medical decisions by logging discrete observable events and state changes in a patient's health trajectory in the form of diagnoses, medications, laboratory measurements, procedures, and other data. These structured data, however, do not explicitly indicate the reasoning behind the medical decisions that led to the logged observable events. While simple reasoning for some medical decisions can be inferred from structured data (e.g., a new prescription following an abnormal lab measurement), intricacies such as patient-specific context (e.g., patient unable to tolerate side effect X), justifications based on differential diagnoses (e.g., likely Y, given Z, despite A), guideline-based decisions (e.g., skipping B because of C), and temporal aspects of decisions (e.g., pause in treatment due to D) are not explicitly captured. Clinical notes, however, often provide extensive information about medical decisions, including explicit statements about the reasoning for specific medical decisions. The ability to mine clinical notes for mentions of medical decisions represents the first step in automating the comprehensive understanding of a patient's health trajectory.

The MedDec corpus (Elgaar et al., 2024a,b) was designed to facilitate development and evaluation of automated systems for identifying mentions of medical decisions in clinical notes. MedDec supplements a subset of discharge summaries from the MIMIC-III corpus (Johnson et al., 2016b,a) with manually-defined span-level annotations labeled with ten medical decision categories from the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM) (Ofstad et al., 2016). The MedDec corpus defines "medical decision" broadly as "a particular course of clinically relevant actions and/or a statement concerning the assessment of a patient's health as defined in the DICTUM" (Elgaar et al., 2024a). This broad definition provides wide coverage for aspects of medical decision making, but also challenges automation of this extraction task due to the potential for overlapping categories (e.g., "there is no pericardial effusion" is annotated using "Defining problem" and "Evaluating test result", 76% and 24% of the time, respectively).

The Medical Decision Extraction, Analysis, and Classification Task (MedExACT) Shared Task @ ACL 2026 leveraged the MedDec corpus to prompt development of automated systems for identifying medical decisions from 9 of the 10 DICTUM categories in clinical text. The demographic makeup of the corpus and the intentional design of the evaluation metric also provides a test bed for studying generalization and robustness of extraction approaches across different demographic groupings. This manuscript describes the development of an automated approach to extracting mentions of medical decisions from clinical text using an ensemble of different classifiers to capture variance in the data and balance token-level and span-level performance generally while not penalizing specific demographic groups.

2 Methods

2.1 Data Preparation

Manual review of MedDec corpus annotations and documents resulted in the implementation of several data preprocessing steps. Analysis of the demographic groupings in the provided training/validation split of the data revealed a single Hispanic discharge summary in the validation set. Restratification of the data resulted in a more balanced distribution of demographic groups between the training and validation sets. Review of discharge summaries revealed two systematic formatting idiosyncrasies. First, 401 instances of presumably non-ASCII characters across 49 documents manifesting as a string of question marks ("?????") were discovered. These question mark strings were deduced to mainly be missing punctuation characters, and were fixed in a semi-automated way, while insertion of space characters maintained character offsets in the documents so as not to disturb the evaluation. Second, as it has been de-identified, the MIMIC-III corpus contains many instances of placeholders of previously identifiable information (e.g., Dr. [**Last Name (STitle) **]). Early experiments suggested that replacing the placeholders with synthetic data, e.g., "Dr. Smith", resulted in improved classification performance of medical decisions. These data preparation steps were used consistently for all model inputs.

2.2 Base Models

Preliminary experiments with several different models (ELECTRA-base (Clark et al., 2020), BioELECTRA (Kanakarajan et al., 2021), BiomedBERT (Gu et al., 2021), BioClinicalBERT (Alsentzer et al., 2019), and BioClinical-ModernBERT (Sounack et al., 2025)) suggested BiomedBERT was more amenable to the medical decision extraction task than other models, somewhat surprisingly even over models trained on the MIMIC-III corpus. The preprocessing steps to alter MIMIC-III artifacts likely impacted this result. Using BiomedBERT predominantly as a base, we developed an ensemble system comprised of six different architectural variants in an attempt to provide diverse responses to account for corpus variance (Table 1).

2.2.1 Variant 1: BiomedBERT Plain

This variant is designed as a straightforward implementation of BiomedBERT that also serves as a

base model for other variants. It uses token-level BIO tagging (Ramshaw and Marcus, 1995) to encode the DICTUM annotations. Overlapping spans are excluded from the training and validation data inherently by the BIO representation. For the base variant, a 10-fold weighting of Hispanic documents via over-sampling was used due to the single Hispanic document in the original validation set.

2.2.2 Variant 2: BiomedBERT + R-Drop

Variant 2 combines BiomedBERT with regularized dropout (Wu et al., 2021) (R-Drop; $\alpha=3.0$) intended to reduce overfitting. It also incorporates section-to-DICTUM-category priors to leverage the fact that some categories are more likely to appear in certain note sections. The priors are applied to the B-tag logits at inference time to decrease the probability of rare category-section pairs, and increase the probability of frequent category-section pairs.

2.2.3 Variant 3: BiomedBERT + DAPT

Domain-adaptive pre-training (DAPT) (Gururangan et al., 2020) was employed with BiomedBERT to construct Variant 3. BiomedBERT was further pre-trained using masked language modeling for 50k steps on the entire MIMIC-III corpus of discharge summaries (~59k notes). This pre-training step was entirely unsupervised and did not use any MedDec annotations or task-specific labels. As the MedDec dataset is derived from MIMIC-III, the pre-training corpus necessarily includes documents later used for downstream evaluation. This setup is consistent with common DAPT practice, where large unlabeled corpora are reused across pre-training and evaluation. Excluding test documents from DAPT, however, would provide a stricter separation and represents a potential refinement for future work. The resulting model checkpoint was also used as the base model for Variants 4 and 5.

2.2.4 Variant 4: BiomedBERT + DAPT + Global Pointer

Variant 4 takes a span-based approach that is fundamentally different from the BIO representation used by the other five variants. Instead of labeling each token with B/I/O, and subsequently determining spans, the Global Pointer approach (Su et al., 2022) directly scores every possible (start, end) token pair for each DICTUM category. The primary motivation for including Variant 4 is to increase architectural diversity.

2.2.5 Variant 5: DAPT + R-Drop + Section Embeddings

Variant 5 builds on previous variants by combining DAPT with R-Drop and section priors, and adds section embeddings that are trained jointly with the model. The section embeddings encode which note sections (see Section 2.3) each token belongs to and are added to the token embeddings before the transformer layers. The intuition is that tokens in one section (e.g. "Discharge Medications") should be treated differently than tokens in a different section (e.g., "Assessment and Plan").

2.2.6 Variant 6: BioClinical-ModernBERT + R-Drop

This is the only variant not based on BiomedBERT. The BioClinical-ModernBERT architecture uses a different attention mechanism and is trained on a broader spectrum of biomedical and clinical data than BiomedBERT. The motivation is again to add diversity in responses as this model should make different errors than the other five variants. R-Drop is incorporated to minimize overfitting and section priors are also used.

2.3 Input Representation

Due to the limited context window for BERT-based models (512 tokens), discharge summaries must be split into chunks for processing. Document sections were iteratively derived from manual analysis of the training data, with canonical section names and their variations encoded as case-insensitive regular expressions, matched only on lines shorter than 80 characters to reduce false positives from narrative text. In total, 23 canonical document section types (Appendix Table 4) were defined and used to segment notes, enabling consistent alignment of semantically equivalent sections (e.g., "Discharge Medications", "Medications on Discharge") across documents. Text appearing prior to the first matched section header was labeled as "Preamble", and unmatched headers were treated as likely subsections and assigned to the most recent preceding canonical section.

Analysis of the training data showed that 99.7% of medical decision annotations fall within a single section, motivating a section-aligned chunking scheme in which input window boundaries respect section boundaries. Sections longer than 512 tokens are split into consecutive, non-overlapping sub-chunks. Short adjacent sections that fit entirely

in the 512 token window are merged into a single chunk.

2.4 Bias mitigation

Demographic subgroup bias mitigation focused on race/ethnicity subgroups based on early experiments. Three different demographic weighting schemes were used. Initial experiments, influenced by the single Hispanic document in the original validation set, applied 10-fold document-level oversampling of Hispanic documents (A in Table 1). One variant makes use of uniform weighting across all documents (B in Table 1), due to an unintended configuration override. Two variants weight the Hispanic, African American, and Other groups by 15x, 8x, and 6x, respectively (C in Table 1), designed to approximately equalize sampling rates across underrepresented groups. Sex and language proficiency were not explicitly modeled, however, the evaluation metric's explicit inclusion of worst-group performance guided model selection and threshold tuning across all experiments.

2.5 Ensemble composition

The full ensemble comprises 30 models: 5 random seeds for each of the 6 architectural variants. Each model processes every input chunk to classify mentions of DICTUM categories in the text. Candidate annotations are normalized using the `refine_span` function from the official evaluator code. Three different matching strategies are employed (Table 2). Run A employed a per-category voting scheme which accounts for some categories being easier to detect than others. It combines weighting for each variant with separate thresholds for each category, as well as pruning underperforming individual models. In the end, 26 of the 30 models are used, and the six variants are weighted 1/1/1/3/3/3 for a total of 56 votes. Category thresholds range from 8/56 for the rare Deferment category to 18/56 for the more frequent Drug category. Run B used a confidence-weighted voting scheme using the probability of the B-tag at the span's first token to scale each model's vote. Run B optimized for worst group score in an attempt to further mitigate demographic bias. Run C used a single global threshold with equal weights; at least 8 of 30 models must predict a span with the same category and normalized text. All voting thresholds were tuned on the re-stratified validation set.

#	Variant	Head	DAPT	R-Drop	Sec. Emb.	Sec. Prior	Wt. Grp.
1	BiomedBERT	BIO					A
2	+ R-Drop	BIO		✓		✓	B
3	+ DAPT	BIO	✓				A
4	+ DAPT (Glob.Ptr)	GP	✓				A
5	+ DAPT+RDrop+Sec	BIO	✓	✓	✓	✓	C
6	ModernBERT+RDrop	BIO		✓		✓	C

Table 1: Ensemble component architectures (5 seeds each, 30 models total). Head: BIO = token-level BIO tagging, GP = Global Pointer span extraction. DAPT = domain-adaptive pre-training on MIMIC-III. Sec. Emb. = learned section embedding. Sec. Prior = section-conditional logit adjustment at inference. Demographic oversampling weight groups: A = Hispanic 10 \times ; B = uniform; C = Hispanic 15 \times , African American 8 \times , Other 6 \times . ModernBERT = BioClinical-ModernBERT. All variants use section-aligned chunking with max length 512.

3 Results

Three ensemble systems comprising 5 randomly seeded models from each of the 6 architectural variants were constructed. Preliminary experiments showed that an ensemble of 5 seeds outperformed the base BiomedBERT model. This pattern was repeated for each of the other variants and amplified by combining the seeds from all six variants. The diversity of responses and errors made by the different seeds of the different variants contributed to more consistent performance across demographic groups. Table 2 summarizes the performance of the three systems against both the re-stratified validation set and the held-out test set. While aggregate metrics provide an overall view of performance, per-category results (Appendix Table 6) reveal substantial variation across DICTUM classes, with more frequent classes generally achieving higher performance than rarer classes. Table 3 details the per-demographic performance for the three systems. Of note, the simplest voting strategy (Run C) generalized best to the test set with a negligible difference in Final Score between the re-stratified validation and test sets, while the more sophisticated voting schemes of Runs A and B overfit to the validation set demographics.

Table 5 (Appendix) presents a quasi-ablation of the system components on the validation set. Among individual components, DAPT provided the largest improvement over the baseline (+0.024), followed by R-Drop (+0.014). Multi-seed ensembling of a single variant yielded modest gains relative to individual models (+0.027-0.029). The largest improvements came from combining architecturally diverse variants: the 15-model ensemble of Variants 1-3 reached 0.570, and each subsequent architecture increased overall performance incrementally, culminating in the 30-model ensemble at

0.597. Notably, the gains from architectural diversity (+0.045 for 15 models over the single-model baseline) substantially exceeded those from seed diversity alone (+0.027 for 5 seeds of one variant), suggesting that the ensemble benefits more from combining architectures that make different errors than from simply averaging multiple runs of the same model.

4 Discussion and Conclusion

Our ensemble approach to handling diversity in corpus annotation as well as in demographic variability proved to be effective at generalization in that it achieved the highest score on the test set. In practice, the simpler voting scheme of Run C was likely the difference maker as it did not experience noticeable overfitting to the validation data. Although Run B explicitly optimized for worst-group performance (primarily across race/ethnicity), this strategy did not generalize as effectively to the test set, suggesting that fairness-aware optimization may be sensitive to sampling variability in small datasets. Future work should explore more robust fairness-aware learning strategies, including explicit modeling of additional demographic attributes such as sex and language proficiency, subgroup calibration, and worst-group-aware training objectives, to better balance overall and subgroup performance. Ensemble diversity itself appears to improve worst-group robustness, even without extensive explicit subgroup modeling.

This work presents a robust baseline system that leverages conventional transformer-based approaches to identify mentions of medical decisions in text. Future work could involve several refinements and extensions to further improve robustness and generalizability. First, the heuristic voting strategy could be replaced with a learned span-

		Span F1	Token F1	WG	Final
<i>Restratified Validation (n=53)</i>					
Run A	26-model, per-cat thresh	.540	.683	.600	.606
Run B	30-model, conf-weighted	.546	.670	.601	.605
Run C	30-model, equal 8/30	.534	.677	.589	.597
<i>Test (n=48)</i>					
Run A	26-model, per-cat thresh	.547	.653	.547	.574
Run B	30-model, conf-weighted	.553	.654	.555	.579
Run C	30-model, equal 8/30	.542	.667	.589	.597

Table 2: Summary metrics for the three submitted runs. WG = Worst Group base score. Final Score = $\frac{1}{2}$ Base + $\frac{1}{2}$ [WG], where Base = $\frac{1}{2}$ [Span F1] + $\frac{1}{2}$ [Token F1]. Bold indicates the highest value in each column.

Subgroup	Run A		Run B		Run C	
	Val	Test	Val	Test	Val	Test
Female	.600	.614	.601	.618	.595	.628
Male	.619	.592	.612	.595	.612	.589
White	.610	.603	.605	.603	.606	.599
African Am.	.605	.547	.601	.555	.595	.590
Hispanic	.629	.648	.622	.635	.626	.660
Asian	.634	.607	.638	.600	.623	.589
Other	.602	.598	.603	.625	.589	.636
English	.614	.615	.609	.617	.612	.610
Non-English	.609	.577	.608	.582	.596	.595

Table 3: Base score ($\frac{1}{2}$ [Span F1] + $\frac{1}{2}$ [Token F1]) by demographic subgroup. **Bold** indicates the worst-performing subgroup for each run and split.

level fusion model to aggregate predictions across diverse ensemble members in a more principled manner, as explored in prior work on span-based system combination for named entity recognition (Fu et al., 2021). Second, fairness-aware learning could be strengthened through the strategies discussed above. Third, although domain-adaptive pre-training followed common practice, excluding evaluation documents from the pre-training corpus would provide a stricter separation and merits further investigation. Fourth, extending the extraction framework to natively support overlapping MedDec annotations may better capture the multi-label nature of clinical decision statements. Finally, deeper integration of document structure, such as through section-conditioned representations or decoding, may further exploit the strong concentration of specific medical decision categories within particular document sections observed in the data.

Although not discussed above, the exact span requirement in the evaluation metric proved challenging, especially considering some of the annotation inconsistencies (e.g., some annotations in a list include the bullet or number, and some do not) observed in the corpus. Significant effort went into analyzing and attempting to ameliorate such in-

consistencies, though none of the approaches were ultimately used. Restratifying the provided training and validation sets proved to be critical for developing models that generalize. More broadly, our results suggest that architectural diversity within an ensemble can compensate for limited and variable training data, and that restraint in threshold tuning is essential when validation sets are small.

Acknowledgments

We thank the MedExACT shared task organizers and the MedDec corpus architects for providing the data and evaluation infrastructure, as well as the reviewers for their insightful comments. The MIMIC-III database was made available by PhysioNet. (Goldberger et al., 2000) AI-assisted coding tools (Anthropic Claude) were used during system development.

Ethical Considerations

For any classification task, it is critical that the approach is designed to generalize over different population subgroups, especially those groups traditionally under-represented in medical studies. The MedExACT evaluation metric explicitly penalizes systems that perform poorly on specific

demographic subgroups. Our system design addressed demographic subgroups primarily through weighted sampling of documents from low frequency demographic subgroups.

MIMIC-III data was stored and processed exclusively on a HIPAA-compliant server managed by University of Colorado Anschutz Medical Campus Office of Information Technology. System development using AI coding assistance was conducted on a local machine without access to the clinical data; code was transferred to the server via a private repository.

Limitations

Due to time constraints of the shared task, system development was driven largely by iterative experimentation rather than systematic ablation and tuning of different components. Once the ensemble approach proved effective for a single model (i.e., 5 random seeds improved over the single initial seed), development focused on expanding architectural diversity to handle annotation variance in the MedDec corpus, rather than optimizing individual model configurations. In the submitted runs, some leave-one-out experiments were performed to determine if removal of specific model seeds would improve overall performance (Run A), but this run overfit to the validation demographics. The best performing run (Run C) resulted from monotonic additions to the system that did not degrade performance on the "Worst-Group Base" score when evaluated on the restratified validation set. A prime avenue for future work would involve systematically ablating the contribution of each architectural component and optimizing the composition of the ensemble itself.

Despite the range of demographic categories captured by the MedDec corpus, the comparatively few examples of many of the categories, specifically the race/ethnicity categories, limit the ability to train models that will truly generalize well over a larger sample. This may have played a role in why the worst-performing demographic subgroup differed between validation and test for two of our three submissions (Table 3).

Finally, our use of 30 models in an attempt to capture annotation variance in the corpus, though it did achieve the highest score of all submitted results, is likely not a practical solution for production deployment. Processing each note 30 different times across an entire health data warehouse is

likely too inefficient without further optimization, such as pruning of redundant ensemble members.

References

- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024a. **MedDec: A dataset for extracting medical decisions from discharge summaries**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16442–16455, Bangkok, Thailand. Association for Computational Linguistics.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024b. **MedDec: Medical Decisions for Discharge Summaries in the MIMIC-III Database**. *PhysioNet*. Version 1.0.0.
- Jinlan Fu, Xuan-Jing Huang, and Pengfei Liu. 2021. Spanner: Named entity re-/recognition as span prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7183–7195.
- Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8342–8360.
- Alistair Johnson, Tom Pollard, and Roger Mark. 2016a. **MIMIC-III Clinical Database**. *PhysioNet*. Version 1.4.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016b. MIMIC-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Kamal Raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. Bioelectra: pre-trained biomedical text encoder using discriminators. In *Proceedings of the 20th workshop on biomedical language processing*, pages 143–154.

Eirik H Ofstad, Jan C Frich, Edvin Schei, Richard M Frankel, and Pål Gulbrandsen. 2016. What is a medical decision? a taxonomy based on physician statements in hospital encounters: a qualitative study. *BMJ open*, 6(2):e010098.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third workshop on very large corpora*.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J Pollard, Eric Lehman, Alistair EW Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp. *arXiv preprint arXiv:2506.10896*.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition. *arXiv preprint arXiv:2208.03054*.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, and 1 others. 2021. R-drop: Regularized dropout for neural networks. *Advances in neural information processing systems*, 34:10890–10905.

Appendix

#	Section Type
1	Allergies
2	Assessment and Plan
3	Attending Note
4	Brief Hospital Course
5	Chief Complaint
6	Discharge Condition
7	Discharge Diagnosis
8	Discharge Disposition
9	Discharge Instructions
10	Discharge Medications
11	Family History
12	Follow-up
13	History of Present Illness
14	Imaging / Radiology
15	Major Surgical / Invasive Proc.
16	Medications on Admission
17	Past Medical History
18	Past Surgical History
19	Pertinent Results
20	Physical Exam
21	Procedures
22	Review of Systems
23	Social History
24	Other / Unknown

Table 4: The 23 canonical section types identified in MIMIC-III discharge summaries, used for section-aligned chunking and section-conditioned inference.

Configuration	n	Final	Δ
<i>Single model (best seed)</i>			
BiomedBERT (V1)	1	.525	—
+ R-Drop (V2)	1	.539	+0.014
+ DAPT (V3)	1	.549	+0.024
<i>5-seed ensemble</i>			
V1 \times 5 seeds	5	.552	+0.027
V3 \times 5 seeds	5	.554	+0.029
<i>Multi-architecture ensemble</i>			
V1–3	15	.570	+0.045
+ V4 (Global Ptr)	20	.585	+0.060
+ V5 (Sec-aware)	25	.592	+0.067
+ V6 (ModernBERT)	30	.597	+0.072

Table 5: Quasi-ablation on the validation set showing the contribution of individual components (top), multi-seed ensembling (middle), and progressive addition of architectural variants (bottom). n = number of models. Final = Final Score. Δ is relative to the single BiomedBERT baseline.

#	Category	Gold	Run A			Run B			Run C		
			P	R	F1	P	R	F1	P	R	F1
1	Contact	361	.311	.543	.396	.344	.529	.417	.388	.438	.411
2	Gathering info	32	.222	.062	.098	.273	.094	.140	.273	.094	.140
3	Defining problem	2304	.534	.618	.573	.533	.630	.577	.512	.614	.559
4	Treatment goal	20	.529	.450	.486	.533	.400	.457	.421	.400	.410
5	Drug	1536	.584	.677	.627	.575	.689	.627	.556	.691	.616
6	Therapeutic procedure	700	.409	.490	.446	.486	.457	.471	.393	.484	.434
7	Evaluating test result	867	.420	.512	.462	.402	.502	.447	.405	.491	.444
8	Deferment	7	.000	.000	.000	.091	.143	.111	.000	.000	.000
9	Advice/precaution	274	.441	.730	.549	.462	.719	.563	.544	.653	.594
	All (micro)	6101	.491	.600	.540	.500	.601	.546	.488	.588	.534

Table 6: Per-category Span precision, recall, and F1 on the validation set for each submitted run. Gold shows the number of gold-standard spans. Note the class imbalance as some medical decision categories are sparsely represented.

LAMAR at MedExACT 2026: Agreement-Driven Large Language Model Ensembles for Clinical Decision Extraction from Discharge Summaries

Monrada Chiewhawan^{*1,2}, Keetawan Limaroon^{*3}, Titipat Achakulvisut¹,

¹Department of Biomedical Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom, Thailand,

²Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand,

³Department of Computer Engineering, Faculty of Engineering,

King Mongkut's University of Technology Thonburi, Bangkok, Thailand,

Correspondence: titipat.ach@mahidol.ac.th

Abstract

Clinical decision extraction from discharge summaries detects contiguous text spans expressing medical decisions and assigns each to predefined categories. In this paper, we propose an ensemble approach using large language models for clinical decision extraction from discharge summaries in the MedDec dataset with XML-like inline tag annotations. The ensemble consists of Qwen3.5-4B models trained under three different settings: (1) Dynamic Fine-tuning (DFT) with LoRA on the original training set, (2) DFT with LoRA then GRPO reinforcement on the original training set, and (3) DFT with LoRA on the original training set augmented with pseudo-labels. We aggregated predictions for each document by category using weights derived from inter-model agreement. Agreement-driven ensembles further enhanced performance across all metrics, yielding a 8.31% gain in Overall F_1 over the baseline and securing second place on the test leaderboard. Subgroup analysis further confirms that performance remains consistent across demographic groups, with no disproportionate degradation on underrepresented populations. We release our code at <https://github.com/biodatlab/medexact-lamar>.

1 Introduction

Clinical decision extraction focuses on retrieving relevant medical decisions from unstructured medical texts. Ofstad et al. (2016) categorized clinical decisions into ten categories, providing a standardized taxonomy for developing extraction techniques and models.

MedExACT is a shared task focused on medical decision extraction from ICU discharge summaries using the MedDec dataset (Elgaar et al., 2026, 2024). MedDec contains 451 discharge summaries from MIMIC-III (Johnson et al., 2016) annotated under the DICTUM guideline. However,

^{*}Equal contribution.

its population skews White, male, and English-speaking, making consistent performance across demographic groups difficult to achieve.

Encoder-only transformer models are used as baselines for this task. Their small context windows and token-wise classification limit performance, as they struggle to capture context across documents. Despite RoBERTa (Liu et al., 2019) being the strongest baseline with a Base Score of 0.5301 and Overall F_1 of 0.5111, the relatively low Span F_1 (0.4363) indicates that accurate decision boundary extraction remains a challenge. Moreover, zero-shot LLaMA-3-8B-Instruct (Grattafiori et al., 2024) reported in MedDec showed limited effectiveness, likely due to challenges in handling long contexts and generating structured outputs. These results suggest considerable room for improvement and highlight the difficulty of the task.

We propose an ensemble approach for clinical decision extraction based on inline XML-like tagging. The ensemble consists of models trained under three complementary settings: (1) DFT with LoRA on the original training set, (2) DFT with LoRA then GRPO on the original training set, and (3) DFT with LoRA on the original training set augmented with pseudo-labels. Our system achieved a Span F_1 of 0.5257, a Token F_1 of 0.6750, and a Worst Group F_1 of 0.5881, yielding an Overall F_1 score of 0.5942 on the leaderboard.

2 Related Work

2.1 From Sequence Labeling to Generative Extraction

Encoder models have long been the backbone of named-entity recognition, with BERT (Devlin et al., 2019) and its variants such as RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), and ALBERT (Lan et al., 2020), setting strong baselines through masked language modeling. However, these models share a fundamental limitation. For example,

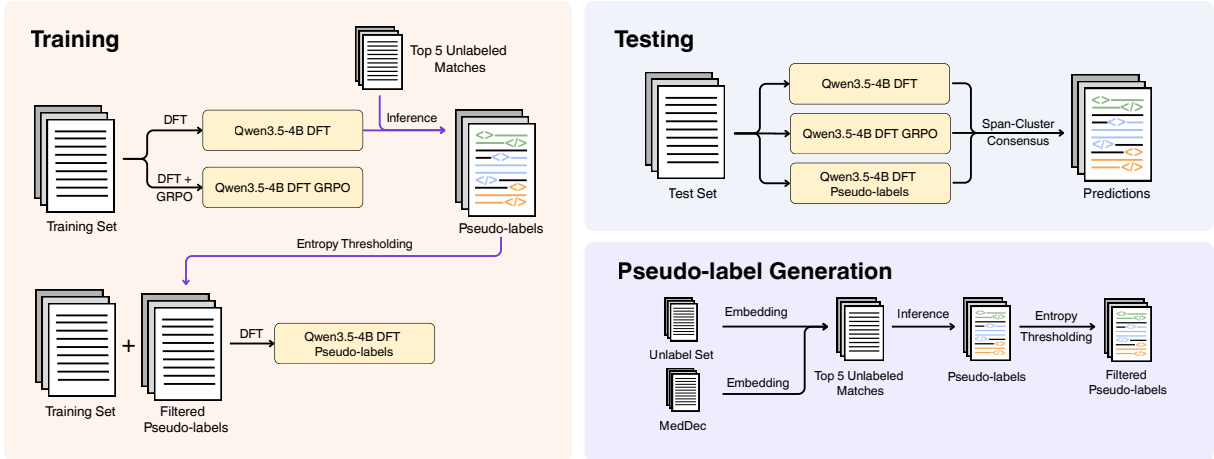


Figure 1: An overview of our system in training, testing, and pseudo-label generation. (Left) Training combines direct fine-tuning (DFT) and Group Relative Policy Optimization (GRPO) on Qwen3.5-4B, augmented with entropy-filtered pseudo-labels from unlabeled data. (Right) At test time, predictions from all three model variants are aggregated via Span-Cluster Consensus.

BIO tagging schemes assign exactly one label per token, making it impossible to represent overlapping spans regardless of model size or architecture. This is a critical shortcoming for clinical NER, given that 4.2% of labeled tokens in MedDec (Elgaar et al., 2024) belong to more than one span simultaneously.

Large language models offer an alternative approach. NER can be reframed as a generation task where the model rewrites the input with entity markers inserted inline, placing no structural restriction on overlapping spans. Wang et al. (2025) demonstrated this paradigm with GPT-NER using special token markup, achieving performance comparable to supervised baselines. Hu et al. (2024) showed that wrapping entities in HTML `` tags maps cleanly to standard evaluation schemes. However, with long, domain-specific, and complex documents like discharge summaries, we hypothesize that fine-tuning these models can serve as an effective approach for clinical decision extraction.

2.2 Supervised Fine-Tuning, Parameter-Efficient Adaptation, and Generalization

Supervised Fine-Tuning (SFT) trains LLMs on labeled examples to produce desired outputs. Since full fine-tuning is computationally expensive, Low-Rank Adaptation (LoRA) (Hu et al., 2022) and Weight-Decomposed Low-Rank Adaptation (DoRA) (Liu et al., 2024) offer a practical alternative by inserting small trainable matrices that capture task-specific changes with fewer param-

eters. However, SFT tends to memorize rather than generalize, where uncertain tokens receive disproportionately large gradients, pushing the model to overfit. Dynamic Fine-Tuning (DFT) (Wu et al., 2026) addresses this by scaling each token’s loss according to model confidence, producing balanced updates that generalize substantially better across challenging tasks.

2.3 Reinforcement Learning with Verifiable Rewards

Reinforcement Learning with Verifiable Rewards (RLVR) replaces the learned reward model in classical RLHF (Ouyang et al., 2022) with verifiable reward functions. Group Relative Policy Optimization (GRPO) (Shao et al., 2024), improving on PPO (Schulman et al., 2017), derives training signals by comparing outputs within a sampled group using rule-based rewards, making it well-suited for structured output tasks with strict metrics. However, broader application of GRPO can lead to entropy collapse, training instability, or diluted signal under multiple rewards. DAPO (Yu et al., 2025) addresses instability by maintaining output diversity, while GDPO (Liu et al., 2026) handles multi-reward settings by standardizing each reward independently before combination. We adopt GRPO and integrate DAPO and GDPO to mitigate these respective challenges.

2.4 Span Aggregation via Model Agreement

Individual models often disagree on span boundaries, so we aggregate predictions across models

to improve annotation reliability. Minimum Bayes Risk (MBR) decoding selects the output that maximizes agreement across a set of candidates, following a hypothesis that the output should be consistent with the others (Bickel and Doksum, 2015). It has shown to improve generation quality across NLP tasks. Heineman et al. (2024) adapted this intuition to multi-prompt decoding of a single model, leveraging varied inputs to encourage diverse prediction distributions. We adapt MBR decoding for clinical decision extraction, but extend prior work by aggregating outputs from multiple differently trained models and by performing refinement at the span-cluster level instead of choosing one complete prediction (Section 4.4).

3 Dataset

We use MedDec (Elgaar et al., 2024) as our dataset. For model development, we follow the original split of 350 training, 53 validation, and 48 test samples. The dataset consists of discharge summaries annotated with decision labels and span boundary offsets. We convert the original dataset into inline XML-like tags that support overlapping spans (Figure 2). This is necessary because 4.2% of tokens overlap. Therefore, this task encourages LLMs to reason through the input and apply inline-tag annotations where relevant.

Our work covers the 9 DICTUM classes defined in (Elgaar et al., 2026). The label distribution is highly imbalanced, with *Defining problem* accounting for 39% of training and validation spans, while *Deferment* accounts for under 0.2%. The dataset also has a demographic imbalance. For example, the Hispanic group contains just 1 sample each in validation and test, and Asian patients are entirely unrepresented in validation, with only 2 training samples (Appendix A). This makes the task challenging as it can strongly skew Worst Group F_1 scores and, in turn, affect Overall F_1 . (Section 5.1).

During validation, a test run suggested a possible category misassignment in the ground-truth annotations. We locally reassigned the affected section to better reflect our model’s true performance (Appendix B). Additionally, we used MIMIC-III discharge summaries outside the MedDec set for pseudo-label generation (Section 4.3).

MedDec Annotations

```
{ "decision": "Right ventricular chamber size is normal with mild global free wall hypokinesis", "category": "Category 7: Evaluating test result", "start_offset": 6224, "end_offset": 6303 }, { "decision": "The aortic valve leaflets (3) are mildly thickened but aortic stenosis is not present", "category": "Category 7: Evaluating test result", "start_offset": 6305, "end_offset": 6390 }, { "decision": "but aortic stenosis is not present. There is no aortic valve stenosis. No aortic regurgitation is seen", "category": "Category 3: Defining problem", "start_offset": 6356, "end_offset": 6458 }
```

Inline Tags (Ours)

```
[...] <evaluate_result>Right ventricular chamber size is normal with mild global free wall hypokinesis</evaluate_result>. <evaluate_result>The aortic valve leaflets (3) are mildly thickened</define_problem>but aortic stenosis is not present</evaluate_result>. There is no aortic valve stenosis. No aortic regurgitation is seen</define_problem>. [...]
```

Figure 2: Converting original offsets (left) into inline tags that support containment and partial overlap (right).

4 Methodology

4.1 Dynamic Fine-tuning (DFT)

We train our first model variant (Model 1) using Dynamic Fine-Tuning (DFT) (Wu et al., 2026), which improves generalization by rescaling token gradients based on model confidence. Standard Supervised Fine-Tuning (SFT) applies a uniform cross-entropy loss, which often assigns disproportionately large gradients to tokens the model is uncertain about, leading to unstable optimization and overfitting. DFT addresses this by weighting each token’s loss by its current generation probability. Formally, given an input sequence x and a target reference sequence $y^* = (y_1^*, \dots, y_T^*)$ of length T , the DFT objective is defined as:

$$\mathcal{L}_{\text{DFT}}(\theta) = -\frac{1}{T} \sum_{t=1}^T \text{sg}(\pi_{\theta}(y_t^* | y_{<t}^*, x)) \cdot \log \pi_{\theta}(y_t^* | y_{<t}^*, x), \quad (1)$$

where t is the current token index, π_{θ} is the model’s policy parameterized by θ , $y_{<t}^*$ denotes the preceding target tokens $(y_1^*, \dots, y_{t-1}^*)$, and $\text{sg}(\cdot)$ is the stop-gradient operator. This promotes balanced updates and encourages the model to learn robust extraction patterns rather than memorizing surface-level data.

4.2 Reinforcement Learning with Verifiable Rewards

4.2.1 Training Objective

We train our second model variant (Model 2) by initializing the policy with Model 1 (DFT), providing a robust starting point that allows the policy

to focus directly on optimizing extraction performance. We use a significantly reduced LoRA rank, exploiting the fact that policy gradient updates derived from sparse rewards inherently occupy a low-rank subspace. We optimize the policy using the Group Relative Policy Optimization (GRPO) (Shao et al., 2024) framework, which estimates advantages across sampled groups without a separate value model. We implement two targeted enhancements: First, we adopt the DAPO (Yu et al., 2025) approach at the token level to mitigate entropy collapse. By applying an asymmetric clipping range ($\epsilon_{low} = 0.20, \epsilon_{high} = 0.28$), DAPO safely broadens the trust region for stable exploration. Second, we integrate GDPO (Liu et al., 2026) to handle multi-objective rewards, standardizing each reward component independently across the group before summation. This preserves fine-grained distinctions between candidate generations and ensures each objective contributes equitably to the final gradient update.

4.2.2 Reward Functions

We translate our evaluation metrics into a composite reward system to optimize extraction performance. Relying on a single metric risks producing either sparse training signals or reward exploitation. To address this, we design a multi-objective reward that balances strict boundary evaluation, flexible word-level credit, and a hard fidelity constraint. Each component is bounded within $[0.0, 1.0]$, and the overall reward for each generation is defined as their combination:

1. **Fidelity Reward:** We design this component to strictly prevent hallucination. We strip all generated tags and compare the remaining text to the original input. Using a character-level similarity ratio, we assign a full reward of 1.0 for a near-perfect match where the ratio ≥ 0.99 , and a reward of 0.0 for a ratio below 0.90, with linear scaling in between. This safeguards that the model behaves purely as a sequence tagger.
2. **Token F_1 Reward:** We use this component to provide flexible, word-level credit. We extract the predicted text spans, split them into individual words, and measure their overlap with the ground-truth words for each category. This allows the model to receive positive feedback for identifying relevant medical concepts
3. **Span F_1 Reward:** We treat this component as the primary objective and strictest metric. We evaluate whether the model precisely identifies the entire text span of a clinical decision, granting a reward only when the extracted string perfectly matches the ground truth within its category. By combining this strict target with the more forgiving Token F_1 reward, we guide the model toward predicting exact span boundaries.

even when exact span boundaries are slightly misaligned.

4.3 Pseudo-label Generation

We train our third model (Model 3) by augmenting the training set with pseudo-labels generated from unlabeled clinical text. We identify semantically relevant unlabeled instances from 59,201 MIMIC-III discharge summaries outside of MedDec. We embed both the unlabeled pool and the complete MedDec set using Qwen3-Embedding-4B (Zhang et al., 2025). Then, we compute cosine similarity between each of the 451 MedDec documents and the unlabeled pool to retrieve the top 5 most similar candidates. We assign each unlabeled document only to its highest-scoring match, yielding a refined pool of 2,255 candidate samples. We then run inference on this pool using Qwen3.5-4B DFT (Model 1), using entropy as a proxy for prediction confidence based on the observed Pearson correlation of -0.4816 with Base Score on the validation set. A higher entropy threshold includes more examples but noisier labels. We experiment with two cut-offs at the 10th percentile (P_{10}) and 15th percentile (P_{15}), adding 221 and 331 pseudo-labeled samples, respectively. We combine each set with the original training set and fine-tune Model 3 using DFT.

4.4 Span-Cluster Consensus

We aggregate predictions by category and document in four stages. First, we greedily match spans above a minimum pairwise IoU threshold and compute pairwise soft Span F_1 agreement between models, then normalize these scores so that model weights sum to one. Second, we cluster all spans using a minimum cluster IoU threshold and score each cluster by the summed weights of its contributing models, removing clusters below the minimum cluster support. Third, we select final boundaries via a weighted vote over start–end offset pairs. Fourth, we merge same-category dupli-

cates above a minimum duplication IoU threshold, keeping the longer span. Figure 3 illustrates each stage.

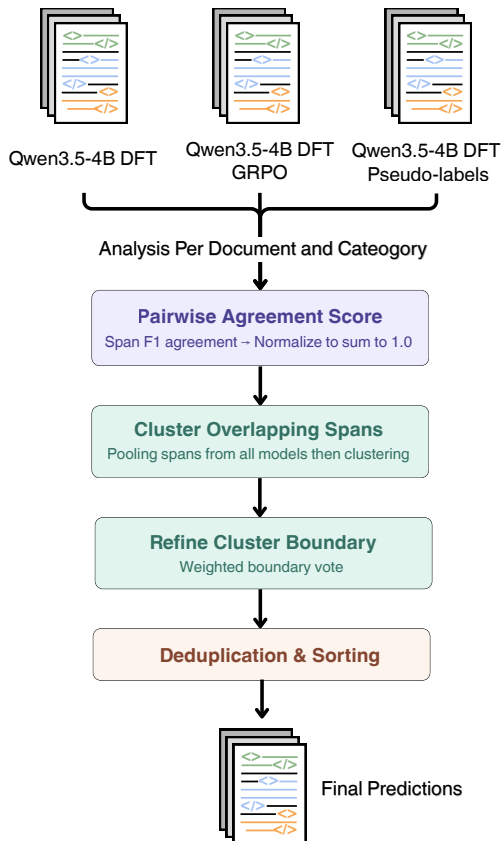


Figure 3: Overview of the Span-Cluster Consensus, comprising pairwise model agreement, cluster weighted boundary refinement, and span deduplication.

To find the best ensemble configuration, we performed a systematic search on the validation set over four fusion hyperparameters: minimum pairwise IoU, minimum cluster IoU, minimum cluster support, and minimum duplication IoU, each within a continuous range of [0.0, 1.0]. We used the Tree-structured Parzen Estimator (TPE) algorithm (Watanabe, 2023) with a warm-start trial seeded from our heuristic baseline to accelerate convergence. The search directly maximized the Overall F_1 score computed by the official evaluation script, ensuring that the selected configuration optimizes for both accuracy and demographic robustness jointly.

5 Experimental Setup

5.1 Evaluation Metrics

We evaluate with the official script and report Token F_1 , Span F_1 , Base Score, Worst Group F_1 , and Overall F_1 . Token F_1 measures word-level

overlap on non-overlapping tokens, while Span F_1 requires exact span matches after word-boundary normalization. Both metrics are macro-averaged across the nine DICTUM labels per document and then across our internal validation or test set. Base Score is the mean of Token F_1 and Span F_1 :

$$\text{Base Score} = \frac{\text{Span } F_1 + \text{Token } F_1}{2} \quad (2)$$

Worst Group F_1 is the lowest Base Score across sex, ethnicity, and language groups:

$$\text{Worst Group } F_1 = \min_{i \in \{1, \dots, 8\}} (\text{Base Score}_i) \quad (3)$$

and Overall F_1 is the average between Base Score and Worst Group F_1 :

$$\text{Overall } F_1 = \frac{(\text{Base Score} + \text{Worst Group } F_1)}{2} \quad (4)$$

5.2 Baselines and Proposed Models

5.2.1 Encoder Baselines with BIO Tagging

As baselines, we evaluate a range of encoder-based models trained with BIO tagging for span detection, including models with stronger biomedical vocabulary coverage such as BioMedBERT (Gu et al., 2020), and models designed for longer clinical documents such as Clinical Longformer (Li et al., 2022). We also include ModernBERT (Warner et al., 2025), which incorporates architectural improvements such as RoPE positional encoding and an extended context window of 8,192 tokens. We additionally retrain RoBERTa and ELECTRA, the best-performing models from Elgaar et al. (2024), under our experimental configuration.

5.2.2 Generative Extraction Baseline

We evaluate two generative baselines. First, we fine-tune Qwen3.5-4B (Qwen Team, 2026) using standard SFT with autoregressive cross-entropy loss on the same training set and prompt as our proposed models (Appendix C). Comparing this baseline against Model 1 directly isolates the benefit of DFT over standard SFT. Second, we prompt GPT-5 (Singh et al., 2025) in a zero-shot setting using the same prompt, establishing an upper reference for generative extraction without task-specific training.

5.2.3 Proposed Models

We use Qwen3.5-4B (Qwen Team, 2026) as our base model for its reasoning capability and extended context window, which processes full discharge summaries without truncation. We train three variants based on strategies in Section 4:

1. **Model 1 (DFT Model):** We fine-tuned Qwen3.5-4B with the DFT objective, establishing it as a primary extraction model.
2. **Model 2 (DFT + RLVR Model):** We initialized from Model 1 and applied RLVR with DAPO and GDPO to further improve performance. Three reward combinations were tested: Span and Token F_1 (Model 2), Span F_1 only, and Token F_1 only. The first combination was also evaluated with metrics other than F_1 , specifically $F_{0.5}$ and precision.
3. **Model 3 (DFT with Pseudo-label Model):** We fine-tuned Qwen3.5-4B with DFT on the original training set augmented with pseudo-labels generated by Model 1. We experimented with two entropy cutoffs including P_{10} (221 samples) and P_{15} (331 samples) to evaluate the tradeoff between pseudo-label quantity and quality.

Finally, predictions from all three models are aggregated using Span-Cluster Consensus (Section 4.4) and compared against two simpler baselines: majority voting (median start/end offsets, majority-voted category) and minimum-entropy selection (the single per-document prediction with lowest generation entropy).

5.3 Implementation Details

We implemented our pipeline using the Unsloth library (Daniel Han and team, 2023) and the Transformers Reinforcement Learning (TRL) framework (von Werra et al., 2020) across all training stages. All models use Qwen3.5-4B (Qwen Team, 2026) as the base model and AdamW (Loshchilov and Hutter, 2017) as the optimizer. Due to the differing computational demands, the DFT stages were trained on a single NVIDIA A100 (80GB) GPU, whereas RLVR was trained across four NVIDIA A100 (40GB) GPUs using Distributed Data Parallel (DDP) (Li et al., 2020). All models start with DFT via LoRA, where Model 1 and 2 were trained on a tag-converted training set and Model 3 with the addition of pseudo-labels. This stage used Rank-Stabilized LoRA (rsLoRA) (Kalajdziewski, 2023)

with a rank of 256, $\alpha = 32$, 2 epochs, a learning rate of 2×10^{-4} , a batch size of 8, and a cosine learning rate scheduler. Model 2 further applies Reinforcement Learning from Verifiable Rewards (RLVR) trained for 1 epoch using the DAPO loss with standard LoRA, a lower learning rate of 5×10^{-5} , a larger batch size of 32, rollout size of 64, a constant scheduler, and a reduced LoRA rank and α of 2 (Schulman and Lab, 2025). For efficient inference, we utilized the vLLM engine (Kwon et al., 2023) with a temperature of 0. The ensemble hyperparameter search used the Optuna library with 5,000 trials using the default TPE sampler (Akiba et al., 2019).

6 Results

We evaluated various setups of model development across the Span F_1 , Token F_1 , and the composite scores. The following section reports the key results and their implications with respect to our internal validation set, unless specified otherwise. Moreover, our interpretation of the results focuses primarily on Base Score performance, since the Overall F_1 may be subject to the imbalanced split caused by the Worst Group F_1 as discussed in Section 3. Accordingly, we treat the Overall F_1 and Worst Group F_1 as supplementary context.

6.1 Encoder and Generative Baselines

All fine-tuned LLMs outperform both encoder and generative baselines (Table 1). Among encoder models, Clinical Longformer achieved the highest performance, with a Span F_1 of 0.4327, Token F_1 of 0.6297, and a Base Score of 0.5312. GPT-5, by contrast, yielded a competitive Span F_1 of 0.5179 but fell sharply on a Token F_1 of 0.2230 and a Base Score of 0.3210. Our LoRA SFT model outperforms both baselines across all three metrics, improving over Clinical Longformer by 1.96% in Token F_1 and 5.58% in Base Score, while also exceeding GPT-5 by 0.68% in Span F_1 and 26.60% in Base Score.

6.2 Fine-tuning Approaches

LoRA was the more efficient adapter, outperforming DoRA at the same rank and fine-tuning settings in both DFT (Base Score 0.5934 vs. 0.5826) and SFT (0.5870 vs. 0.5817) (Table 1). DFT also performed better than SFT in both LoRA (0.5934 vs. 0.5870) and DoRA (0.5826 vs. 0.5817). Although LoRA DFT had a lower Worst Group F_1 than LoRA SFT, the Worst Group F_1 for both methods

Approach	Model & Configuration	Span F_1	Token F_1	Base Score	Worst Group F_1	Overall F_1
Baselines	RoBERTa (BIO Tag)	0.4171	0.6233	0.5202	0.4416	0.4809
	ELECTRA (BIO Tag)	0.4061	0.5890	0.4976	0.4105	0.4540
	BioMedBERT (BIO Tag)	0.4126	0.5967	0.5047	0.4063	0.4555
	Clinical Longformer (BIO Tag)	0.4327	0.6297	0.5312	0.4609	0.4960
	ModernBERT (BIO Tag)	0.3282	0.5070	0.4176	0.3787	0.3981
Zero-shot	GPT-5-2025-08-07 (Medium Reasoning)	0.5179	0.2230	0.3210	0.3025	0.3118
Fine-tuning	Qwen3.5-4B (DoRA SFT, $r = 256$)	0.5250	0.6385	0.5817	0.4994	0.5406
	Qwen3.5-4B (DoRA DFT, $r = 256$)	0.5177	0.6475	0.5826	0.5151	0.5488
	Qwen3.5-4B (LoRA SFT, $r = 256$)	0.5247	0.6493	0.5870	<u>0.5440</u>	<u>0.5655</u>
	Model 1: Qwen3.5-4B (LoRA DFT, $r = 256$)	0.5299	0.6570	<u>0.5934</u>	0.5225	0.5580
Model Size	Qwen3.5-9B (LoRA DFT, $r = 256$)	0.5179	0.6582	0.5881	0.5044	0.5462
Rank Ablation	Qwen3.5-4B (LoRA DFT, $r = 128$)	0.5152	0.6464	0.5808	0.4716	0.5262
	Qwen3.5-4B (LoRA DFT, $r = 512$)	0.5273	0.6504	0.5889	0.4695	0.5292
RLVR Reward	Qwen3.5-4B (GRPO, Token F_1)	0.5258	<u>0.6577</u>	0.5917	0.5241	0.5579
	Qwen3.5-4B (GRPO, Span F_1)	0.5298	0.6573	0.5935	0.5261	0.5598
	Qwen3.5-4B (GRPO, Span and Token Precision)	0.5296	0.6523	0.5909	0.5249	0.5579
	Qwen3.5-4B (GRPO, Span and Token $F_{0.5}$)	0.5289	0.6560	0.5925	0.5309	0.5617
	Model 2: Qwen3.5-4B (GRPO, Span and Token F_1)	0.5273	0.6565	0.5919	0.5320	0.5619
Pseudo-labels	Qwen3.5-4B (Train set + P_{15})	0.5283	0.6460	0.5871	0.4783	0.5327
	Model 3: Qwen3.5-4B (Train set + P_{10})	0.5307	0.6481	0.5894	0.5022	0.5458
Ensemble	Majority vote	0.5298	0.6521	0.5910	0.5210	0.5560
	Minimum entropy	0.5307	0.6481	0.5894	0.5022	0.5458
	Final Model: Span-Cluster Consensus	<u>0.5300</u>	0.6611	0.5955	0.5518	0.5737

Table 1: Performance comparison of baselines, proposed configurations (**Model 1, 2, and 3**), ablations, and final ensemble on the internal validation set. **Bold** and underline indicate best and second-best results.

was the Hispanic subgroup, which contained only one sample and was therefore a less stable evaluation point. In all other subgroup analyses, LoRA DFT performed better than LoRA SFT. Therefore, we chose LoRA DFT as Model 1.

6.3 Adapter Rank and Model Size

Qwen3.5-4B was the optimal model size, outperforming Qwen3.5-9B (Base Score 0.5934 vs. 0.5881) (Table 1). At a fixed alpha of 32, an adapter rank of 256 achieved the highest Base Score (0.5934), compared with rank 128 (0.5808) and rank 512 (0.5889).

6.4 Reinforcement Learning Configuration

The Token F_1 -only reward variant achieved the highest Token F_1 score (0.6577), while the Span F_1 -only variant achieved the best Span F_1 score (0.5298) (Table 1). Notably, the Span F_1 -only variant also maintained a comparatively high Token F_1 score (0.6573), resulting in the highest Base Score among the variants at 0.5935. In addition to F_1 , we explored other evaluation metrics, including $F_{0.5}$, which weights precision more heavily than recall, and precision, which measures the fraction of predicted positive instances that are truly positive. Among these metrics, $F_{0.5}$ appears to perform best, followed by F_1 and then precision, with Base Scores of 0.5925, 0.5919, and 0.5909, respectively.

6.5 Pseudo-labels Thresholds

When defined as the entropy percentile threshold, P_{10} outperformed P_{15} across all metrics (Table 1). Model 3 (P_{10}) achieved the highest Span F_1 of 0.5307 and a Token F_1 of 0.6481, while also misclassifying fewer spans as non-decision, particularly in minority classes such as *Gathering additional information*, *Treatment goal*, and *Deferment*.

6.6 Ensemble Models

The ensemble outperformed the individual components and all other variants on all but one metric, achieving the best Token F_1 (0.6611), Base Score (0.5955), Worst Group F_1 (0.5518), and Overall F_1 (0.5737) (Table 1). Its Span F_1 of 0.5300 was the second highest, narrowly behind Model 3 (0.5307).

Rank	Team	Span F_1	Token F_1	Base	Worst F_1	Overall
1	billbaumgartner	0.5419	0.6667	0.6043	0.5886	0.5965
2	LAMAR (Ours)	0.5257	0.6750	0.6003	0.5881	0.5942
3	Otter	0.5181	0.6666	0.5924	0.5695	0.5809
4	viahes	0.5237	0.6541	0.5889	0.5723	0.5806
5	ahmed_ayman	0.4900	0.6796	0.5848	0.5601	0.5724
-	Baseline	0.4363	0.6238	0.5301	0.4922	0.5111

Table 2: Official Leaderboard of MedExACT Shared Task (Elgaar et al., 2026). Our proposed system (LAMAR) achieves highly competitive performance, securing 2nd place overall.

6.7 Test Set Performance

Our system ranked second on the test leaderboard (Table 2). Compared with the first-place system, we achieved an Overall F_1 that was 0.23% lower, but a 0.83% higher Token F_1 . We also consistently outperformed the baseline across all scores with a 7.02% improvement on Base Score. As shown in Table 3, our method performs consistently well across all demographic subgroups with the Worst Group F_1 at 0.5881 in the Non-English subgroup.

Category	Subcategory	Span F_1	Token F_1	Base Score
Gender	Male	0.5117	0.6678	0.5897
	Female	0.5495	0.6850	0.6173
Ethnicity	White	0.5216	0.6726	0.5971
	African American	0.4962	0.6926	0.5944
	Hispanic	0.5761	0.6794	0.6277
	Asian	0.5319	0.7090	0.6205
	Other	0.5538	0.6651	0.6094
Language	English	0.5303	0.6853	0.6078
	Non-English	0.5156	0.6605	0.5881

Table 3: Test set performance of our proposed system across demographic subgroups.

7 Qualitative and Error Analysis

Our system missed approximately 40% of ground truth spans overall, and around 30% of non-decision spans were misclassified as *Defining problem*. Minority classes were the most difficult to predict. The most extreme case was Model 2 (DFT + GRPO), which correctly identified only 2 out of 41 *Gathering additional information* spans (4.88%). This likely reflects the influence of class imbalance across data splits (Appendix A) on the model’s behavior. The most common decision-to-decision misclassification was *Evaluating test results* being predicted as *Defining problem*, occurring at a rate of 6.8 – 9.1%. We hypothesize that this is due to the fact that evaluating a test result often requires inferring the current state of a condition, thereby blurring the boundary between test-result evaluation and defining clinical problems. The qualitative example from the validation set is illustrated in Figure 4.

8 Discussion

The results collectively demonstrate that encoder baselines perform well on Token F_1 by design, while zero-shot GPT-5 highlights that LLMs can effectively extract span-level information but still struggle in token-level settings. Inline-tag fine-tuning bridges this gap, as even LoRA SFT outperforms the baselines across metrics. Among the fine-

Ground Truth	Prediction
<p>FINDINGS: <evaluate_result> There is an enhancing mass in the left parietal dural mass, which extends to the calvarium </evaluate_result>. <evaluate_result> There is minimal, if any, mass effect on the posterior frontal and anterior parietal parenchyma </evaluate_result>. <evaluate_result> No intraparenchymal metastatic lesions are identified </evaluate_result>. <evaluate_result> Gliosis with evidence of laminar necrosis is again seen in the medial right occipital lobe </evaluate_result>, likely <define_problem> sequela of a chronic infarction </define_problem>.</p>	<p>FINDINGS: <evaluate_result> There is an enhancing mass in the left parietal dural mass, which extends to the calvarium </evaluate_result>. <evaluate_result> There is minimal, if any, mass effect on the posterior frontal and anterior parietal parenchyma </evaluate_result>. <define_problem> No intraparenchymal metastatic lesions are identified </define_problem>. <define_problem> Gliosis with evidence of laminar necrosis is again seen in the medial right occipital lobe, likely sequela of a chronic infarction </define_problem>.</p>

Figure 4: Most common decision-to-decision misclassification: *Evaluating test results* being predicted as *Defining problem*.

tuning strategies, Qwen3.5-4B LoRA with DFT rank 256 consistently emerged as the strongest single-model configuration. The reinforcement learning results behaved as theoretically expected, with each reward type optimizing its corresponding metric, reinforcing the importance of aligning reward signals with evaluation objectives. The pseudo-label experiments further revealed sensitivity to label noise: a tighter entropy threshold (P_{10}) produced cleaner pseudo-labels and meaningfully improved minority-class recall. Finally, the ensemble’s strong overall performance, driven by complementary prediction patterns between models trained with DFT, DFT + GRPO, and DFT with pseudo-labels demonstrates that ensembling effectively compensates for individual model weaknesses. Test-set results further show that the system generalizes consistently across demographic subgroups, with the Worst Group F_1 not arising from disproportionately small subgroups.

9 Conclusion

Our work demonstrates that the agreement-driven ensemble consistently outperforms both encoder-based and zero-shot generative baselines across all metrics. This suggests that combining predictions from models with different output distributions can better capture complementary representations. We further show that performance can be improved through fine-tuning with DFT, reinforcement learning with verifiable rewards, and pseudo-label augmentation, highlighting the value of combining complementary modeling strategies with targeted training enhancements.

Limitations

Our study has several limitations spanning methodology, data, and experimental scope. First, full-document generation with inline tags is more computationally expensive at inference time than classifier-based BIO tagging, which may limit practicality in clinical settings. Although we use a fidelity reward to reduce hallucinations, this mechanism has not been tested across diverse inputs, and its robustness remains uncertain. Second, our experiments are limited to Qwen3.5-4B and a single dataset, MedDec (451 documents), leaving generalizability an open question. Third, we observed overlapping spans within the same category, including cases with slightly different offsets and cases with identical boundaries. Cleaning these annotations led to worse performance than retaining the original data, likely because overlapping spans are also present in the validation set. This issue warrants further investigation. Fourth, pseudo-label thresholding remains a challenge. Our results show slightly negative correlations, suggesting the current strategy does not reliably distinguish useful pseudo-labels from noisy ones. Because entropy captures only model confidence, it may be insufficient as the sole filtering criterion. Finally, time constraints limited our exploration of reinforcement learning policies, and the current setup may not reflect the optimal policy or ensemble configuration. Although Model 2 shows notable potential for reward-based improvement, this study prioritized ensemble methods, and a deeper investigation of reward configuration is left for future work.

Acknowledgments

We thank Pattaramanee Arsomngern for her thoughtful feedback on this manuscript, and to Kunat Pipatanakul and Chompakorn Chaksangchai for their discussions on model development.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Peter J Bickel and Kjell A Doksum. 2015. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC.

Michael Han Daniel Han and Unsloth team. 2023. [Unsloth](#).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024. Meddec: A dataset for extracting medical decisions from discharge summaries. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16442–16455.

Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Mehrnaz Sadrolashrafi, Mitra Mohtarami, Adrian Wong, Hadi Amiri, and Leo A. Celi. 2026. [Overview of medical decision extraction, analysis, and classification task \(medexact\) 2026](#). In *The 25th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, San Diego, California, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. [Domain-specific language model pretraining for biomedical natural language processing](#).

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: Decoding-enhanced bert with disentangled attention](#). In *International Conference on Learning Representations*.

David Heineman, Yao Dou, and Wei Xu. 2024. Improving minimum bayes risk decoding with multi-prompt. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22525–22545.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, and 1 others. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi,

- and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.
- Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. *Albert: A lite bert for self-supervised learning of language representations*. In *International Conference on Learning Representations*.
- Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and 1 others. 2020. Pytorch distributed: Experiences on accelerating data parallel training. *arXiv preprint arXiv:2006.15704*.
- Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.
- Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. 2026. *Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization*. *Preprint*, arXiv:2601.05242.
- Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Eirik H Ofstad, Jan C Frich, Edvin Schei, Richard M Frankel, and Pål Gulbrandsen. 2016. What is a medical decision? a taxonomy based on physician statements in hospital encounters: a qualitative study. *BMJ Open*, 6(2):e010098.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Qwen Team. 2026. *Qwen3.5: Towards native multi-modal agents*.
- John Schulman and Thinking Machines Lab. 2025. *Lora without regret*. *Thinking Machines Lab: Connectionism*. <https://thinkingmachines.ai/blog/lora/>.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Galouédec. 2020. *TRL: Transformers Reinforcement Learning*.
- Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. Gpt-ner: Named entity recognition via large language models. In *Findings of the association for computational linguistics: NAACL 2025*, pages 4257–4275.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. *Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.
- Shuhei Watanabe. 2023. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127*.
- Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. 2026. *On the generalization of SFT: A reinforcement learning perspective with reward rectification*. In *The Fourteenth International Conference on Learning Representations*.

Qiyong Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. [DAPO: An open-source LLM reinforcement learning system at scale](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Exploratory Data Analysis

We examined the demographic breakdown of the train, validation, and test splits. Table 4 highlights the severe underrepresentation of some groups, particularly across ethnicities, which can substantially affect the Worst Group F_1 score. Table 5 provides additional context on the prevalence of each decision category in the dataset, offering insight into observed model behavior.

B Internal Validation Set

We reassigned all spans from a section shown in Figure 5 from *Treatment goal* to *Defining problem* in the original validation set for internal validation. The original span boundaries were preserved, and only the category labels were changed.

C Decision Extraction Prompt

We designed the inline tag extraction prompt shown in Figure 6. This prompt was used for supervised fine-tuning, RLVR fine-tuning, and GPT-5 zero-shot evaluation.

D Encoder Model Training Configurations

For the encoder baselines, we adopted MedDec’s sliding-window training and non-overlapping window inference strategy while modifying several training hyperparameters. The window size for each model was set to its maximum supported input length: 512 for RoBERTa, ELECTRA, and BioMedBERT; 4096 for Clinical Longformer; and 8192 for ModernBERT. With these settings, RoBERTa, ELECTRA, and BioMedBERT were trained on a single NVIDIA A100 80GB GPU, whereas Clinical Longformer and ModernBERT required greater computational resources and were

trained using 4 NVIDIA A100 40GB GPUs. All models were trained for 250 epochs with a learning rate of 2×10^{-5} , an effective batch size of 32, and a cosine learning rate scheduler.

Category	Subcategory	Train	Validation	Test
Gender	Male	204	27	28
	Female	146	26	20
Ethnicity	White	244	43	35
	African American	31	7	4
	Hispanic	21	1	1
	Asian	10	0	2
	Native Hawaiian	1	0	0
	Other	43	2	6
Language	English	197	35	28
	Non-English	153	18	20

Table 4: Dataset distribution by category and subcategory across train, validation, and test sets.

Class	Train	%Train	Validation	%Validation
Contact related	2293	5.28	289	4.12
Gathering additional information	376	0.87	41	0.58
Defining problem	17015	39.19	2766	39.39
Treatment goal	107	0.25	46	0.66
Drug related	10930	25.17	1989	28.33
Therapeutic procedure related	5311	12.23	825	11.75
Evaluating test results	5806	13.37	828	11.79
Deferment	84	0.19	10	0.14
Advice and precaution	1496	3.45	228	3.25

Table 5: Number of spans per class in the train and validation sets.

Excerpt of the Internal Validation Set

[...] **Review of systems:** (+) Per HPI (-) Denies fever, chills, night sweats, recent weight loss or gain. Denies headache, sinus tenderness, rhinorrhea or congestion. Denies cough, shortness of breath, or wheezing. Denies chest pain, chest pressure, palpitations, or weakness. Denies nausea, vomiting, diarrhea, constipation, abdominal pain, or changes in bowel habits. Denies dysuria, frequency, or urgency. Denies arthralgias or myalgias. Denies rashes or skin changes.

Past Medical History: EtOH Abuse Cirrhosis Hepatitis C: No prior treatment Diabetes Mellitus 2 - 20 + years Tobacco Use Depression Hypertension GERD Pancreatitis Diverticulitis Hemorrhoids Atypical chest pain

Social History: - Tobacco: 1 ppd x 20+ years - Alcohol: 6-12 beers daily - Illicits: None

Family History: No history of bleeding disorders or abdominal bleeding. Both parents still living. **Physical Exam:** **Vitals:** T: 97 BP: 127/54 P: 112 R: 18 O2: 96/RA **General:** Alert, oriented, no acute distress **HEENT:** Sclera anicteric, MMM, oropharynx clear **Neck:** supple, JVP not elevated, no LAD **Lungs:** Clear to auscultation bilaterally, no wheezes, rales, ronchi **CV:** Regular rate and rhythm, normal S1 + S2, no murmurs, rubs, gallops **Abdomen:** soft, non-tender, non-distended, bowel sounds present, no rebound tenderness or guarding, no organomegaly **GU:** no foley **Ext:** warm, well perfused, 2+ pulses, no clubbing, cyanosis or edema [...]

Figure 5: Reassigned spans in the internal validation set. Only spans in this section originally categorized as *Treatment goal* were reassigned to *Defining problem*.

Decision Extraction Prompt

You are an expert specializes in extracting clinical decisions from a patient's discharge summary.

YOUR TASK

Given an input discharge summary, return the EXACT SAME text, but with specific phrases wrapped in inline tags to mark clinical decisions.

IMPORTANT:

- Do NOT add, remove, or rephrase any text outside the tags.
- Preserve all original punctuation, line breaks, and spacing.
- EVERY opening tag MUST have a corresponding closing tag (e.g., `<drug_decision>Aspirin 81 mg daily</drug_decision>`).
- These tags CAN overlap or nest in one another, as long as they are VALID TAGS.

DECISION CATEGORIES & TAGS

Use the following tags exactly as defined:

1. `<define_problem>` : diagnostic conclusions, health state evaluations, etiological inference, or prognostic judgment.
2. `<drug_decision>` : decisions to start, stop, continue, withhold, or modify medications.
3. `<evaluate_result>` : interpretation of clinical findings or test results.
4. `<contact_related>` : admissions, discharges, follow-ups, or referrals to other hospitals.
5. `<therapeutic_procedure>` : decisions to perform, plan, or refrain from procedures.
6. `<advice_and_precaution>` : patient instructions, advice, or precautions.
7. `<gather_info>` : decisions to order tests and investigations or consult another colleague.
8. `<treatment_goal>` : therapeutic goals, aims, or treatment objectives.
9. `<defer_decision>` : delaying judgment or action for now.

ANNOTATION RULES

1. **Boundary:** Annotate spans that capture the full clinical decision. Prefer longer spans than short words.
2. **Comprehensiveness:** The output should be comprehensively annotated. Extract as many valid decisions as possible.
3. **Exclusions:** DO NOT annotate document headers or labels (e.g., "Admission Date:", "Discharge Date:", "Physical Exam:").
4. **Overlapping Spans:** Spans may overlap or belong to multiple categories. Wrap each span independently with all applicable tags.

Example (nested): `<drug_decision>continue warfarin for <treatment_goal>stroke prevention</treatment_goal></drug_decision>`

Example (partial overlap): `<define_problem>The next previous examination suggested <evaluate_result>atelectasis</define_problem> - density in the left base cannot be evaluated</evaluate_result>`

OUTPUT FORMAT

Return ONLY the fully annotated text. Ensure all tags are properly closed. Do not include any explanations.

INPUT TEXT

{discharge_summary}

Figure 6: Decision extraction prompt used in our system and for zero-shot evaluation of GPT-5.

CS_Metro at PsyDefDetect: Detecting Psychological Defense Mechanisms in Mental Health Dialogues with Summarization-Enhanced Transformer Ensembles

Oarisa Rebayet*, Radiul Walee*, Symom Hossain Shohan,
Kawsar Ahmed and Mohammed Moshiul Hoque

Department of Computer Science and Engineering
Chittagong University of Engineering and Technology

u2104129@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

Detecting psychological defense mechanisms in supportive conversations is essential for assisting mental health practitioners. Natural language processing techniques are increasingly integral to such systems, enabling automated classification of defense levels to better understand help-seeker behavior and resistance patterns. In PsyDefDetect at BioNLP 2026, we address the task of nine-class defense level classification on the PSYDEFCONV corpus. We propose a three-stage pipeline combining LLM-based dialogue summarization, domain-specific transformer fine-tuning, and rule-based ensemble prediction. Additionally, we evaluate three mental health domain-specific transformers (Mental-BERT, Mental-RoBERTa, Mental-XLNet) alongside fine-tuned LLMs (Qwen3-4B, Qwen3-1.7B, Mistral-7B) under different input conditions. Experimental results on the released test-set gold labels show that our ensemble approach achieves the best performance, reaching 34.69% macro F1 and surpassing the baseline by 4.69 percentage points. On the official PsyDefDetect Leaderboard 1 (labels 1–8), the submitted system achieved a Macro-F1 score of 23.46%, ranking 15th out of 21 teams, while on Leaderboard 2 (labels 0–8), it achieved 30.04%, securing 14th place. These findings demonstrate that domain-specific transformers substantially outperform generic LLM fine-tuning on this specialized clinical task.

1 Introduction

Psychological defense mechanisms are automatic cognitive and behavioral strategies that individuals employ to manage emotional distress and interpersonal conflict (Vaillant, 1992). While defenses serve adaptive functions, their rigid or excessive use is strongly associated with poor mental health outcomes, reduced help-seeking behavior, and resistance to therapeutic intervention (?). In clinical

and supportive conversations, the ability to automatically identify defense levels in a speaker’s utterances could assist mental health practitioners in monitoring client progress, tailoring interventions, and reducing annotation burden in large-scale clinical studies. Despite the growing application of natural language processing (NLP) to mental health (Chancellor and Choudhury, 2020), the automatic detection of psychological defense mechanisms remains largely underexplored. The task is inherently challenging, as psychological defense mechanisms are highly dependent on contextual information, and remain difficult to reliably assess even for trained clinicians.

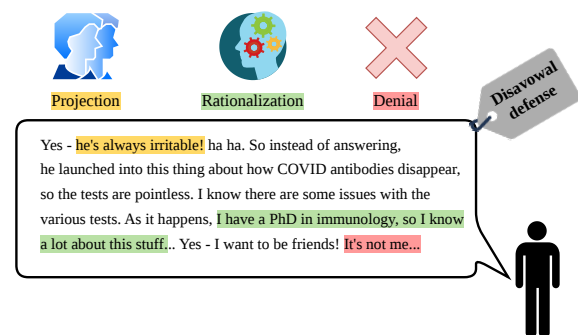


Figure 1: An example of defense level classification. Highlighted spans show the linguistic signals that indicate a Disavowal defense (Label 3) in the target utterance.

The recently introduced PSYDEFCONV corpus (Na et al., 2026b), comprising 200 therapeutic dialogues across nine defense levels, establishes a benchmark for this problem. However, even strong language models achieve only around 30% macro F1 in zero-shot and fine-tuning settings, and they consistently tend to overpredict mature defenses due to class imbalance. As illustrated in Figure 1, the task requires identifying such defense-indicative signals within helpseeker utterances to assign one of nine DMRS-defined defense levels.

To address these challenges, we participated in Psy-DefDetect@BioNLP 2026 (Na et al., 2026a), a shared task on detecting psychological defense levels in supportive conversations. Our contribution are as follows:

- We propose a dialogue summarization strategy using Qwen3-4B-Instruct (Team, 2025) to reduce conversational noise and provide global context for utterance-level defense classification.
- We systematically evaluate mental health domain-specific transformer models and fine-tuned LLMs, providing comparative analysis across all nine defense levels. Our results show that transformers with summarization guidance achieve more competitive performance than raw LLM fine-tuning on this specialized clinical task.
- We introduce a rule-based ensemble that leverages per-class model strengths identified through validation analysis, achieving more balanced predictions across defense levels.

2 Literature Review

Modeling mental health conversations poses unique challenges for NLP systems due to their context dependent nature, and has recently been explored extensively in the context of large language models for psychotherapy (Na et al., 2025). The ESConv corpus (Liu et al., 2021) laid the foundation for this area, providing 1,053 dialogues annotated with eight support strategies. Extended ESConv and ESConv-SRA (Madani and Srihari, 2025) generate strategy conditioned continuations and analyze how large language models (LLMs) maintain coherence across turns. Multi-Strategy ESConv (Bai et al., 2026) demonstrates that LLMs outperform supervised models in producing multi-strategy replies. Additionally, Zheng et al. (2023a) present AugESC, which finetunes GPT-J 6B on ESConv to complete dialogue threads, while Zheng et al. (2023b) create ExTES via ChatGPT in-context generation. Zhang et al. (2024) introduce ESD-CoT, extending ESConv with explicit reasoning chains.

Defense mechanisms are unconscious strategies individuals employ to manage emotional distress (Freud, 1936). The Defense Mechanism Rating Scales (DMRS) (Perry et al., 1993) is a validated, empirically grounded instrument that arranges defenses into a seven level hierarchy from Level 1

(Action defenses) to Level 7 (High Adaptive defenses) (Perry and Henry, 2004; Di Giuseppe and Perry, 2021). The PSYDEFCONV corpus (Na et al., 2026b), the first conversational dataset annotated with DMRS defense levels comprises 2,336 labeled utterances across nine defense classes. However, strong language models achieve only around 31% macro F1 in fine tuning settings, with a marked tendency to overpredict the dominant High Adaptive class (51.8% of the dataset) due to severe class imbalance.

Previous work on defense mechanism detection has either focused on supporting strategies in emotional support conversations or applied general mental health classification techniques without explicitly modeling defensive functioning. We address this gap through a three-stage pipeline: LLM-based dialogue summarization to reduce noise, domain specific transformer fine tuning with class imbalance handling, and a rule-based ensemble to counteract overprediction of the dominant class.

3 Dataset and Task Description

PsyDefDetect@BioNLP 2026 (Na et al., 2026b) is a shared task focused on analyzing psychological defense mechanisms in supportive conversations. The task requires identifying the defense level of a target utterance within its multi-turn dialogue context, spanning 9 categories described in Appendix 4. Table 1 summarizes the corpus statistics.

Corpus Statistic	Train	Test
Dialogues	200	189
Utterances	1,864	472
Average turns per dialogue	22.3	16.9
Average words per utterance	18.9	18.3
Seeker-to-Supporter ratio	1.04:1	1.06:1

Table 1: Dataset statistics for train and test splits.

4 System Overview

This study addresses psychological defense mechanism classification from therapeutic dialogues as a nine-class prediction problem. Our pipeline comprises three sequential stages: LLM-based dialogue summarization, domain-specific transformer fine-tuning, and a rule-based ensemble that combines model predictions. Figure 2 represents the architecture of the system. The implementation and source code are publicly available on GitHub.¹

¹https://github.com/0arisaR/CS_Metro-at-PsyDefDetect-BioNLP-2026

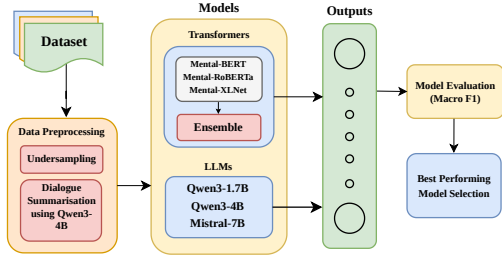


Figure 2: Overview of the proposed methodology pipeline, including dialogue summarization, transformer-based classification, LLM fine-tuning, and evaluation.

4.1 Data Preprocessing

The dataset presents two primary challenges: irrelevant or noisy information that can misguide the model and a notable class imbalance. To mitigate these issues, we apply dialogue summarization to extract the most relevant content from each conversation and use class imbalance handling techniques to ensure more balanced learning across all classes.

Dialogue Summarization: Each dialogue was summarized using **Qwen3-4B-Instruct** to capture the main intent of the conversation. The dialogues were first formatted by concatenating speaker turns as "Speaker: text" and then passed to the model with instructions to produce a concise summary within 100 words. The structured prompt used for this summarization step is provided in Appendix 7.

Class Imbalance Handling: The training data exhibits a strong class imbalance, with Label 7 (High-Adaptive) originally comprising 968 samples, making it disproportionately dominant. To mitigate this effect, we apply targeted undersampling, restricting Label 7 to a maximum of 550 samples. However, since undersampling alone does not fully resolve the imbalance across the remaining classes, we further compute inverse-frequency class weights on the undersampled data.

$$w_c = \frac{N}{9 \times \text{count}(c)} \quad (1)$$

where N is the total number of training samples and $\text{count}(c)$ is the number of samples in class c .

4.2 Models

We evaluate two categories of models for multi-label classification:

Domain-Specific Transformer Models We employ transformer models which are pretrained on mental-health related corpora such as Mental-BERT-base and Mental-RoBERTa-base (Ji et al., 2022), and Mental-XLNet-base (Ji et al., 2023).

Fine-tuned Large Language Models We further fine-tune instruction-tuned LLMs, including Qwen3-1.7B, Qwen3-4B (Team, 2025), and Mistral-7B² (Jiang et al., 2023), to evaluate their effectiveness on the task.

4.3 Training

The models are trained end-to-end using the cross-entropy loss function. All experiments follow a unified hyperparameter configuration, as summarized in Table 2.

Hyperparameter	Value
Epochs	20
Learning rate	2×10^{-5}
Batch size	16
Weight decay	0.01

Table 2: Training hyperparameters.

4.4 Rule-Based Ensemble

We construct a rule-based ensemble based on per-class performance observed on the validation set. The models exhibit complementary strengths across labels, with predictions selected from the model that is relatively more reliable for a given class. In cases where no clear preference emerges, the ensemble defaults to the most stable model. Final predictions are stored along with individual model outputs for evaluation.

5 Results and Discussion

Table 3 summarizes the performance of all models on the test set across labels (0–8). The results correspond to evaluation on the released test-set gold labels. In the official PsyDefDetect Leaderboard 1 (labels 1–8), our submission achieved a Macro-F1 score of 23.46%, ranking 15th out of 21 teams, while in Leaderboard 2 (labels 0–8), it achieved a Macro-F1 score of 30.04%, securing 14th place.

²<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>

<i>Transformers</i>					
Model	Approach	Acc	Rec	Prec	F1
MentalBERT	Raw	56.36	28.80	28.25	28.36
	Summ	52.97	33.32	31.53	31.21
	Δ	-3.39	+4.52	+3.28	+2.85
MentalRoBERTa	Raw	57.84	33.21	31.79	32.20
	Summ	54.45	31.56	30.66	29.57
	Δ	-3.39	-1.65	-1.13	-2.63
MentalXLNet	Raw	46.40	23.17	21.69	22.17
	Summ	51.06	33.38	27.94	29.65
	Δ	+4.66	+10.21	+6.25	+7.48
Ensemble	Summ	51.48	40.00	36.00	34.69
<i>LLMs (Few-shot, 9-example)</i>					
Llama-3.2-3B-Instruct	Raw	30.72	22.15	18.64	14.65
	Summ	31.14	22.99	19.51	18.25
	Δ	+0.42	+0.84	+0.87	+3.60
Qwen3-4B-Instruct	Raw	20.34	21.19	21.83	16.15
	Summ	29.03	21.72	24.95	19.48
	Δ	+8.69	+0.53	+3.12	+3.33
Phi-4-mini-instruct	Raw	33.47	18.14	20.29	16.73
	Summ	34.96	23.60	22.27	19.46
	Δ	+1.49	+5.46	+1.98	+2.73
<i>LLMs (Fine-tuned)</i>					
Qwen3-4B	Raw	63.77	31.31	36.73	30.04
Qwen3-1.7B	Raw	54.45	30.19	25.27	25.58
Mistral-7B	Raw	47.03	29.39	26.56	26.59

Table 3: Performance comparison of transformer-based models and LLMs (both fine-tuned and few-shot), reported in %. 'Summ' denotes summarisation of raw conversations; 'Raw' uses full dialogue history; few-shot uses 9 examples per label.

Summarization reduces input noise but affects transformer models differently across architectures. The results suggest that dialogue summarization has a non-uniform effect on transformer-based models. MentalBERT and MentalRoBERTa show a small decrease in accuracy (-3.39 for both), while recall and F1 increase, indicating that summarization helps recover relevant information but may slightly affect precision. In contrast, MentalXLNet improves across all metrics, with a notable gain in F1 (+7.48) and recall (+10.21), suggesting better use of compressed dialogue context.

Model architecture influences sensitivity to compressed dialogue representations. The results show that transformer variants respond differently to summarization. MentalRoBERTa performs best on raw dialogues but degrades more with summarized inputs, whereas MentalXLNet benefits more consistently from compression. MentalBERT remains comparatively stable but achieves lower over-

all performance. Overall, stronger contextual representations appear more effective at handling compressed inputs.

Large language models leverage longer context windows but remain competitive rather than outperforming transformer ensembles. Fine-tuned LLMs showed strong performance, with Qwen-4B (QLoRA) performing best, followed by Qwen-1.7B and Mistral-7B. Unlike encoder-based transformers, these models were trained on raw dialogues, as their larger context windows reduce the need for summarization and make them less sensitive to irrelevant conversational context. However, despite this advantage, LLMs did not surpass the transformer ensemble, suggesting that carefully optimized transformer models with summarization remain highly effective for this specialized classification task.

Few-shot learning remains below fine-tuned baselines. Although summarization improves performance, the best few-shot F1 score (19.48) remains well below the transformer ensemble (34.69) and fine-tuned Qwen3-4B (30.04). This gap suggests that limited in-context examples are insufficient for capturing the nuanced patterns in mental health dialogues, emphasizing the importance of task-specific fine-tuning.

5.1 Per-Class Behaviour Analysis

Figure 3 illustrates the class-wise performance distribution across all models. In particular, No Defense and High-Adaptive exhibit higher accuracy across all models, indicating that these classes dominate the decision space. In contrast, defense categories such as Disavowal and Neurotic show low performance.

The observed distribution suggests a strong bias toward high-frequency, while minority and psychologically nuanced categories remain unstable across architectures. This indicates that the classification boundary is primarily shaped by data imbalance rather than model capacity.

5.2 Error Analysis

Figure 4 presents the confusion matrix of the ensemble system, highlighting systematic error patterns across all classes. The misclassification structure is highly non-uniform and reveals strong semantic entanglement between several defense mechanism categories.

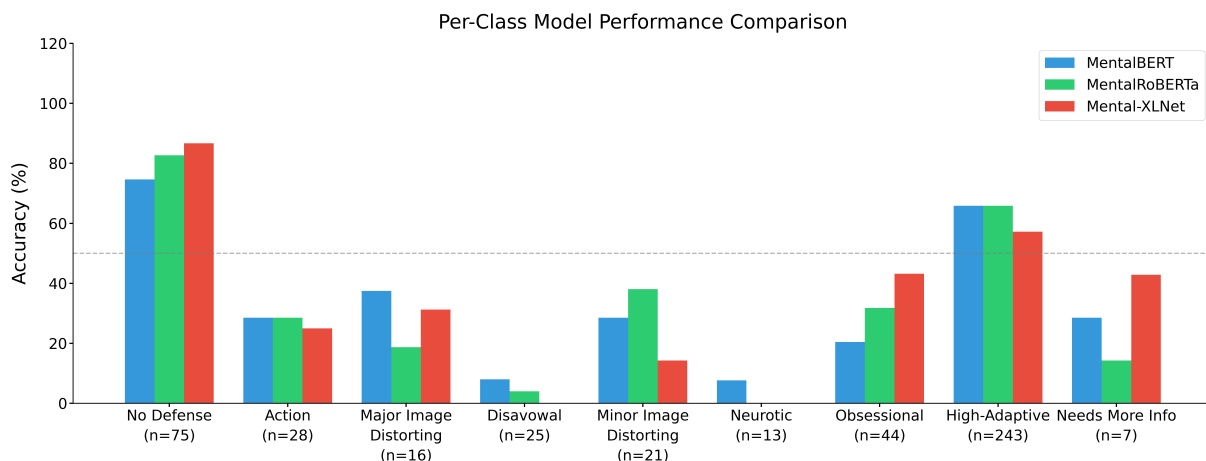


Figure 3: Per-class accuracy comparison across models.

True Class \ Predicted Class	NoDef	Action	MajDist	Disavow	MinDist	Neurotic	Obsess	HighAdapt	NeedInfo
NoDef	55	0	0	0	0	2	1	16	1
Action	0	11	4	0	3	1	5	4	0
MajDist	0	0	5	0	4	0	3	4	0
Disavow	1	5	3	2	8	0	1	5	0
MinDist	0	1	4	0	3	2	3	8	0
Neurotic	1	1	0	0	1	1	0	9	0
Obsess	1	4	1	1	2	2	15	18	0
HighAdapt	10	9	8	1	11	3	21	177	3
NeedInfo	0	1	0	0	0	0	1	4	1

Figure 4: Confusion matrix of ensemble predictions over Labels 0–8.

A clear error pattern is the consistent bias toward High-Adaptive, which behaves as a dominant sink class. Samples from Disavowal, Minor Image Distorting, Neurotic, and Obsessional classes are often misclassified as High-Adaptive, suggesting that uncertain cases are frequently mapped to a high-frequency category.

There is also notable confusion between Neurotic and High-Adaptive, indicating that the model does not clearly separate these two and may treat them as closely related states.

Minor Image Distorting is particularly unstable, with predictions spread across several nearby classes, showing weak and inconsistent learning for this category.

6 Conclusion

This study investigates psychological defense mechanism classification within conversations us-

ing transformer-based models and fine-tuned large language models. MentalBERT, MentalRoBERTa, and MentalXLNet were evaluated under raw and summarized settings alongside an ensemble approach. The results indicate that summarization consistently benefits MentalXLNet, whereas MentalBERT and MentalRoBERTa exhibit mixed behavior. Fine-tuned LLMs, including Qwen-4B, Qwen-1.7B, and Mistral-7B, were evaluated on raw dialogues leveraging their larger context windows. These models do not consistently outperform the transformer ensemble. Future research could explore synthetic augmentation for minority classes, dialogue compression strategies, and cross-lingual transfer to enhance generalization.

Limitations

While this study contributes to psychological defense mechanism classification in supportive conversations, several limitations exist. Summarization may discard subtle emotional cues critical for understanding psychological distress. Resource constraints prevented exploration of larger LLMs, alternative compression techniques, multi-turn context, and cross lingual transfer. Models were evaluated on a single dataset (PSYDEFCONV), which may not generalize across diverse populations. Finally, parameter-efficient fine-tuning may not fully utilize the representational capacity of larger models compared to full fine-tuning.

Ethical Considerations

The system presented in this study are intended to support, not substitute, clinical professionals. Predictions are subject to potential misclassifica-

tion and may inherit biases present in the training data. Real-world deployment necessitates human oversight and compliance with applicable data protection regulations.

Acknowledgements

We express our sincere gratitude to the organizers of the PsyDefDetect@BioNLP 2026 shared task for conducting the shared task and for providing the dataset and evaluation framework. We also acknowledge the support of the Directorate of Research and Extension and the NLP Lab at Chittagong University of Engineering and Technology (CUET), Bangladesh, for their continuous encouragement and support.

References

- Xin Bai, Guanyi Chen, Tingting He, Rui Fan, Chenlian Zhou, and Yu Liu. 2026. [Emotional supporters often use multiple strategies in a single turn](#). *Neurocomputing*, 693:133897.
- Stevie Chancellor and Munmun De Choudhury. 2020. [Methods in predictive techniques for mental health status on social media: a critical review](#). *npj Digital Medicine*, 3(1):43.
- Mariagrazia Di Giuseppe and J. Christopher Perry. 2021. [The hierarchy of defense mechanisms: Assessing defensive functioning with the defense mechanisms rating scales q-sort](#). *Frontiers in Psychology*, 12:718440. PMID: 34366995.
- Sigmund Freud. 1936. [Inhibitions, symptoms and anxiety](#). *The Psychoanalytic Quarterly*, 5:1–28. Original work published 1936.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. [MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare](#). In *Proceedings of LREC*.
- Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, Erik Cambria, and Jörg Tiedemann. 2023. [Domain-specific continued pretraining of language models for capturing long context in mental health](#). *arXiv preprint arXiv:2304.10447*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.
- Navid Madani and Rohini Srihari. 2025. [Steering conversational large language models for long emotional support conversations](#). In *Proceedings of the Third Workshop on Social Influence in Conversations (SICoN 2025)*, pages 109–123, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. [Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations](#). In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. [You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.
- J. Christopher Perry, Marianne E. Kardos, and Christopher J. Pagano. 1993. [The study of defenses in psychotherapy using the defense mechanism rating scales \(dmrs\)](#). In *The Concept of Defense Mechanisms in Contemporary Psychology: Theoretical, Research, and Clinical Perspectives*, pages 122–132.
- John Perry and Melissa Henry. 2004. [Studying defense mechanisms in psychotherapy using the defense mechanism rating scales](#). *Advances in Psychology*, 136.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- George E. Vaillant. 1992. *Ego Mechanisms of Defense: A Guide for Clinicians and Researchers*. American Psychiatric Press. Original work published June 1, 1992.
- Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. [ESCoT: Towards interpretable emotional support dialogue systems](#). In *Proceedings*

of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. *AugESC: Dialogue augmentation with large language models for emotional support conversation*. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023b. *Building emotional support chatbots in the era of llms*. *ArXiv*, abs/2308.11584.

A Appendix

A.1 Defense Mechanism Labels and Dataset Distribution

Table 4 presents the definition of nine defense mechanism labels along with their class distribution across train and test splits.

Label	Level	Definition	Train	Test
0	No Defense (NoDef)	No defensive content.	296	75
1	Action (Action)	Distress acted out externally.	108	28
2	Major Image Distorting (MajDist)	All-good or all-bad distortions.	61	16
3	Disavowal (Disavow)	Denial, projection, fantasy.	99	25
4	Minor Image Distorting (MinDist)	Idealization or devaluation.	84	21
5	Neurotic (Neurotic)	Motives kept out of awareness.	48	13
6	Obsessional (Obsess)	Logic separates feeling from event.	172	44
7	High-Adaptive (HighAdapt)	Mature emotional coping.	968	243
8	More Info Needed (NeedInfo)	Too ambiguous to annotate.	28	7
Total			1864	472

Table 4: DMRS label taxonomy with train/test distribution (Perry et al., 1993).

A.2 QLoRA Fine-Tuning Configuration

We employ parameter-efficient fine-tuning using QLoRA to adapt a 4-bit quantized Qwen3-4B-Instruct model for the classification task.

Category	Configuration
Quantization	4-bit NF4 with bfloat16 compute
LoRA Rank (r)	16
LoRA Scaling (α)	32
Target Modules	Attention and MLP projections
Epochs	3
Learning Rate	1×10^{-4}
Effective Batch Size	8
Max Sequence Length	512 tokens

Table 5: QLoRA hyperparameter configuration for fine-tuning Qwen3-4B-Instruct.

A.3 Few-Shot Prompt Structure

All evaluated models used a standardized 9-shot prompting framework for defense mechanism clas-

sification. Each prompt included a fixed system instruction, nine in-context examples and a final target query. The complete prompt structure is shown below.

System Prompt

You are a clinical psychologist specializing in psychodynamic defense mechanisms. Classify the psychological defense mechanism in the given therapy exchange.

Defense levels (choose exactly one):

- 0 – No Defense: No clear defense mechanism observed.
- 1 – Action: Acting out, passive aggression, help-rejecting complaining.
- 2 – Major Image-Distorting: Splitting, projective identification, autistic fantasy.
- 3 – Disavowal: Denial, projection, rationalization.
- 4 – Minor Image-Distorting: Devaluation, idealization, omnipotence.
- 5 – Neurotic: Repression, reaction formation, displacement.
- 6 – Obsessional: Intellectualization, isolation of affect, undoing.
- 7 – High-Adaptive: Humor, sublimation, anticipation, altruism, affiliation.
- 8 – Needs More Info: Not enough context to classify.

Your response must end with exactly: **Label: (digit)**

Example: Label: 3

No explanation after the label.

A.4 Few-Shot Examples

To construct the few-shot prompting framework, one representative demonstration example was selected for each defense mechanism category (labels 0–8). Each example comprises an LLM-generated session summary, a target utterance, and its corresponding gold-standard label. The demonstrations were formatted using a unified chat-based structure to maintain consistency across all evaluated models. The complete set of nine in-context examples is presented in Table 6.

Table 6: Nine-shot in-context examples. *Session summary* is an LLM-generated condensation of the full conversation history; *Target utterance* is the text the model must label.

L	Class	Session summary	Target utterance
0	No Defense	The dialogue begins with a simple greeting from the Seeker. Overall conclusion: no meaningful points, insights, or conclusions can be drawn due to lack of content beyond a greeting.	<i>"hi"</i>
1	Action	The seeker feels isolated and lethargic due to the cold season and lack of family or friends, living in a hotel during the pandemic. They struggle with motivation to work or write and have no social connections. The supporter suggests planning Christmas activities to break the cycle of loneliness, but the seeker acknowledges they have no support system currently. Key insight: isolation and lack of social connection significantly impact emotional well-being, highlighting the need for intentional connection despite current circumstances.	<i>"I live in a hotel currently. No family or friends around, especially with the virus."</i>
2	Major Image-Distort	The seeker expresses feelings of depression and self-doubt, struggling to feel like a good person and believing they are a failure. The supporter validates their pain, offers empathy, and affirms the seeker's potential to become anyone they desire. The key insight is that while the supporter is not a professional, they provide emotional support and encourage self-worth, suggesting that personal transformation is possible despite current feelings of failure.	<i>"I am not sure of that. I feel like a failure."</i>
3	Disavowal	The seeker experiences overwhelming anxiety and depression without a clear trigger, fueled by overthinking. The supporter emphasizes the importance of professional help and shares that engaging in enjoyable hobbies helps distract from racing thoughts. Both acknowledge that seasonal changes, like winter, can worsen symptoms. Key insight: consistent self-care, hobbies, and professional support are vital for managing anxiety and depression on bad days.	<i>"That makes a lot of sense. Winter makes it tougher too I think."</i>
4	Minor Image-Distort	The seeker feels isolated and lethargic due to the cold season, lack of family, and pandemic-related restrictions. Living in a hotel with no nearby support, they struggle with motivation and loneliness, despite being a writer. While they try dating sites, they find them impersonal. The supporter suggests planning a positive future event, like a Christmas gathering next year, to break the cycle of isolation and build hope.	<i>"I try the dating sites, but they seem too far away or impersonal. Just the current season has me down."</i>
5	Neurotic	The seeker expresses deep frustration and emotional distress over recurring negative events, particularly their mother's situation in Mexico with a deported partner and financial mismanagement. The supporter validates their feelings and asks about past efforts. Key insight: the seeker struggles with a sense of helplessness and cycles of worsening circumstances, highlighting emotional exhaustion and a lack of perceived progress in personal or familial situations.	<i>"Hm. I tried moving on. And it just appears she gets worse."</i>
6	Obsessional	The seeker expresses stress, anxiety, and financial struggles due to the pandemic and job market challenges. They feel overwhelmed and emotionally drained, seeking any work opportunity. The supporter empathizes, acknowledges shared difficulties, and asks about the seeker's preferred type of work. Key insight: widespread economic hardship is impacting mental and financial well-being, and emotional support and understanding are crucial. The conversation highlights the need for empathy and tailored support in times of crisis.	<i>"anything really atm and sorry to hear that—it's tough everywhere, a lot of people having a hard time def."</i>
7	High-Adaptive	The seeker struggles with chronic back pain, anxiety, and depression for eight years, compounded by grief over their mother's death and pandemic-related delays in medical care. The supporter emphasizes the importance of talking through emotions, journaling, and seeking support. The dialogue highlights resilience, the value of emotional support, and hope amid ongoing challenges.	<i>"Thank you, I will look there for others going through the same things, it doesn't sound silly at all and you are an amazing person too. Thank you for listening and your thoughtful help."</i>
8	Needs More Info	The seeker feels shamed by friends for not donating to charity during Christmas, due to financial struggles from COVID-19. The supporter emphasizes that donations should come from genuine capacity, not guilt. The seeker should contribute based on personal values and circumstances, not others' expectations—pride in self-worth is essential.	<i>"Yeah..."</i>

A.5 Prompt Design for Dialogue Summarization

Table 7 presents the structured prompt used for LLM-based dialogue summarization. It defines a template in which a raw dialogue is provided as input and the model generates a concise summary capturing the key information. The prompt emphasizes clarity and semantic consistency, ensuring that the generated summary reflects the core content of the dialogue without introducing unnecessary details.

Instruction Prompt
Please summarize the following dialogue in a clear and concise manner, capturing the main points, key insights, and overall conclusions. Your summary should be within 100 words, focusing on the core ideas without unnecessary details. Dialogue: {dialogue} Summary:

Table 7: Structured prompt used for LLM-based dialogue summarization.

A.6 Example of Dialogue Summarization

To illustrate the effectiveness of the proposed prompt, an example of input dialogue and the corresponding generated summary is presented in Table 8.

Example: Input–Output Pair
Dialogue: <ul style="list-style-type: none">• Supporter: hi• Seeker: Hey You there?• Supporter: Hi I stepped out, how are you?• Seeker: I'm okay I guess. How are you?• Supporter: Good, I am good. Is there anything I can be of an assistance today?• Seeker: I don't know. I've been dealing with a lot this last month. Generated Summary: <p>The dialogue begins with a friendly greeting and mutual check-in. The supporter expresses well-being and offers assistance, while the seeker reports experiencing significant challenges over the past month. This indicates emotional strain and a need for support. Overall, the interaction reflects a supportive conversational exchange.</p>

Table 8: Example of input dialogue and corresponding generated summary using the proposed prompt.

A.7 Prediction Examples

Table 9 illustrates sample predictions for the dialogue classification task. In this table, sample text inputs are presented along with their corresponding summaries, actual labels, and predicted labels from the ensemble model. This provides a qualitative view of how the model performs across different input examples by comparing the true labels with the predicted outputs.

Summary	Current Text	Actual Defense	Predicted Defense
The seeker reports chronic severe back pain causing significant anxiety and depression. The supporter offers assistance, indicating a need for emotional and possibly medical support. Key insight: persistent physical pain is deeply impacting mental health, highlighting the interconnection between physical suffering and emotional well-being. Conclusion: Addressing both the back pain and its psychological effects is essential for holistic recovery.	Hi I'm experiencing a lot of anxiety and depression because of constant severe back pain.	High-Adaptive	High-Adaptive
The seeker feels depressed, attributing it to pandemic-related isolation and lack of connection with loved ones. Despite fears of vulnerability, the supporter encourages open communication, noting that sharing emotions can relieve sadness and strengthen relationships. The key insight is that vulnerability is healthy and worthy—true love accepts one as they are. The seeker resolves to be brave and express their feelings to their boyfriend.	That is true. I should be brave enough to find out if he will be there when the chips are down. I am going to do it.	High-Adaptive	Obsessional
The seeker expresses feelings of depression and lack of motivation, attributing it to being stuck in one place. The supporter asks what caused the depression, indicating a need to explore underlying causes. The core insight is that the seeker feels emotionally stagnant and unmotivated, possibly due to a lack of change or progress in their current situation. The key takeaway is that identifying the root cause of stagnation is essential for addressing the depression.	I don't know. Maybe stuck at one place	Neurotic	Neurotic

Table 9: Sample predictions with actual and predicted defense levels.

Sparse Category Routing and Fairness-Aware Optimization for Medical Decision Extraction

Ahmed Elshehaby*, Mohamed Abdalla*, Youssef Mohamed†
Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI)
{ahmed.elshehaby, mohamed.abdalla, youssef.mohamed}@mbzuai.ac.ae

Abstract

Extracting structured medical decisions from ICU discharge summaries is hard because of long documents, severe category imbalance across nine DICTUM decision types, and a fairness-aware evaluation that penalizes inconsistent performance across demographic subgroups. We present our system for the MedExACT 2026 shared task (Elgaar et al., 2026), which fine-tunes BiomedBERT with a composite loss combining label-smoothed cross-entropy, a soft token-F1 auxiliary term, and R-Drop regularization. At inference time we apply a deterministic ensemble: half-offset sliding-window augmentation across four window configurations, dual-branch logit aggregation from the same checkpoint, per-category length calibration on the Anchor Branch, and sparse routing to a context-weighted specialist branch for Cat7, which has strong span-length evidence, and Cat4, which is included only as a rare-category heuristic. In a single-seed validation run, adding R-Drop improved Overall_F1 by 1.24 points over the CE + soft-F1 baseline, with a larger 1.70-point gain on Worst-Group F1; these gains should be interpreted as single-run observations rather than variance-estimated effects. Our best submission achieves Span F1 of 0.4900, Token F1 of 0.6796, and an official Overall_F1 of 0.5724, with the African American subgroup as the Worst-Group bottleneck at Base_Score 0.5601.

1 Introduction

ICU clinicians record dozens of decisions per patient in discharge summaries, from drug orders and referrals to test-result evaluations and disposition plans. The DICTUM taxonomy (Ofstad et al., 2016) defines ten decision types. The MedDec dataset (Elgaar et al., 2024) provides span-level

annotations over 451 de-identified MIMIC-III discharge summaries (Johnson et al., 2016) with inter-annotator agreement of Cohen’s $\kappa = 0.74$, and the MedExACT 2026 shared task restricts scoring to nine of these as in-scope (Categories 1–9). The MedExACT 2026 shared task (Elgaar et al., 2026) challenges participants to extract these spans and classify each into one of the nine categories under a fairness-aware metric that penalizes inconsistent performance across demographic subgroups.

The task is harder than standard clinical NER for three reasons. Notes average 1,571 whitespace tokens, well beyond the 512-token context of standard BERT encoders. The nine in-scope shared-task categories are severely imbalanced in the official 350-document training split: Category 8 (Deferment) has only 84 annotated spans, whereas Category 3 (Defining problem) has 17,015. And the ranking metric (Overall_F1) is the average of aggregate Base_Score and Worst-Group F1, so improving the aggregate at the expense of any single demographic subgroup directly costs points.

We describe a system based on a single fine-tuned BiomedBERT checkpoint (Figure 1). Training uses a composite loss that adds soft token-F1 and R-Drop regularization to label-smoothed cross-entropy. Inference applies four half-offset sliding-window views to reduce boundary truncation, filters Anchor-branch predictions for three categories by refined span length, and then replaces Cat7 and Cat4 predictions with a context-weighted specialist branch. Cat7 routing is supported by span-length analysis, while Cat4 routing is a rare-category heuristic with weak distributional evidence. The system achieves Overall_F1 of 0.5724 on the official test set.

*Equal contribution.

†Corresponding author.

2 Related Work

Clinical NER and span extraction. Prior work on MedDec established that fine-tuned token classifiers far outperform zero-shot LLMs: MedDecXtract (Elgaar et al., 2025) reached 34.8 span F1 with fine-tuned RoBERTa vs. 4.8 for one-shot Llama-3.1-8B. BIO labeling over pretrained transformers is the standard approach for flat clinical NER (Gu et al., 2022), with domain-specific pretraining consistently outperforming general encoders. SpanNER (Fu et al., 2021) and the boundary-smoothing loss of Zhu and Li (2022) point to span-level and boundary-aware modeling as a more direct fit for this task: among detected spans, the dominant failure mode is boundary drift rather than category confusion (Section 7). Overlap-capable extraction (Li et al., 2021) would also apply since many labeled MedDec documents contain overlapping spans, but BIO tagging cannot represent these.

Long-document inference. ERNIE-Doc (Ding et al., 2021) addresses full-document context via retrospective attention. Practical sliding-window inference with half-offset TTA was the most reliable option in our setting.

Regularization and fairness-aware selection. R-Drop (Liang et al., 2021) was our single largest training gain in the validation run. Group DRO (Sagawa et al., 2020) and Just Train Twice (Liu et al., 2021) target training-time worst-group robustness; Group DRO destabilized training in our setting (proxy span F1: 15.4 vs. 39.9 baseline), so our fairness-aware contribution is limited to selecting checkpoints and post-processing settings by Overall_F1 rather than introducing an effective training-time fairness method.

3 Task and Evaluation

Given a full ICU discharge summary, systems must detect contiguous text spans expressing medical decisions and assign each span one of nine in-scope DICTUM categories (Elgaar et al., 2026). The evaluation combines performance and subgroup robustness. The Base_Score averages span- and token-level F1:

$$\text{Base_Score} = \frac{\text{Span F1} + \text{Token F1}}{2} \quad (1)$$

Worst-Group F1 is the minimum subgroup Base_Score across nine demographic partitions (sex: Female/Male; race: White, African American,

Encoder	P	R	F1
BiomedBERT fulltext	36.2	42.8	39.2
BioClinical-ModernBERT-lg	34.7	44.7	39.1
BiomedBERT abstract	36.0	42.6	39.0
BioClinical-ModernBERT-base	33.3	42.2	37.2
SciBERT	33.8	41.2	37.1
Bio Discharge Summary BERT	33.8	40.7	36.9
BlueBERT	31.5	40.4	35.4
Bio-GottBERT	30.2	39.9	34.4
Clinical_ModernBERT	10.6	22.4	14.4

Table 1: Backbone comparison (span-level precision, recall, F1 on proxy scorer, 5,000 training steps, seed 42).

Hispanic, Asian, Other; language: English/Non-English):

$$\text{Worst-Group F1} = \min(\{\text{Base_Score}_i\}_{i=0}^8) \quad (2)$$

The final ranking metric is:

$$\text{Overall_F1} = \frac{\text{Base_Score} + \text{Worst-Group F1}}{2} \quad (3)$$

A model that improves aggregate Base_Score while degrading any subgroup scores lower on Overall_F1 than one that simply holds the worst group steady.

4 System Description

4.1 Backbone and Task Formulation

We treat medical decision extraction as BIO token labeling, assigning each token a label from $\{B-k, I-k, O\}$ where $k \in \{1, \dots, 9\}$, yielding 19 labels. We use BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext (Gu et al., 2022) as the encoder, pre-trained on both PubMed abstracts and full texts. In a preliminary comparison across multiple biomedical and clinical encoders (Table 1), BiomedBERT fulltext produced the highest span F1 (39.2%) and was selected as the foundation for all subsequent experiments.

A BIO-constrained Viterbi decoder enforces legal tag transitions at inference time: the start state allows only O or any $B-*$; $I-k$ is permitted only after $B-k$ or $I-k$ (same-category continuation only). BIO tagging is imperfect for this dataset: when two overlapping spans share a token, only one annotation can be represented, so some gold spans are silently dropped during training.

4.2 Text Normalization and Preprocessing

MIMIC-III discharge summaries contain formatting artifacts that affect both tokenization and span

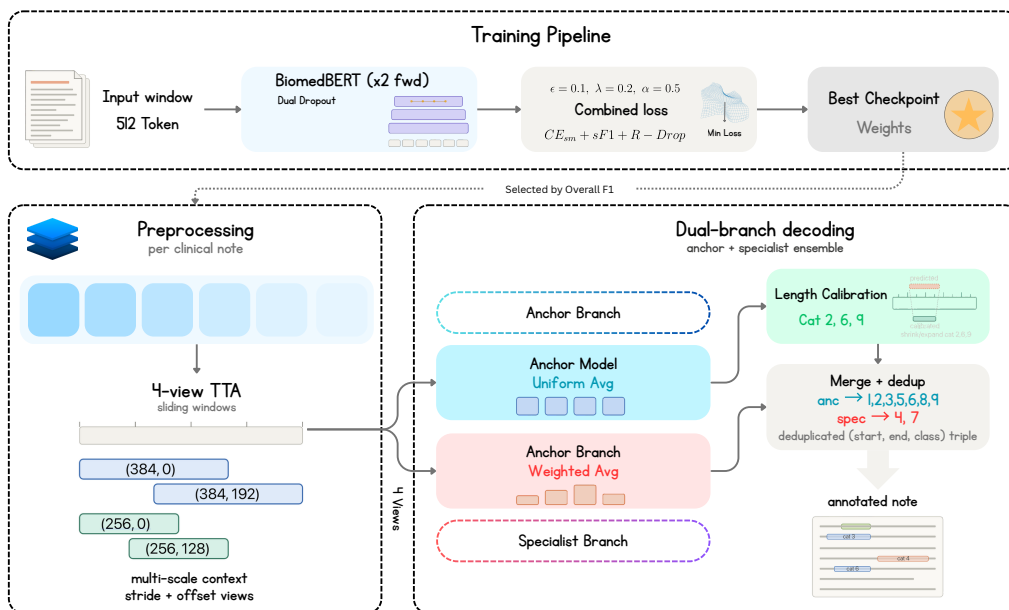


Figure 1: System pipeline. **Training:** 512-token windows (biased toward annotated regions) pass twice through BiomedBERT under independent dropout. A composite loss is backpropagated for 5500 steps; the best checkpoint by Overall_F1 is kept. **Inference:** four half-offset sliding-window views feed an Anchor Branch (uniform avg) and a Specialist Branch (weighted avg). Length calibration is applied to the Anchor Branch only, filtering predictions for Categories 2, 6, and 9 by refined character length. The merge step keeps Categories 1, 2, 3, 5, 6, 8, 9 from the calibrated Anchor Branch and substitutes Categories 4 and 7 from the Specialist Branch, followed by deduplication over (s, e, cat) triples.

boundary alignment. We apply a three-step preprocessing pipeline before training.

Character normalization. Unicode whitespace variants (non-breaking spaces U+00A0, figure spaces U+2007, narrow no-break spaces U+202F, BOM U+FEFF) are mapped to ASCII space. Typographic quotes (U+2018/U+2019/U+201C/U+201D) are folded to ASCII equivalents, and en-dashes, em-dashes, and minus signs (U+2013/U+2014/U+2212) are replaced with hyphens. Bullets (U+2022) are likewise mapped to hyphens. All replacements are single-character to single-character so that character offsets are preserved without remapping.

Boilerplate filtering. An extractive line-level filter removes administrative headers (*Admission Date, Discharge Date, Sex, Service, Attending*, etc.) and section markers (*Discharge Medications, Discharge Disposition, Followup Instructions*) that never contain decision spans. Blank lines and decorative separators (____, -, ==) are also removed. Lines that overlap any gold annotation span are always kept, regardless of boilerplate matching, so no annotated text is lost. A bidirectional character-

level offset map tracks the mapping between original and shortened documents, allowing predictions on the shortened text to be projected back to original offsets for evaluation.

Whitespace compaction. Runs of consecutive spaces or newlines that do not overlap an annotation span are collapsed: multi-space runs become a single space, and runs containing a newline become a single newline. Runs within annotation spans are left untouched to preserve byte-identical span text. This step reduces document length by 8–15% on average without affecting any gold span content.

Span refinement at evaluation. The official evaluator applies a `refine_span` routine during matching, which expands boundaries that cut through an alphanumeric word and strips leading and trailing punctuation tokens using NLTK’s `TreebankWordTokenizer`, preserving de-identified PHI markers (`[**.*]**`). In the final routed inference recipe, this routine was used only to compute refined character length for category-specific filtering (Cat2, Cat6, Cat9) on the Anchor Branch; the final submitted offsets remained the

original decoded span boundaries.

4.3 Training Objective

The composite loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}^{\text{smooth}} + \lambda \mathcal{L}_{\text{soft-F1}} + \alpha \mathcal{L}_{\text{R-Drop}} \quad (4)$$

with label smoothing $\epsilon=0.10$, $\lambda=0.20$, $\alpha=0.50$.

Label-smoothed cross-entropy. Applied over all BIO positions to reduce overconfidence on the dominant O label, which accounts for $>85\%$ of tokens.

Soft token-F1. A differentiable macro soft-F1 over token-category probabilities, computed by summing $P(B-k) + P(I-k)$ for each category and forming soft true positives, false positives, and false negatives:

$$\mathcal{L}_{\text{soft-F1}} = 1 - \frac{1}{9} \sum_{k=1}^9 \frac{2 \text{TP}_k + \epsilon}{2 \text{TP}_k + \text{FP}_k + \text{FN}_k + \epsilon} \quad (5)$$

This matches the training signal to the token-level component of the evaluation metric.

R-Drop. Each window passes through the model twice under independent dropout masks. A symmetric KL divergence is added between the two token-logit distributions on the 19 BIO label positions of valid (non-padding) tokens:

$$\mathcal{L}_{\text{R-Drop}} = \frac{1}{2} [\text{KL}(p_1 \| p_2) + \text{KL}(p_2 \| p_1)] \quad (6)$$

where p_1, p_2 are the softmax distributions from the two forward passes. The primary losses are averaged across both passes before backpropagation.

Figure 2 illustrates the interaction of the three loss terms. Training uses online patch sampling with a positive-window bias probability of 0.70, drawing 512-token windows anchored near annotated spans (with ± 10 -character fallback for tokenizer boundary alignment) rather than feeding full documents. Mixed precision (fp16 autocast/GradScaler) is enabled on CUDA, with gradient clipping at max norm 1.0.

4.4 Long-Document Inference and TTA

A single 512-token sliding window with a fixed stride misses spans that fall near window boundaries, a non-trivial risk given that documents average 1,571 whitespace tokens. We use a four-configuration TTA ensemble with window configurations applied in order: $(384, 0)$, $(384, 192)$,

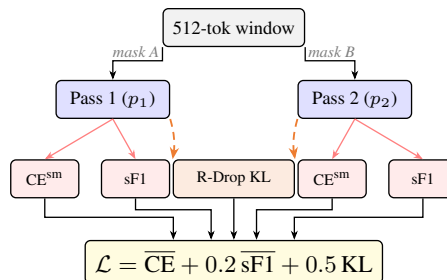


Figure 2: R-Drop training. Each window passes through BiomedBERT twice under different dropout masks. Both passes produce CE and soft-F1 losses; the R-Drop term adds a symmetric KL penalty between p_1 and p_2 . Losses are averaged across passes (overbar) before back-propagation.

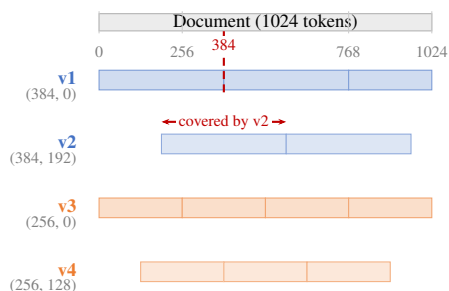


Figure 3: Half-offset TTA tiling for a 1024-token document. Each row shows sliding windows for one view (w, o) . **Blue**: 384-token windows (v1–v2); **orange**: 256-token windows (v3–v4). The dashed line marks where v1 splits at token 384; v2’s half-offset places this position mid-window (red arrow), so every token is covered under at least two different boundary alignments.

$(256, 0)$, $(256, 128)$, where the notation (w, o) denotes window size w and starting half-offset o . Figure 3 shows how the four views tile across a document. Token logits from overlapping windows within each view are averaged in *logit space* (before softmax and before decoding) at the document level. The four views cover each document position under at least two different boundary alignments, reducing edge-truncation artifacts without retraining.

4.5 Sparse Category-Routing Ensemble

Statistical analysis of gold character-level span lengths provides a clear routing motivation for Cat7, but not for Cat4. Figure 7 visualizes these distributions. Cat7 (Evaluating test result) has a mean span of 79.9 characters and a median of 60, compared to a median of 28 across the anchor-pool categories 3, 5, and 6. A Mann-Whitney test confirms this separation for Cat7 ($p < 10^{-16}$, rank-

biserial = -0.40). Cat4 has much weaker length evidence ($p = 0.21$); we include it in the Specialist Branch only because it is extremely rare (107 training spans, 0.25% of all training annotations) and because development runs suggested that the context-weighted branch could recover occasional Cat4 spans. We therefore treat Cat4 routing as an exploratory low-frequency-category heuristic, not as a statistically justified span-length intervention.

The **Anchor Branch** handles categories 1, 2, 3, 5, 6, 8, and 9 via uniform aggregation over the four TTA views, i.e. view weights $[0.25, 0.25, 0.25, 0.25]$ in logit space. The **Specialist Branch** handles categories 4 and 7 via weighted aggregation with weights $[0.4, 0.4, 0.1, 0.1]$ in view order, biasing toward broader context windows. Broader routing searches over all nine categories consistently overfit the 53-document validation set; two-category routing was the most stable intervention found.

4.6 Character-Level Span-Length Calibration

Before sparse category routing, the Anchor Branch is filtered with category-specific character-length thresholds derived from training span statistics, after which Categories 4 and 7 are replaced by Specialist-branch predictions and the final span set is deduplicated. Thresholds are applied to *refined character length* after a text-boundary repair step that expands boundaries cut through alphanumeric words and strips leading/trailing non-content tokens (using `TrebankWordTokenizer`), preserving PHI-style `[**...**]` markers. Three categories have constraints, applied only to the Anchor Branch. Thresholds were chosen by a small grid search on the validation split, seeded near the corresponding tail percentile of training-span lengths (p95 for upper bounds, p05 for lower bounds): category 2 drops spans longer than 98 characters (training p95 is 91 chars; the validation optimum sat slightly above this percentile); category 6 drops spans shorter than 3 characters; category 9 drops spans shorter than 24 characters (training p05 is 32 chars; 24 was preferred on validation because it avoided deleting short but valid spans). The final span set is deduplicated over exact (s, e, cat) triples. No overlap suppression or containment deletion is applied, since many labeled documents contain overlapping gold span pairs and blunt deletion damages recall.

4.7 Fairness-Aware Model Selection

We use Overall_F1 as the selection criterion for checkpoint selection, TTA weight tuning, and calibration threshold search. This aligns development decisions with the shared-task metric, so aggregate gains are not preferred when they reduce Worst-Group F1. It does not impose fairness constraints during optimization. Group DRO (Sagawa et al., 2020) was explored during development but achieved only proxy span F1 of 15.4%, most likely because category-level grouping was too coarse. Our fairness-aware component is therefore model and hyperparameter selection by Overall_F1, not a successful training-time fairness method. R-Drop (added for regularization) raises Worst-Group F1 more than Base_Score in our single validation run, by 1.70 vs. 0.77 points (Section 7), but this should be read as an observed robustness gain rather than proof of a fairness-optimized training recipe.

5 Experimental Setup

5.1 Data

The MedDec resource (Elgaar et al., 2024) contains 451 de-identified MIMIC-III discharge summaries (Johnson et al., 2016). For MedExACT, we restrict training and evaluation to the nine in-scope categories (Categories 1–9), excluding Categories 10 and 11 present in the broader MedDec annotations. We use 350 documents for training, 53 for validation, and 48 held-out test documents released by the shared-task organizers for final evaluation. The public release contains 43,640 training and 7,044 validation annotations. Restricting to the nine in-scope shared-task categories yields the 43,418 training spans in Table 2 and 7,022 validation spans; the 222 training annotations and 22 validation annotations excluded are Categories 10 and 11, plus a small number of annotations released as TBD. Table 2 shows the category distribution in the training set.

5.2 Implementation Details

We train for 5,500 forward/backward iterations with gradient accumulation 2, corresponding to 2,750 optimizer updates and 275 linear warmup steps (10% of optimizer updates). Maximum sequence length is 512 tokens. Training time ranges from 38 minutes (CE variants) to 64 minutes (R-Drop) on a single NVIDIA T4 (16 GB VRAM). The full hyperparameter list is in Appendix A; compute and reproducibility details, including the train-

Cat	DICTUM Type	Spans	%
1	Contact related	2,293	5.28
2	Gathering additional information	376	0.87
3	Defining problem	17,015	39.19
4	Treatment goal	107	0.25
5	Drug related	10,930	25.17
6	Therapeutic procedure related	5,311	12.23
7	Evaluating test result	5,806	13.37
8	Deferment	84	0.19
9	Advice and precaution	1,496	3.45
Total		43,418	100.00

Table 2: Category distribution in the official training split (350 documents, 43,418 in-scope spans across Categories 1–9). Categories 10 and 11 are present in the broader MedDec annotations but are out of scope for MedExACT and were excluded. Categories 2, 4, and 8 are extremely rare, together accounting for only 1.31% of all spans.

ing and inference cost of R-Drop and 4-view TTA, are in Appendix B. All experiments use a fixed random seed (42); no variance estimates are available, so small deltas and component gains should not be interpreted as statistically reliable. Validation scores reported in Table 3 use an internal proxy scorer that matches spans by normalized text rather than strict character offsets, and are not directly comparable to the official evaluator used in Table 5.

5.3 Baselines

We compare against the RoBERTa baseline from Elgaar et al. (2024), which achieves span F1 of 34.8 and token-level accuracy of 79.9 on the original MedDec test split (that evaluation uses a different split and scorer than MedExACT, so numbers are not directly comparable).

6 Results

6.1 Training Objective Comparison

Table 3 reports single-seed proxy validation scores for four training objectives. R-Drop + CE + soft-F1 produced the highest scores on all three metrics in this run. The CE + soft-F1 baseline reached Overall_F1 0.4996, 1.24 points below R-Drop, but this comparison lacks a seed sweep or confidence interval. Focal + soft-F1 matched the baseline on Base_Score (0.5441 vs. 0.5493) but fell further on Worst-Group F1 (0.4209 vs. 0.4499). Weighted CE + soft-F1 was the worst configuration across all metrics, with loss scale collapse at initialization (step-0 loss ≈ 0.04 vs. normal ≈ 2.9) destabilizing BIO training.

Training objective	Base	WG	Overall
CE + soft-F1	0.5493	0.4499	0.4996
Weighted CE + soft-F1	0.4652	0.3772	0.4212
Focal + soft-F1	0.5441	0.4209	0.4825
R-Drop + CE + soft-F1	0.5570	0.4669	0.5120

Table 3: Proxy validation scores for training objective comparison (0–1 scale). WG = Worst-Group F1. Overall = (Base + WG)/2. Scores use an internal proxy scorer; official test results are in Table 5. For reference, the RoBERTa baseline of Elgaar et al. (2024) achieves 34.8 span F1 on the original MedDec test split, which uses a different evaluation protocol and is not directly comparable.

Approach	Proxy Base	Δ
Baseline (BiomedBERT-ft, CE)	41.54	–
Focal loss ($\gamma=2.0$)	16.78	–24.8
Weighted CE (inv. freq.)	17.82	–23.7
Group DRO ($\eta=0.01$)	15.31	–26.2
BioLinkBERT-large (333M)	35.73	–5.8
Sliding window conv ($w=3,5,7$)	38.94	–2.6
Threshold calibration (per-cat)	31.67	–9.9

Table 4: Failed approaches tested during development. Proxy Base is reported on a 0–100 scale (i.e., percentage points), consistent with the baseline value of 41.54. Δ is the absolute difference from the CE baseline.

Table 4 lists additional configurations explored during development that did not improve over the CE baseline and were not carried forward.

6.2 Official Test Results

Table 5 shows official test set results. Our system placed fifth in the shared task with Overall_F1 of 0.5724. Token F1 of 0.6796 is the strongest among the top-five final-score submissions. We attribute this primarily to the soft token-F1 auxiliary loss, which optimizes the same token-level score that the official scorer reports. The result does not by itself validate the sparse routing design; routing and length calibration are better viewed as post-processing heuristics motivated by boundary and rarity analyses. The 19-point gap between Token F1 (0.6796) and Span F1 (0.4900) was consistent across all configurations and reflects the structural mismatch between token-level BIO labeling and exact character-span evaluation.

7 Analysis

7.1 Effect of R-Drop on Subgroup Robustness

The R-Drop term in Equation 4 accounts for 1.24 Overall_F1 points in the single-seed validation

Metric	Score
Span F1	0.4900
Token F1	0.6796
Base_Score	0.5848
Worst-Group F1	0.5601
Overall_F1	0.5724

Table 5: Official test set results. Base_Score = (Span F1 + Token F1)/2. Worst-Group subgroup is African American. Overall_F1 = (Base + WG)/2.

comparison (Table 3). The gain is larger on Worst-Group F1 (+1.70 points: 0.4499 to 0.4669) than on Base_Score (+0.77: 0.5493 to 0.5570). This is consistent with the intuition behind R-Drop (Liang et al., 2021): the KL consistency term reduces disagreement across dropout masks, which may help harder inputs where the model is less confident. Because only a single random seed (42) was used, statistical significance cannot be confirmed and the size of the R-Drop gain may be unstable.

7.2 Demographic Subgroup Performance

Table 6 and Figure 4 report per-subgroup test scores. Figure 4a shows a 10.1-point Base_Score gap between the best group (Hispanic, 0.6611) and the worst (African American, 0.5601). The African American subgroup has the lowest Span F1 (0.4639) and one of the lowest Token F1 scores (0.6562). Figure 4b also shows that the \approx 19-point aggregate Token–Span gap holds broadly across subgroups (per-subgroup range 14.7–21.8 points), so the gap is primarily a system-wide property rather than one concentrated in the worst group. Structural analysis reveals that African American notes are harder: mean length 1,756 tokens vs. 1,570 for White notes, 140.3 annotations per note vs. 121.6, and a Cat7 share of 15.4% vs. 11.6% ($p = 0.0095$, Mann-Whitney). Readability (Flesch-Kincaid: $p = 0.50$) and abbreviation density ($p = 0.59$) did not differ significantly, ruling out lexical complexity as the primary driver.

7.3 Boundary Error Analysis

Figure 5 breaks down span matches per category into exact matches, near-misses (correct category but shifted offsets), and complete misses. Out of 7,022 in-scope gold spans in the official validation split, 3,373 (48.0%) are exact matches, 1,449 (20.6%) are near-misses, and 2,200 (31.3%) are completely missing. Appendix C reports the same data as per-category recall.

Subgroup	Span	Token	Base
Hispanic	0.5856	0.7366	0.6611
Female	0.5279	0.7149	0.6214
Other	0.5414	0.6879	0.6146
Asian	0.4800	0.6982	0.5891
English	0.4903	0.6872	0.5887
White	0.4812	0.6782	0.5797
Non-English	0.4893	0.6690	0.5792
Male	0.4677	0.6544	0.5611
Afr. Am.	0.4639	0.6562	0.5601

Table 6: Official test set scores per demographic subgroup, sorted by Base_Score descending. The African American subgroup is the Worst-Group bottleneck. The 10.1-point gap between the best (Hispanic) and worst group motivates future work on training-time fairness interventions.

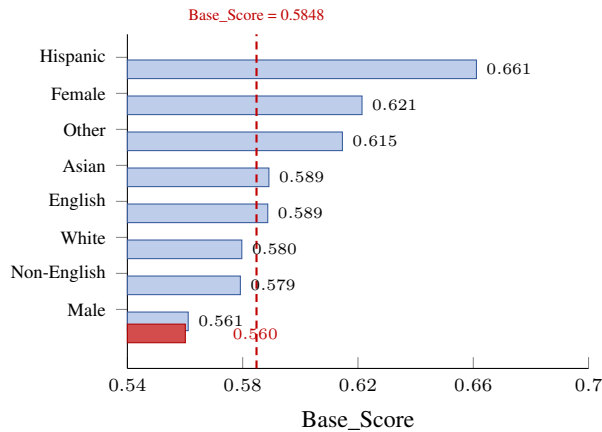
Left-boundary errors dominate: pred_too_short_left (642 instances) is the most frequent type, followed by pred_too_long_left (355). Figure 6 shows the full breakdown. Numbered bullets (“1.”, “2.”, “3.”) are the most common missed left-boundary tokens – the model excludes list-item prefixes that are inside the gold annotation boundary.

7.4 Span-Length Distribution and Routing Motivation

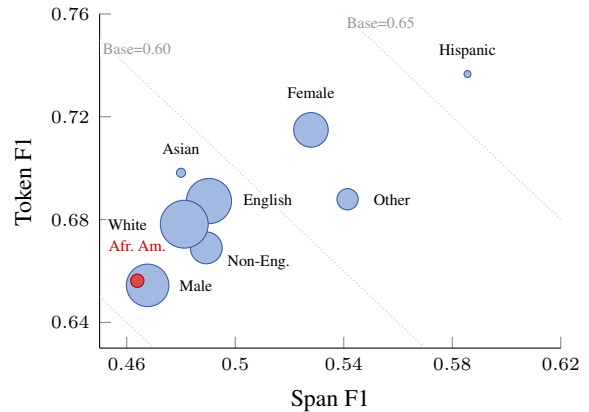
Figure 7 shows that Cat7 spans are much longer and more variable than other frequent categories. Cat4 does not show a significant length shift ($p = 0.21$) and is routed to the specialist branch only as a rare-category heuristic (107 training spans), where the weighted aggregation marginally recovers low-frequency spans in development runs. Cat9 (Advice and precaution) contains relatively long spans, motivating its 24-character minimum-length filter, which removes spurious short predictions.

7.5 Ablation Study

Table 7 reports single-seed incremental proxy validation scores as each component is added. These scores are *not* directly comparable to the official test results because the proxy scorer matches spans by normalized text content while the official scorer requires exact character offsets. R-Drop is the only component with a clear positive proxy delta in this run (+1.24 Overall, +1.70 WG), although this remains a single-seed observation. Adding TTA actually reduces proxy Base from 0.5570 to 0.5408, which at first looks counterproductive. The explanation is that TTA moves predicted span boundaries to better positions on average, which the official



(a) Base_Score by subgroup, sorted. The dashed line is the overall test-set Base_Score (0.5848); the red bar is African American, the worst group.



(b) Span F1 against Token F1 by subgroup, with marker area proportional to n . Dotted diagonals are lines of equal Base_Score. The ≈ 19 -point aggregate Token–Span gap holds broadly across subgroups (per-subgroup range 14.7–21.8 points), so the gap is primarily a system-wide property rather than one concentrated in any one subgroup.

Figure 4: Test performance by subgroup. (a) Base_Score per group, sorted, with the overall test-set Base_Score shown as a dashed line and the worst group in red. (b) Span F1 against Token F1 for the same groups, with marker area proportional to group size and dotted diagonals marking equal Base_Score. Hispanic ($n=89$) scores highest and African American ($n=313$) sets the Worst-Group F1 at 0.5601.

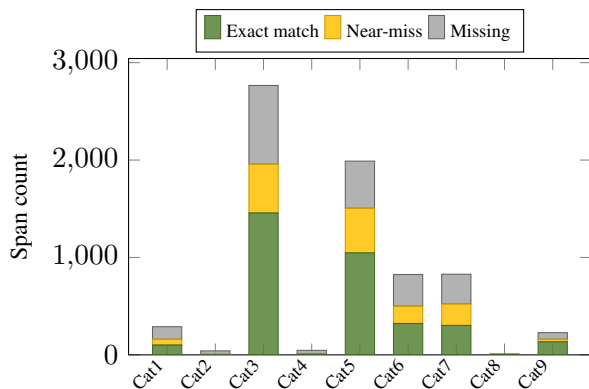


Figure 5: Per-category span match breakdown on the official validation split (53 documents; 7,022 in-scope gold spans across Categories 1–9). Near-misses (correct category, incorrect offsets) account for 20.6% of all gold spans, confirming boundary recovery as the dominant error mode. Categories 2, 4, and 8 are nearly invisible to the model due to extreme rarity.

character-offset scorer rewards but the proxy text-match scorer penalizes when the normalized content was already correct. TTA appeared beneficial under the official scorer, but a clean controlled final-system ablation was not recoverable from the saved artifacts. Length calibration shows no change on the proxy scorer because the constraints act on character lengths, which the text-match scorer ignores; its value is measured by the official evaluator. The

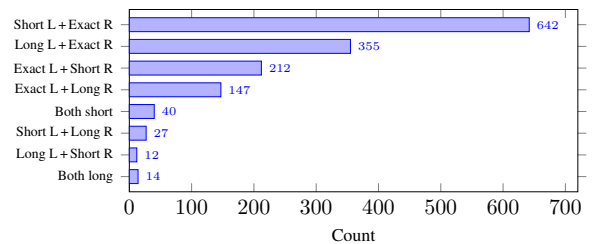


Figure 6: Directional breakdown of 1,449 boundary errors. Left-boundary errors (short or long on the left) account for 71.4% of near-misses. L = left boundary, R = right boundary.

final sparse-routing step changes Overall_F1 by only +0.0010 relative to length calibration on the proxy scorer, so it should not be interpreted as a robust validation gain.

7.6 Qualitative Error Analysis

The two categories routed to the Specialist Branch exhibit distinct failure patterns. Cat7 (Evaluating test result) spans are long (mean 79.9 chars) and often straddle chunk boundaries; a span like “CXR showing bilateral infiltrates, unchanged from prior” is frequently truncated to the first clause when a chunk boundary falls mid-sentence. The half-offset TTA and context-heavier Specialist Branch were designed to reduce this fragmentation. Cat4 (Treatment goal) is extremely rare (107 training spans); the model routinely predicts zero Cat4 spans on

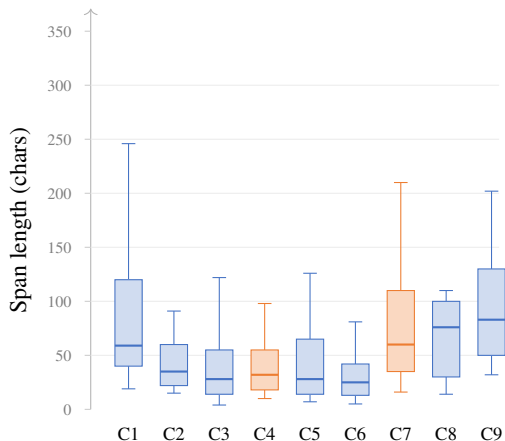


Figure 7: Gold span-length distributions per category (character count). **Orange**: categories routed to the Specialist Branch (C4, C7). **Blue**: Anchor categories. C7 has the widest IQR and highest median among frequent categories ($n > 40$).

Configuration	Base	WG	Overall
CE + soft-F1 only	0.5493	0.4499	0.4996
+ R-Drop ($\alpha=0.5$)	0.5570	0.4669	0.5120
+ Viterbi decoding	0.5570	0.4669	0.5120
+ 4-view TTA	0.5408	0.4640	0.5024
+ Length calibration	0.5408	0.4640	0.5024
+ Sparse routing (Cat 4, 7)	0.5407	0.4660	0.5034

Table 7: Incremental ablation (proxy scorer). TTA and length calibration appear neutral or negative on this scorer but were retained in the final official system: the proxy matches normalized text content and cannot capture exact-offset gains from TTA or length trimming. The sparse-routing delta is very small and is not treated as evidence of a stable improvement.

many notes, and the Specialist Branch provides a distributional shift that marginally recovers some of these low-frequency spans, but this choice is not supported by the length test.

We also observe clinically plausible false positives, suggesting that some errors reflect mismatch between model behavior and annotation guidelines rather than simple confidence miscalibration.

8 Conclusion

We have presented our MedExACT 2026 system for extracting nine-category medical decision spans from ICU discharge summaries. The system uses a single BiomedBERT checkpoint with no extra parameters at inference.

R-Drop was the strongest training decision in our single-seed validation comparison, adding 1.24 Overall_F1 points with a larger 1.70-point gain on Worst-Group F1. The larger gain on WG than on

Base suggests that the KL consistency term may help harder or underrepresented inputs where the model is least confident, but the absence of a seed sweep prevents a stability claim.

Four-view TTA and sparse category routing did not improve proxy validation scores, which are based on text-match rather than offset accuracy. These components are part of the best official system and are consistent with the observed boundary and long-span error patterns, although clean controlled final-system ablations were not recoverable. Within sparse routing, Cat7 has strong length-based support; Cat4 routing is an exploratory rare-category heuristic and should not be treated as statistically justified by span length.

The persistent 19-point gap between Token F1 (0.6796) and Span F1 (0.4900) shows that the model largely finds the right text but consistently gets the exact start and end characters wrong. This is a fundamental limitation of BIO token tagging, which cannot model overlap and has no direct loss on character offsets. Moving to span-boundary or span-classification architectures that directly optimize character-level match criteria is the natural next step.

Limitations

Single seed. All training runs use seed 42; no variance estimates are available and all reported gains should be interpreted as single-run observations. This particularly affects the reported R-Drop gain and the very small sparse-routing delta. A complete evaluation would report means and standard deviations across multiple seeds or confidence intervals from resampling.

Proxy vs. official scorer. Validation scores in Tables 3 and 7 use an internal proxy scorer that matches spans by normalized text rather than strict character offsets. The proxy scorer and the official shared-task evaluator differ in token-level scoring behavior, so validation and test numbers are not directly comparable.

Sparse routing evidence. Cat7 routing is supported by a significant span-length difference and a high near-miss rate, but Cat4 routing is not supported by a significant length difference ($p = 0.21$). Cat4 was included because it is rare and sometimes benefits from a context-weighted branch in development, so this part of the routing recipe should be viewed as heuristic. The Specialist Branch weights

[0.4, 0.4, 0.1, 0.1] were chosen to favor broader context windows but were not selected by a formal grid search.

Fairness-aware optimization scope. Our use of “fairness-aware optimization” should be interpreted as metric-aligned development rather than an effective training-time fairness method. The system’s fairness contribution is selection-time alignment by Overall_F1, not training-time robustness. Group DRO was explored and failed in this formulation. Dedicated subgroup training, lexical error analysis by subgroup, variance estimates, and subgroup-targeted significance testing remain as future work.

Ethics Statement

This work uses the MedDec dataset derived from MIMIC-III, which contains de-identified clinical notes from the Beth Israel Deaconess Medical Center. Access requires completion of a data use agreement and CITI training through PhysioNet. Our system processes only de-identified text and does not attempt re-identification. The fairness-aware evaluation metric used in MedExACT penalizes models that perform unevenly across demographic subgroups, and we report per-subgroup scores in full (Table 6). We note that our system addresses fairness only through model selection, not through training-time interventions, and the 10-point gap between the best and worst subgroups remains an open problem.

Acknowledgments

We thank the MedExACT organizers for providing the MedDec dataset and evaluation infrastructure.

References

Siyu Ding, Junyuan Shang, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. [ERNIE-Doc: A retrospective long-document modeling transformer](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2914–2927, Online. Association for Computational Linguistics.

Mohamed Elgaar, Hadi Amiri, Mitra Mohtarami, and Leo Anthony Celi. 2025. [MedDecXtract: A clinician-support system for extracting, visualizing, and annotating medical decisions in clinical narratives](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: Sys-*

tem Demonstrations), pages 481–489, Vienna, Austria. Association for Computational Linguistics.

- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024. [MedDec: A dataset for extracting medical decisions from discharge summaries](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16442–16455, Bangkok, Thailand. Association for Computational Linguistics.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Mehrnaz Sadrolashrafi, Mitra Mohtarami, Adrian Wong, Hadi Amiri, and Leo A. Celi. 2026. [Overview of medical decision extraction, analysis, and classification task \(medexact\) 2026](#). In *The 25th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, San Diego, California, USA. Association for Computational Linguistics. To appear.
- Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. [SpanNER: Named entity re-/recognition as span prediction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7183–7195, Online. Association for Computational Linguistics.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. [Domain-specific language model pretraining for biomedical natural language processing](#). *ACM Transactions on Computing for Healthcare*, 3(1):1–23.
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. [MIMIC-III, a freely accessible critical care database](#). *Scientific Data*, 3:160035.
- Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2021. [A span-based model for joint overlapped and discontinuous named entity recognition](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4814–4828, Online. Association for Computational Linguistics.
- Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tiejun Liu. 2021. [R-Drop: Regularized dropout for neural networks](#). In *Advances in Neural Information Processing Systems*, volume 34.
- Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. [Just train twice: Improving group robustness without training group information](#). In *Proceedings of the 38th International*

Conference on Machine Learning, pages 6781–6792. PMLR.

Eirik H. Ofstad, Jan C. Frich, Edvin Schei, Richard M. Frankel, and Pål Gulbrandsen. 2016. [What is a medical decision? A taxonomy based on physician statements in hospital encounters: A qualitative study.](#) *BMJ Open*, 6(2):e010098.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. [Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.](#) In *Proceedings of the International Conference on Learning Representations*.

Enwei Zhu and Jinpeng Li. 2022. [Boundary smoothing for named entity recognition.](#) In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7096–7108, Dublin, Ireland. Association for Computational Linguistics.

Appendices

A Hyperparameter Summary

Table 8 and Table 9 summarize the hyperparameters used in our final submitted system.

Training	
Training steps	5,500
Batch size	8 (eff. 16 w/ grad. accum. 2)
Learning rate	2×10^{-5}
Warmup steps	275 (10% of 2,750 optimizer updates)
Optimizer	AdamW
Weight decay	0.01
Epsilon	10^{-8}
Gradient clip	1.0 (max norm)
Max sequence length	512 tokens
Label smoothing ϵ	0.10
Soft-F1 weight λ	0.20
R-Drop weight α	0.50
Positive-window bias	0.70
Random seed	42
Precision	fp16 (autocast / GradScaler)

Table 8: Training hyperparameters for the final submitted system.

Inference & Post-processing	
TTA configurations	(384, 0), (384, 192); (256, 0), (256, 128)
Specialist weights	[0.4, 0.4, 0.1, 0.1] (logit space)
Cat. 2 max span length	98 chars
Cat. 6 min span length	3 chars
Cat. 9 min span length	24 chars
Cats. 1,3,4,5,7,8	unconstrained
Viterbi start state	O or any B -*
I - k allowed after	B - k or I - k only

Table 9: Inference, post-processing, and decoding constraints.

B Compute and Reproducibility

Hardware. All training and inference was run on an environment with two NVIDIA T4 GPUs (16 GB VRAM per card, 32 GB total), using fp16 autocast with GradScaler. The BiomedBERT-base encoder has approximately 110 M parameters; with the additional 19-way linear BIO classification head the full model fits comfortably on a single T4. Individual training runs used a single T4; the second card was reserved for parallel ablation sweeps.

Training cost. Single-pass training runs (every configuration in Table 3 except R-Drop + CE + soft-F1) complete the 5,500 forward/backward iterations in approximately 38 minutes. R-Drop adds a second forward pass per training step; the R-Drop + CE + soft-F1 configuration takes about 64 minutes. Total wall-clock time across the four training-objective comparison runs in Table 3 is under 3 hours on one T4.

Inference cost. Single-view inference (one sliding window with default stride) is the baseline cost unit. The 4-view TTA ensemble passes each document through the encoder with the four window configurations in Table 9, so inference compute scales roughly $4\times$ over single-view. The sparse routing step adds no extra model cost because both branches read from the same checkpoint; length calibration and deduplication are pure Python post-processing.

C Per-Category Validation Breakdown

Figure 5 reports per-category match counts on the 53-document validation split. Table 10 expresses those counts as exact and near-miss recall. Three patterns are visible.

Recall tracks training-set frequency. Categories 2, 4, and 8 have 376, 107, and 84 training spans respectively (Table 2), and reach exact recall of at most 0.152 on validation. Cat 3 alone has 17,015 training spans, more than 200 times Cat 8, and its exact recall is 0.527. A BIO classifier trained on this split cannot close that frequency gap without auxiliary supervision, so the rare categories contribute little to aggregate metrics.

Cat 7 fails at boundaries. Its exact recall (0.365) is 2.5 percentage points below Cat 6 (0.390), but its near-miss rate of 26.6% is the highest of any category. This is consistent with its long span-length distribution (median 60 characters, Figure 7): long

spans are more likely to have boundary errors under the strict span matcher. This is the empirical motivation for routing Cat 7 through the weighted Specialist Branch (Section 4.5).

Cat 9 benefits from length calibration. Its exact recall (0.588) is the highest of any category despite its mid-range training frequency (1,496 spans). The 24-character minimum-length filter (Section 4.6) discards short noisy predictions that would otherwise be scored as false positives under the strict matcher, which improves precision on Cat 9; the threshold was chosen on validation to avoid deleting short but valid spans, so recall is preserved.

Cat	Counts			Recall	
	Exact	Near	Miss	Exact	Exact+near
1	101	59	129	0.349	0.554
2	2	3	36	0.049	0.122
3	1,457	500	809	0.527	0.708
4	7	3	36	0.152	0.217
5	1,047	458	484	0.526	0.757
6	322	178	325	0.390	0.606
7	302	220	306	0.365	0.630
8	1	2	7	0.100	0.300
9	134	26	68	0.588	0.702

Table 10: Per-category validation-split recall, derived from the counts in Figure 5. **Exact** recall treats only strict matches as correct. **Exact+near** counts any predicted span that carries the correct category, regardless of offset. Gold spans per category are the row sums of the *Counts* block.

AlienAnnotators at PsyDefDetect: What Lies Between the Lines: Probing Lightweight Open-Source LLMs for Psychological Defense Mechanism Detection

Siam Rahman Karip and Nahid Hossain
United International University
Dhaka, Bangladesh

Abstract

Detecting psychological defense mechanisms in therapy dialogue is a clinically valuable but computationally underexplored task. We present our systematic analysis for PsyDefDetect, a shared task at BioNLP@ACL 2026, which frames defense detection as a nine-class utterance-level classification problem based on the Defense Mechanism Rating Scale (DMRS). We systematically evaluate six open-source, instruction-tuned small language models (SLMs, ≤ 9 B parameters) in zero-shot and fine-tuning settings, and compare a clinically-grounded prompt against the organizer-provided baseline. Our official submission achieved 59.96% accuracy and 16.28% Macro F1. Post-submission experiments show that fine-tuning combined with 5-fold cross-validation and logit averaging ensemble substantially improves performance, with the best configuration reaching 34.59% Macro F1 and 65.25% accuracy. We find that clinically-grounded prompts outperform bare label definitions, model scale does not consistently improve zero-shot performance, and fine-tuning dramatically recovers even collapsed zero-shot models. Certain defense tiers remain persistently difficult across all settings, pointing to clinical ambiguity at tier boundaries as a more fundamental bottleneck than data imbalance alone.

1 Introduction

Natural language processing is increasingly applied to understand different characteristics from client language in clinical interactions (Voultsiou and Moussiades, 2026; Na et al., 2025). However, many clinically meaningful constructs remain difficult to model computationally, such as, psychological defense mechanisms. Defense mechanisms are unconscious strategies that individuals use to manage internal conflict and distress. Accurate identification of defenses can improve case formulation

and therapeutic outcomes. The Defense Mechanism Rating Scale (DMRS) provides a structured framework for this task. It organizes defenses into hierarchical levels based on adaptiveness (Perry and Henry, 2004).

The PsyDefDetect shared task (Na et al., 2026a) operationalizes this problem as a nine-class classification task, based on the DMRS tiers. The task is challenging by design: the dataset is small, the label distribution is severely imbalanced (up to $34.6\times$ between majority and minority classes), and the tier boundaries require clinical judgment to distinguish. These challenges motivate a systematic study of how small-scale LLMs handle this task.

Our main contributions are as follows:

- A systematic zero-shot evaluation of six small-scale LLMs comparing a clinically-grounded prompt against the organizer-provided baseline, revealing that behavioral tier descriptions consistently outperform bare label definitions, and that larger model size does not consistently improve fine-grained clinical classification.
- A fine-tuning pipeline combining clinically-grounded DMRS prompt design, dialogue-grouped cross-validation, and ensemble strategies for robust low-resource clinical classification.
- A brief error analysis examining per-tier classification difficulty, model confusion patterns across clinically similar defense levels, and the effect of class imbalance on minority tier detection.

2 Task and Data

The task uses PSYDEFCONV (Na et al., 2026b), a dataset of emotional support dialogues annotated with DMRS defense levels. Each instance consists of a multi-turn dialogue and a target utterance produced by the help-seeker, which is assigned

one of eight DMRS defense levels, or flagged as requiring more information (L8: Needs More Information) when the dialogue context is insufficient for a confident classification. The official training set contains 1,864 samples, and the official test set contains 472 samples. The dataset exhibits severe class imbalance, with the "Highly Adaptive" tier (L7) comprising 51.9% of training samples, while the rarest class, "Needs More Information" (L8), accounts for only 1.5%, depicting an overall imbalance ratio of $34.6\times$. Table 1 presents the full label distribution.

ID	Label	Count	%
L0	No Defense	296	15.9
L1	Action	108	5.8
L2	Maj. Image-Distort	61	3.3
L3	Disavowal	99	5.3
L4	Min. Image-Distort	84	4.5
L5	Neurotic	48	2.6
L6	Obsessional	172	9.2
L7	High-Adaptive	968	51.9
L8	Needs More Info	28	1.5
Total		1,864	100.0

Table 1: Training set label distribution. Imbalance ratio is computed relative to the majority class (L7).

3 Methodology

3.1 Model Selection

We evaluate six open source, instruction tuned LLMs with at most 9B parameters, selected to cover a range of model families, scales, and pretraining objectives. From the Gemma family, we include Gemma3-1B-it and Gemma2-9B-it (Team et al., 2024). From the Llama family, we include Llama-3.1-8B-Instruct and Llama-3.2-1B-Instruct (Grattafiori et al., 2024). From the Qwen3 family, we include Qwen3-1.7B and Qwen3-8B (Yang et al., 2025). We primarily focus on general-purpose decoder-only models, as their instruction-following capability enables zero-shot evaluation without task-specific adaptation, which is central to our prompt design comparison.

3.2 Prompt Design

We compare two prompt variants. The **organizer-provided baseline prompt** (Variant A) presents the task instruction alongside bare label names and their constituent defense mechanisms. Our **clinically-grounded prompt** (Variant B) replaces these with behavioral descriptions derived from the DMRS manual (Perry and Henry, 2004), providing

the model with observable verbal cues for each tier. For example, the Disavowal tier is described as: *the speaker denies an obvious reality, externalizes blame, justifies behavior with plausible-sounding logic, or retreats into elaborate private fantasy*. Full prompt texts are provided in Appendix 5.

3.3 Zero-Shot Evaluation

We evaluate all six models in a zero-shot setting using both prompt variants. Each model receives the full dialogue context and target utterance, and is required to output a single digit (0–8) with no examples or additional guidance. We compare Variant A and Variant B across all models to assess the effect of clinically-grounded prompt design on zero-shot classification performance.

3.4 Finetuning

We fine-tune three models: Qwen3-1.7B, Gemma3-1B-IT and Llama-3.2-1B-Instruct. All models are trained using standard causal language modeling loss with the clinically-grounded prompt (Variant B) as the input format. To prevent data leakage across dialogue turns, we apply 5-fold stratified cross-validation grouped by dialogue ID, ensuring that all utterances from the same dialogue appear in the same fold. All models are trained with a learning rate of $5e-5$ with cosine decay to $5e-6$, a warmup ratio of 0.1, and a batch size of 4 (effective 16 with gradient accumulation). Training runs for up to 8 epochs with early stopping patience of 2 epochs based on validation Macro F1, and a weight decay of 0.01.

3.5 Ensemble Strategies

Given the 5-fold cross-validation setup, we explore three ensemble strategies over the fold checkpoints:

- **Logit Averaging:** Raw output logits are averaged across all five folds before taking the argmax.
- **Majority Vote:** The most frequent predicted label across folds is selected, with ties broken by logit confidence.
- **Best Single Fold:** The highest-performing individual fold checkpoint is used alone, serving as a non-ensemble baseline.

3.6 Evaluation Metrics

We report three metrics for all experiments. **Macro F1** is our primary metric, as it weights all classes equally regardless of support, directly capturing

Model	Variant A (Baseline)			Variant B (Clinical)		
	Acc	F1	MAE	Acc	F1	MAE
Qwen3-1.7B	25.85	12.17	2.32	34.32	14.57	2.77
Qwen3-8B	22.67	11.16	3.97	29.24	13.63	3.23
Gemma3-1B-it	18.97	8.01	3.45	21.17	12.10	3.01
Gemma2-9B-it	28.97	15.01	4.29	31.14	18.79	3.10
Llama-3.2-1B-Instruct	10.41	7.14	4.21	9.75	6.93	3.46
Llama-3.1-8B-Instruct	15.77	10.79	2.47	16.74	12.41	2.23

Table 2: Zero-shot results under Variant A (organizer-provided baseline) and Variant B (clinically-grounded) prompts on the official test set. Accuracy and Macro F1 (F1) are given in percentage(%) values. Best Macro F1 per model is **bolded**.

performance on minority tiers. **Accuracy** measures overall correctness but is susceptible to majority-class bias given the severe class imbalance. **Mean Absolute Error (MAE)** treats the DMRS tiers as an ordinal scale and penalizes predictions proportionally to their distance from the true tier, which is clinically meaningful given the hierarchical structure of the DMRS.

4 Results and Analysis

4.1 Zero-Shot Results

Table 2 presents the zero-shot performance of all six models under both prompt variants. Variant B (clinically-grounded) consistently outperforms Variant A (baseline) across most models in terms of Macro F1, demonstrating the benefit of behavioral descriptions over bare label names for this task.

4.2 Fine-Tuning and Ensemble Results

Table 3 presents the test set performance of the three fine-tuned models under three ensemble strategies and our leaderboard submission. Our official leaderboard submission was based on a fine-tuned Qwen3-1.7B model, achieving 59.96% accuracy and 16.28% Macro F1. Ensembling consistently improves over the best single fold across all models, with logit averaging and majority vote yielding comparable results. The results reported here reflect post-submission experiments with improved methodology. The official submission used the organizer-provided baseline prompt (Variant A) and was fine-tuned on a 75/15/15 train/validation/test split of the training set, as the official test set was not available during the submission period. Post-submission, we adopted the clinically-grounded prompt (Variant B), applied 5-fold cross-validation grouped by dialogue ID, and evaluated on the released test set, which accounts for the substantial improvement from 16.28% to 34.59% Macro F1.

Model	Strategy	Acc	F1	MAE
Qwen3-1.7B ¹	Logit Avg	65.25	34.59	0.94
	Maj. Vote	63.35	31.29	1.15
	Best Fold	58.90	33.47	1.28
Llama-3.2-1B ¹	Logit Avg	65.25	34.21	1.16
	Maj. Vote	63.14	33.74	1.25
	Best Fold	59.98	32.63	1.33
Gemma3-1B-it ¹	Logit Avg	59.75	27.85	1.34
	Maj. Vote	56.89	26.51	1.38
	Best Fold	59.30	29.46	1.35
Qwen3-1.7B ²	Official Sub.	59.96	16.28	2.45

Table 3: MAE = Mean Absolute Error. Accuracy and Macro F1 (F1) are given in percentage(%) values.¹Post-submission Fine-tuning and ensemble experiment results on the official test set. ²Official leaderboard submission.

4.3 Analysis

Effect of Model Scale in Zero-Shot setting:

Larger models do not consistently outperform smaller ones in zero-shot. Within the Qwen3 family, the 1.7B model (F1=14.57%) outperforms the 8B model (F1=13.63%). In the Gemma and Llama families, larger models do improve over their smaller counterparts, but only modestly (+6.69pp and +5.48pp respectively). This reflects the highly specialized nature of the task: DMRS tier classification requires expert clinical reasoning that is not well-represented in general pretraining data, limiting the benefit of additional parameters.

Per-Tier Difficulty and Class Learnability:

Across all models and settings, L2 (Major Image-Distorting), L5 (Neurotic), and L6 (Obsessional) remain the hardest tiers, with near-zero F1 in zero-shot and modest improvement after fine-tuning, as shown in Figure 1. These tiers are both low-resource and clinically subtle, their boundaries require distinguishing between cognitively similar defensive patterns (e.g., intellectualization vs. ratio-

nalization), which current small models struggle to capture reliably. L0 (No Defense), by contrast, is consistently well-learned across all fine-tuned models with F1 ranging from 77% to 91%, despite not being the majority class. L7 (High-Adaptive), despite comprising 51.9% of training samples, shows more variable F1 (74%–81%) and is frequently confused with L0, as evidenced by consistent off-diagonal mass in the L7 row of the confusion matrices (Appendix, Figure 17). L8 (Needs More Info) shows partial recovery under fine-tuning, reaching up to 33% F1 with logit averaging, likely because its defining characteristic (insufficient context) is more lexically identifiable than subtle tier distinctions. Figure 17 further illustrates the per-tier improvement from zero-shot to fine-tuning. These patterns suggest that class frequency alone does not determine learnability, and that clinical ambiguity at tier boundaries is a more significant bottleneck than data imbalance.

Ordinal Proximity of Predictions: Despite modest Macro F1 scores, models show meaningful ordinal awareness after fine-tuning. Qwen3-1.7B with logit averaging achieves an MAE of 0.9428, the only configuration to fall below 1.0 (Table 3). This indicates predictions are on average less than one DMRS tier away from the true label. Llama-3.2-1B with logit averaging follows closely with MAE = 1.1695. In contrast, zero-shot models show substantially higher MAE, with Llama-3.2-1B-Instruct reaching 3.4597, indicating near-random tier assignment. This gap confirms that fine-tuning not only improves exact classification but also brings predictions significantly closer to the correct tier on the ordinal scale.

Fine-Tuning Rescues Collapsed Zero-Shot Models: Llama-3.2-1B-Instruct nearly collapses in zero-shot evaluation, achieving only 6.93% Macro F1, effectively predicting low-tier labels for almost all inputs. Similarly, Gemma3-1B-it achieves only 12.10% Macro F1 in zero-shot, with L2, L4, L5, and L8 all at zero F1. Despite these failures, fine-tuning brings Llama-3.2-1B to 32.63% Macro F1 (second best overall) and Gemma3-1B-it to 29.46%, representing improvements of $4.7\times$ and $2.43\times$ respectively. Qwen3-1.7B, which already performs best in zero-shot among the fine-tuned models (F1=14.57%), also benefits the most in absolute terms, reaching 33.47% after fine-tuning. These results demonstrate that zero-shot performance is a poor predictor of fine-tuning potential, and that

even severely collapsed models can be effectively adapted with task-specific training. Tables 2 and 3 show the detailed results.

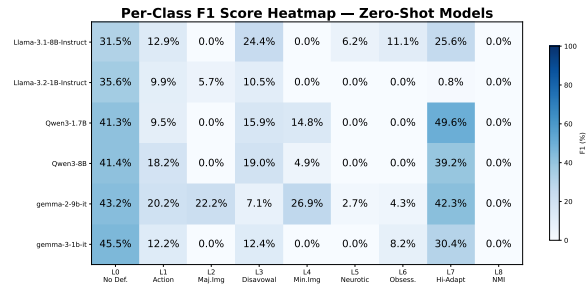


Figure 1: Per-class F1 heatmap for all six zero-shot models. L2 (Major Image-Distorting), L5 (Neurotic), and L6 (Obsessional) show near-zero F1 across all models, reflecting both low support and high clinical ambiguity.

Ensemble Effect: Logit averaging consistently outperforms majority vote and best single fold across fine-tuned models. For Qwen3-1.7B, logit averaging achieves F1=34.59% vs. 32.29% for majority vote and 33.47% for best single fold, confirming that aggregating probability distributions across folds is more effective than hard-label voting. Llama-3.2-1B also showed competitive performance by reaching 34.21% Macro F1, with similar results for Majority Voting. On the other hand, Gemma performed very inconsistently with its best fold performance (29.46% F1) being better than ensembling. Table 3 shows detailed results.

5 Conclusion

We presented a systematic study of small language models for psychological defense mechanism classification under the DMRS framework. Across six models in zero-shot and fine-tuning settings, we find that task-specific fine-tuning combined with logit averaging ensemble is the dominant factor in performance, reaching 34.59% Macro F1 (Qwen3-1.7B), nearly doubling the best zero-shot result of 18.79% (Gemma2-9B-IT). Model scale does not consistently improve zero-shot performance, suggesting that the task’s clinical specificity limits the benefit of additional parameters. Certain defense tiers, particularly L2, L5, and L6, remain persistently difficult across all settings, pointing to clinical ambiguity at tier boundaries as a fundamental challenge beyond data imbalance.

Limitations

Due to GPU constraints, fine-tuning is restricted to models at or below 1.7B parameters. Ensemble strategies are limited to fold-level aggregation within each model, and all evaluations use a single dataset, leaving generalization untested. A broader exploration of prompt strategies, including few-shot and chain-of-thought prompting, remains for future work.

References

- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.
- J. Christopher Perry and Melissa Henry. 2004. [Chapter 9 - studying defense mechanisms in psychotherapy using the defense mechanism rating scales](#). In Uwe Hentschel, Gudmund Smith, Juris G. Draguns, and Wolfram Ehlers, editors, *Defense Mechanisms*, volume 136 of *Advances in Psychology*, pages 165–192. North-Holland.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Evdokia Voultsiou and Lefteris Moussiades. 2026. [A systematic review of large language models in mental health: Opportunities, challenges, and future directions](#). *Electronics*, 15(3).
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Appendix

Variant A: Organizer-Provided Baseline Prompt

You are a Defense Mechanism Rating Scale (DMRS) specialist. Examine the dialogue carefully and select the single most appropriate defense tier. When multiple defenses seem plausible, choose the tier with the strongest supporting evidence; if evidence is weak or contradictory, default to '0' (No defense).

Dialogue context: {conversation}
Target utterance: {current_text}

Labels:

- 0 = No defense
- 1 = Action Defense Level (Acting Out / Help-Rejecting Complaining / Passive Aggression)
- 2 = Major Image-distorting Defense Level (Splitting / Projective Identification)
- 3 = Disavowal Defense Level (Denial / Projection / Rationalization / Autistic Fantasy)
- 4 = Minor Image-distorting Defense Level (Devaluation / Idealization / Omnipotence)
- 5 = Neurotic Defense Level (Displacement / Dissociation / Reaction Formation / Repression)
- 6 = Obsessional Defense Level (Intellectualization / Isolation of Affects / Undoing)
- 7 = Highly Adaptive Defense Level (Affiliation / Altruism / Anticipation / Humor / Self-Assertion / Self-Observation / Sublimation / Suppression)
- 8 = Need More Information

Return the label digit (0-8) ONLY. No additional content is allowed.

Variant B: Clinically-Grounded Prompt

You are a clinician trained in the Defense Mechanism Rating Scale (DMRS; Perry, 1990). Analyze the dialogue and identify the defense mechanism in the target utterance based strictly on observable verbal behavior.

Dialogue context: {conversation}
Target utterance: {current_text}

- 0 = No defense: The speaker communicates directly with no defensive distortion.
- 1 = Action: Acts on impulse, expresses hostility indirectly, or repeatedly seeks yet rejects help.
- 2 = Major Image-distorting: Rigidly splits others into all-good/all-bad, or projects own feelings onto others.
- 3 = Disavowal: Denies reality, externalizes blame, justifies behavior with plausible logic, or retreats into fantasy.
- 4 = Minor Image-distorting: Subtly belittles self/others, idealizes unrealistically, or expresses special invulnerability.
- 5 = Neurotic: Redirects emotion onto safer target, shows emotional blankness, expresses opposite of what is felt.
- 6 = Obsessional: Detaches via abstract reasoning, focuses on trivial details to avoid affect, or attempts to undo thoughts.
- 7 = Highly Adaptive: Seeks support, acts for others, plans ahead, uses humor constructively, asserts needs calmly.
- 8 = Need More Information: Defense is suspected but insufficient evidence to confirm any tier.

Return the digit (0-8) ONLY. No additional content is allowed.

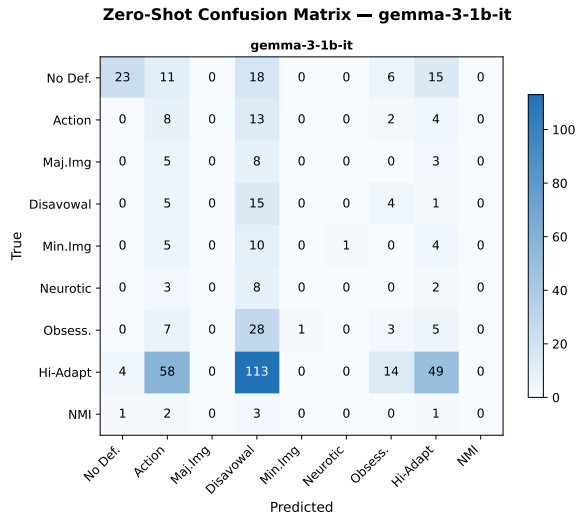


Figure 2: Confusion matrix for Gemma3-1B-IT (Zero-shot).

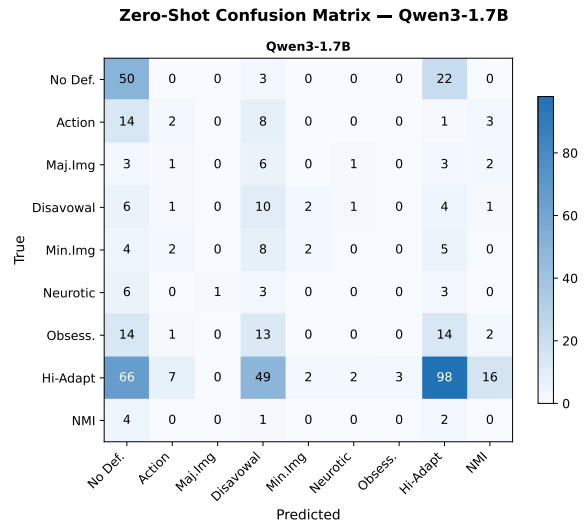


Figure 4: Confusion matrix for Qwen-3-1.7B (Zero-shot).

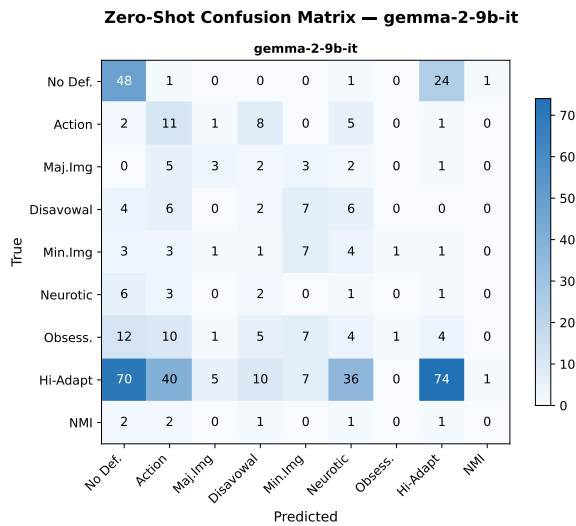


Figure 3: Confusion matrix for Gemma2-9B-IT (Zero-shot).

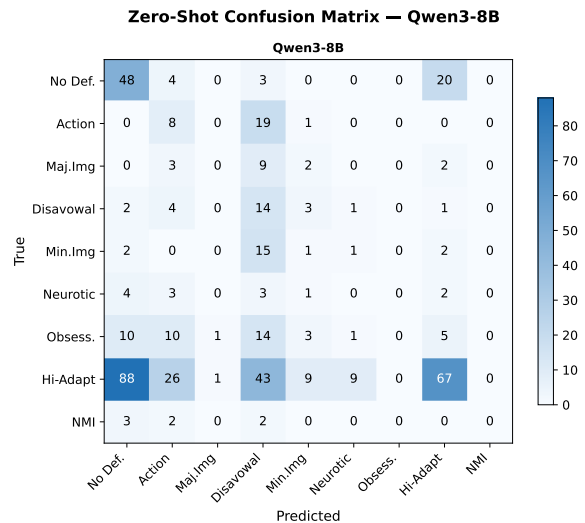


Figure 5: Confusion matrix for Qwen-3-8B (Zero-shot).

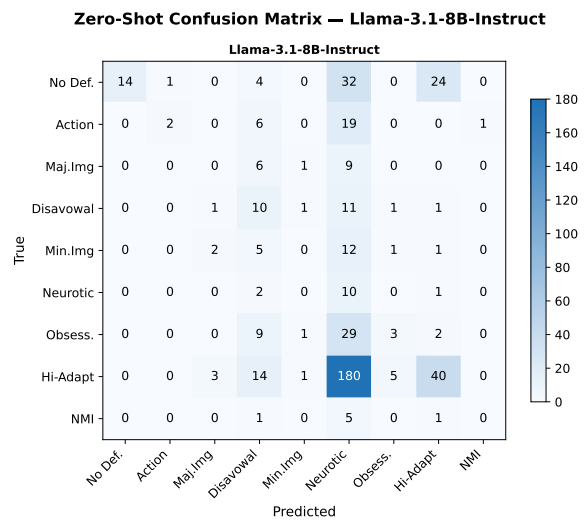


Figure 6: Confusion matrix for Llama-3.1-8B-Instruct (Zero-shot).

Zero-Shot Confusion Matrix — Llama-3.2-1B-Instruct

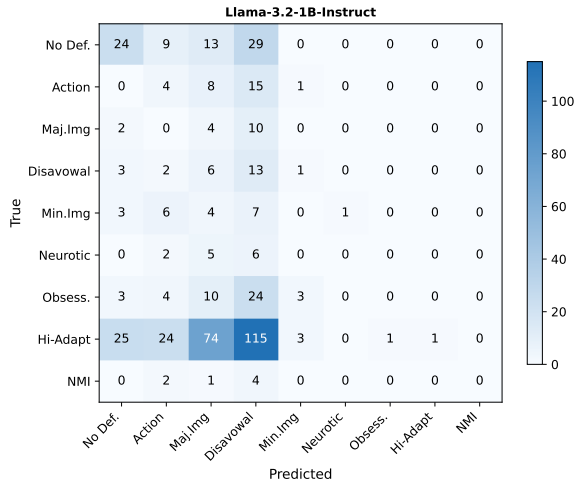


Figure 7: Confusion matrix for Llama-3.2-1B-Instruct (Zeroshot)

Finetuned: Llama-3.2-1B-Instruct — Best Fold (2)

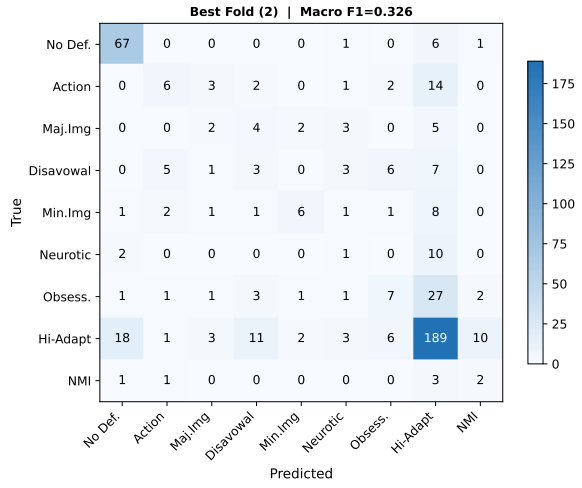


Figure 10: Confusion matrix for Llama-3.2-1B-Instruct (Best Fold).

Finetuned: Qwen3-1.7B — Best Fold (2)

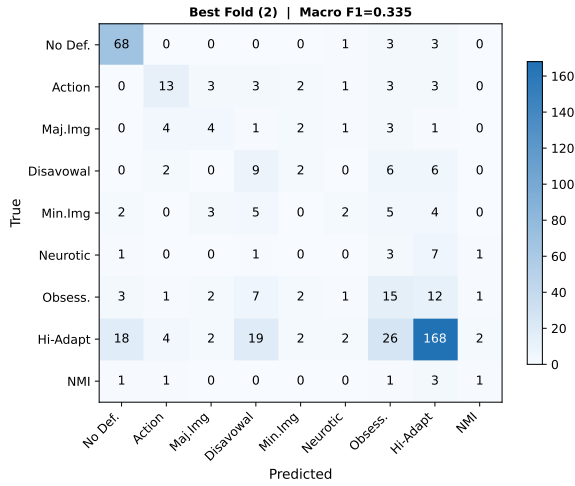


Figure 8: Confusion matrix for Qwen3-1.7B (Best Fold).

Finetuned: Qwen3-1.7B — Majority Vote

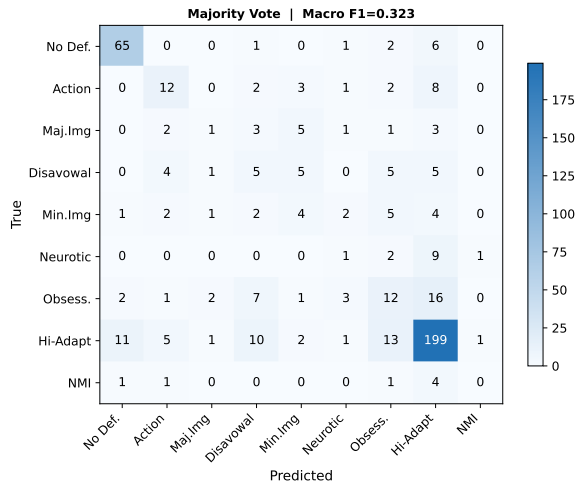


Figure 11: Confusion matrix for Qwen3-1.7B (Ensembling - Majority Voting).

Finetuned: Gemma-3-1B-it — Best Fold (1)

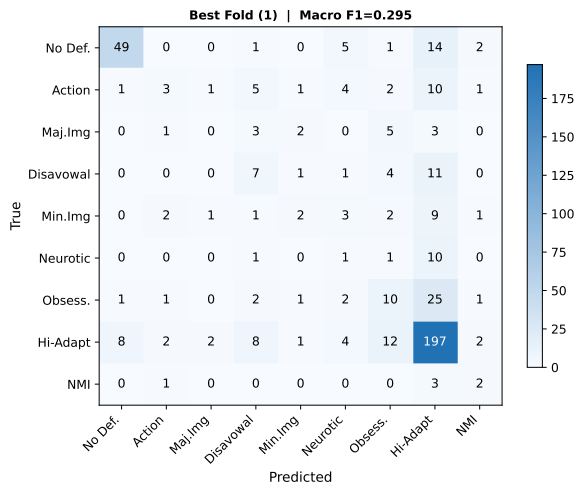


Figure 9: Confusion matrix Gemma-3-1B-it (Best Fold).

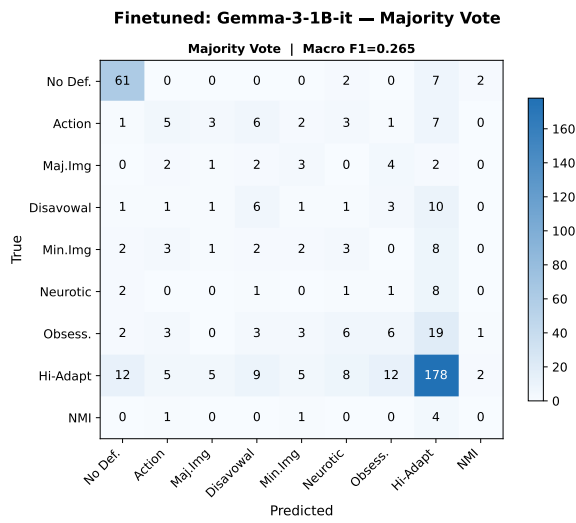


Figure 12: Confusion matrix for Gemma-3-1B-it (Ensembling - Majority Voting).

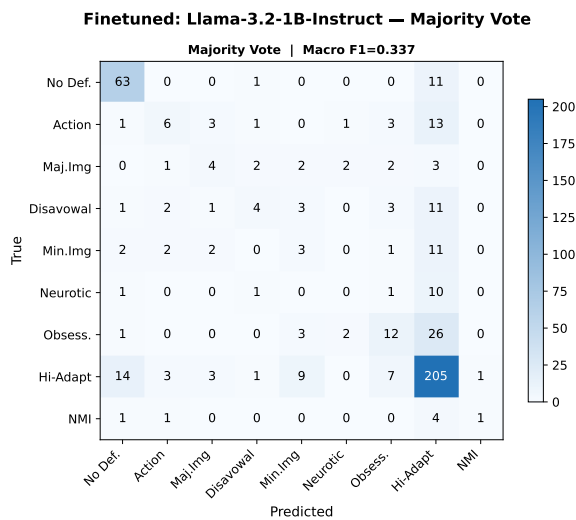


Figure 13: Confusion matrix for Llama-3.2-1B-Instruct (Ensembling - Majority Voting).

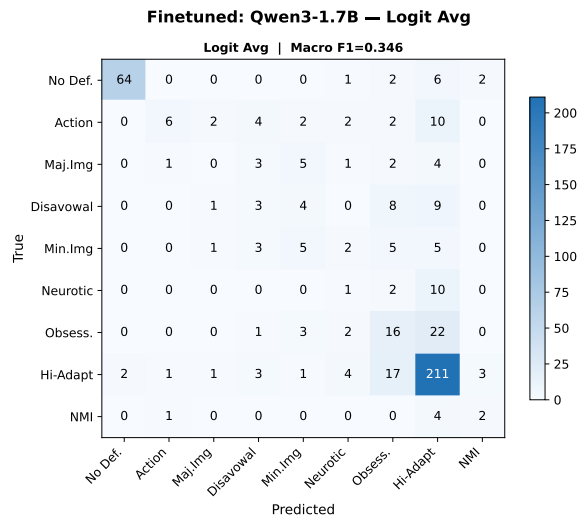


Figure 14: Confusion matrix for Qwen3-1.7B (Ensembling - Logit Averaging).

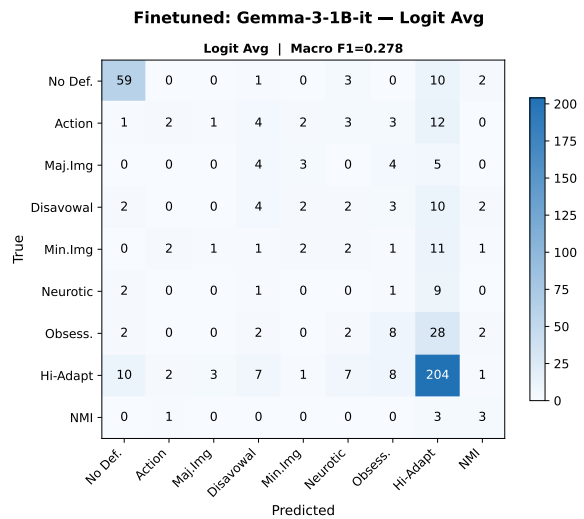


Figure 15: Confusion matrix for Gemma-3-1B-it (Ensembling - Logit Averaging).

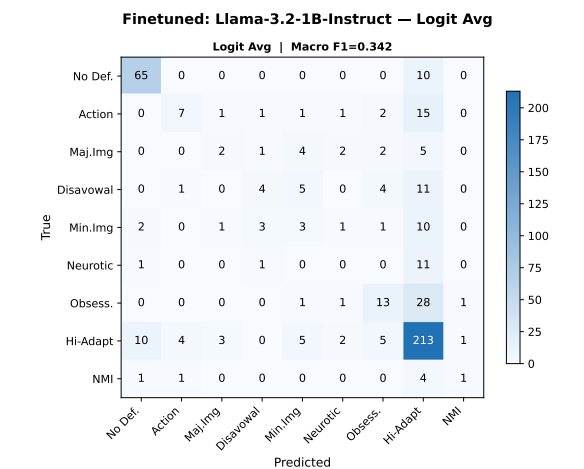


Figure 16: Confusion matrix for Llama-3.2-1B-Instruct (Ensembling - Logit Averaging).

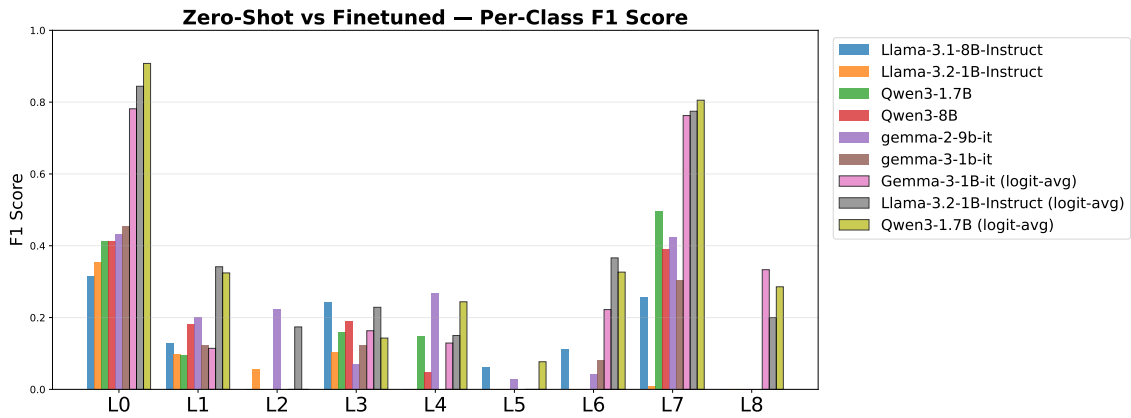


Figure 17: Per-class F1 comparison between the best zero-shot model (Gemma2-9B-IT) and the best fine-tuned model (Qwen3-1.7B, logit averaging). Fine-tuning improves most tiers, though L2, L5, and L6 remain difficult.

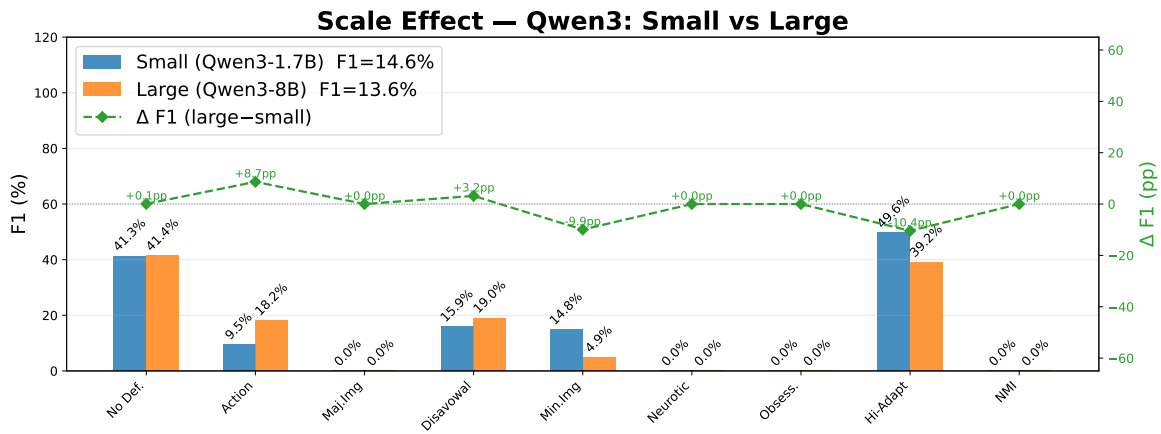


Figure 18: Per-class F1 comparison between Qwen3-1.7B and Qwen3-8B in zero-shot. The larger model underperforms on several tiers.

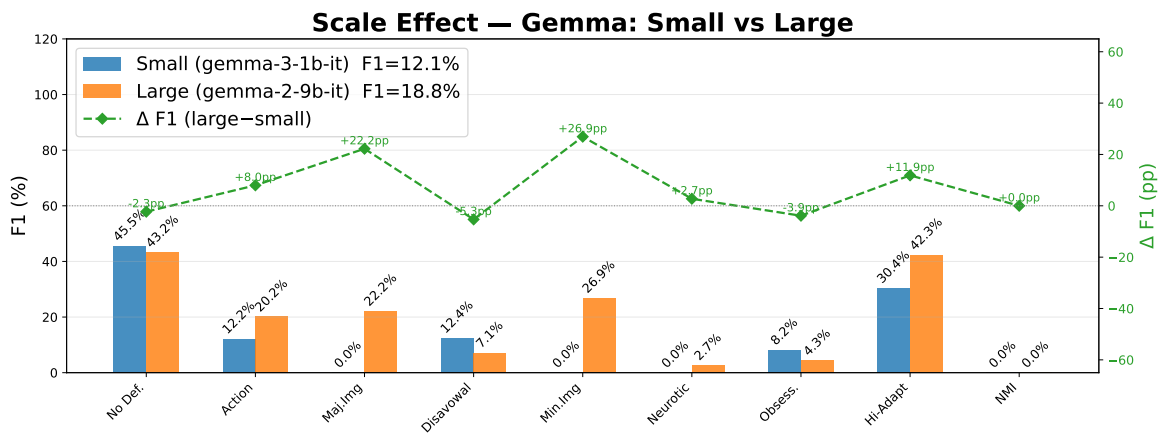


Figure 19: Per-class F1 comparison between Gemma3-1B-IT and Gemma2-9B-IT in zero-shot.

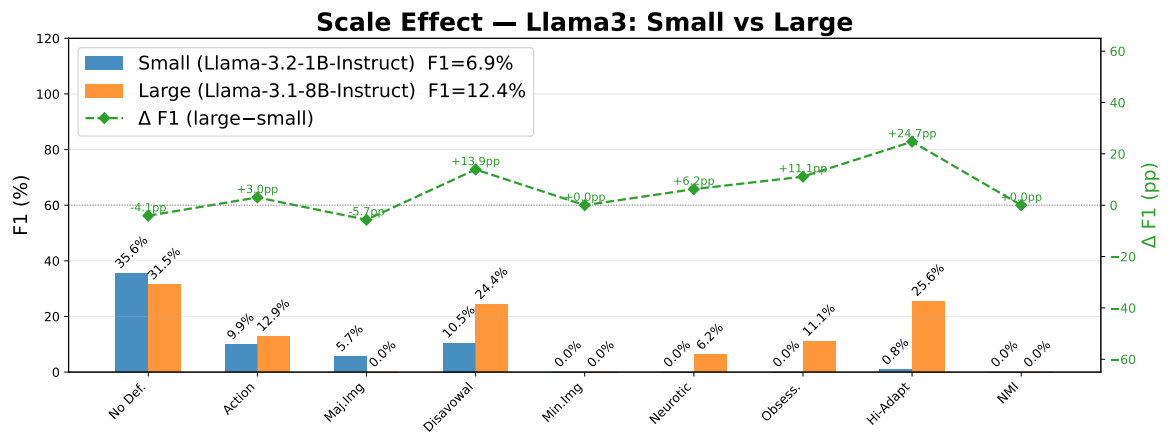


Figure 20: Per-class F1 comparison between Llama-3.2-1B-Instruct and Llama-3.1-8B-Instruct in zero-shot.

Team Aurum at MedExACT 2026@ACL: Data Augmentation and Clinical Longformer Fine-Tuning for Medical Decision Extraction

Vinay Babu Ulli
Oogwai Analytics,
Bangalore, India
ullivinaybabu@gmail.com

Jyoti Kumari
Department of Linguistics,
Banaras Hindu University, India
jyoti@bhu.ac.in

Anindita Mondal
Language Technologies
Research Center,
IIT Hyderabad, India
anindita.mondal@research.iit.ac.in

Abstract

This paper describes the system submitted by team **Aurum** to the Medical Decision Extraction, Analysis, and Classification Task (MedExACT) at BioNLP 2026. The task requires the extraction and classification of contiguous text spans representing medical decisions from lengthy ICU discharge summaries. To address the dual challenges of long document lengths and severe class imbalance within a limited training set of 350 notes, we propose a two-pronged strategy. First, we employ a tripartite data augmentation pipeline utilizing rule-based entity replacement, LLM-based contextual paraphrasing, and synthetic note generation to expand the training data to over 2,300 notes. Second, we fine-tune a domain-specific Clinical Longformer model equipped with a sliding-window inference mechanism and Focal Loss to handle sequences up to 2,048 tokens while focusing on rare decision categories. Paired with a targeted post-processing module, our system achieved a Final Score of 0.5251, demonstrating high token-level detection (Token F1: 0.6311) and strong stability across patient demographics.

1 Introduction

Clinical discharge summaries document critical medical decisions, ranging from diagnostic assessments to therapeutic procedures, that shape downstream patient care. However, these decisions are embedded in dense, unstructured, and often lengthy free-text narratives (Agrawal et al., 2022). The MedExACT 2026 shared task (Elgaar et al., 2026) aims to advance the automatic extraction of these decisions by framing it as a joint span detection and classification problem. Crucially, systems are evaluated not only on average performance but also on their worst-group robustness across demographic lines.

Extracting medical decisions from the MedDec dataset (Elgaar et al., 2024) presents two primary

bottlenecks. First, the training set is relatively small (350 notes), leading to rapid model overfitting. Second, discharge summaries average between 2,000 and 6,000 tokens in length. Standard Transformer encoders limited to 512 tokens (e.g., ClinicalBERT, ELECTRA) inevitably truncate 60–80% of the text, discarding vital sections like the *Hospital Course*.

To overcome these limitations, we engineered a pipeline that combines extensive **Data Augmentation** with **Clinical Longformer Fine-Tuning**. Our augmentation strategies expand linguistic and entity diversity, while the Longformer architecture (Beltagy et al., 2020) ensures that decisions located deep within the document are successfully processed.

2 Task Description and Background

The MedExACT 2026 task is built upon the MedDec dataset (Elgaar et al., 2024), derived from the MIMIC-III critical care corpus (Johnson et al., 2016). The utility of this dataset extends beyond static benchmarks, having recently supported the development of MedDecXtract, an interactive downstream tool designed to assist clinicians in extracting, visualizing, and annotating medical decisions in real-time (Elgaar et al., 2025).

Task Definition: Given a full discharge summary, systems must detect contiguous text spans that express medical decisions and assign each span one of nine DICTUM (Ofstad et al., 2016) decision categories (*Contact related, Gathering information, Defining problem, Treatment goal, Drug, Therapeutic procedure, Evaluating test result, Deferment, Advice and precaution*) or a *None* label when no decision is present.

Note: Some MedDec annotations may include Category 10 (Legal/insurance related) and Category 11 (Others). These categories are out of scope for MedExACT@ACL 2026 and are ignored by the official evaluator.

Decision Type (lr)2-3 (lr)4-9 (lr)10-11	Sex		Race					Lng. Proficiency		
	Male (n=259)	Female (n=192)	White (n=327)	AA (n=42)	Hispanic (n=25)	Asian (n=15)	NH (n=1)	Other (n=21)	En (n=260)	Non-En (n=45)
Defining Problem	39.2	38.8	39.5	37.5	38.0	36.4	30.9	38.6	38.7	39.2
Drug	26.0	25.1	25.7	24.4	25.0	27.5	19.1	27.0	26.1	25.6
Evaluation	12.9	13.6	12.6	16.6	13.3	12.7	25.5	12.8	13.1	13.9
Therapeutic proc.	12.2	12.4	12.4	12.5	11.7	13.2	10.6	12.2	12.0	12.0
Contact	4.9	5.2	5.0	4.6	6.0	5.4	8.5	4.3	4.8	5.1
Advice	3.4	3.5	3.5	3.2	4.2	3.3	0.0	3.9	3.9	3.0
Gathering info	0.8	0.9	0.8	0.7	1.2	1.3	5.3	0.9	0.9	0.6
Treatment goal	0.3	0.3	0.3	0.3	0.4	0.2	0.0	0.2	0.2	0.4
Deferment	0.2	0.2	0.2	0.2	0.2	0.0	0.0	0.1	0.2	0.2
Legal/Insurance	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total Count	33,054	24,235	41,666	5,684	3,264	1,737	94	3,078	37,026	6,295

Table 1: Percentage of annotated spans for each decision category across protected variables in the MedDec dataset. n is the number of discharge summaries for each category and the last row shows the total count of decisions per variable.

Dataset and Demographics: The dataset exhibits extreme class imbalance. As shown in Table 1, *Defining Problem* and *Drug* decisions dominate the annotations, accounting for roughly 65% of all spans. Conversely, categories such as *Treatment Goal*, *Gathering info*, and *Deferment* each represent less than 1% of the data. Table 1 details the distribution of these spans across protected variables, which is fundamental to the task’s worst-group robustness evaluation.

3 Methodology

We formulate the medical decision extraction task as a token-level sequence tagging problem. We apply a standard BIO (Begin, Inside, Outside) labeling scheme to the 9 in-scope DICTUM categories, resulting in 19 distinct prediction classes (9 categories \times 2 + 1 ‘O’-tag). Our methodology comprises a multi-strategy data augmentation pipeline, a domain-adapted Clinical Longformer backbone, specialized loss functions, and targeted post-processing.

3.1 Tripartite Data Augmentation Pipeline

With only 350 training notes and models containing over 140 million parameters, overfitting and memorization are significant risks. To prevent the model from simply memorizing the training set, we designed a three-pronged augmentation strategy to inject entity, linguistic, and volume diversity, thereby expanding the training corpus to over 2,300 notes.

Strategy 1: Rule-Based Entity Replacement.

To increase entity diversity, we utilized curated medical dictionaries to replace specific drugs, con-

ditions, and therapeutic procedures with clinically similar alternatives (e.g., swapping *vancomycin* with *meropenem*, or *pneumonia* with *bronchitis*). Replacements were executed with a 30% probability per entity. Crucially, because string lengths vary, our algorithm dynamically recalculates and shifts the exact character offsets for all subsequent annotations in the document. This zero-cost method generated roughly 1,050 augmented notes.

Strategy 2: LLM Contextual Paraphrasing.

To increase linguistic diversity and make the model robust to different physician writing styles, we leveraged the Claude Sonnet Large Language Model (LLM) via the OpenRouter API. We parsed the training documents into decision spans and the non-annotated “gaps” between them. We prompted the LLM to rewrite only the gap texts while keeping the annotated spans strictly verbatim. This forces the model to learn the semantic boundaries of decisions regardless of surrounding syntax (e.g., altering “*was diagnosed with [SPAN]*” to “*workup was consistent with [SPAN]*”). This process yielded approximately 700 paraphrased notes.

Strategy 3: Synthetic Note Generation.

Categories such as *Treatment Goal* (Cat 4), *Gathering Info* (Cat 2), and *Deferment* (Cat 8) are severely starved, appearing in less than 1% of annotations. We prompted the LLM to generate entirely synthetic discharge summaries targeting these rare categories using the following prompt template:

“Generate a realistic hospital discharge summary for a patient with [Disease]. Include 40-80 medical decisions across categories 1-9. Return the note text AND a JSON list of annotations with exact character offsets.”

For the sampling strategy, we utilized a temperature of 0.7 and Top-P of 0.9 to ensure varied but clinically plausible outputs. Because LLM-generated character offsets are notoriously inaccurate (often hallucinating string lengths), we implemented strict programmatic filtering: the system searched the generated text for the exact predicted string. Annotations that could not be perfectly substring-aligned were discarded. This targeted generation successfully increased the representation of starved categories (Cat 2, 4, 8) from under 1% to approximately 6% of the total training spans, yielding 200 high-quality synthetic notes.

Augmentation Quality and Validation. To validate the quality of our augmented data, we manually inspected a random sample of 50 notes from each strategy. Entity replacement maintained near 100% clinical plausibility as it relied strictly on curated medical dictionaries (e.g., matching drug classes). LLM paraphrasing preserved the exact decision spans, introducing minor syntactic artifacts in only $\sim 4\%$ of the surrounding gaps. For the synthetic generation, manual inspection revealed an initial LLM offset error rate of roughly 20–25%; however, our strict substring-matching filter effectively eliminated these misalignments, dropping the final annotation error rate in the synthetic set to under 2% while maintaining coherent clinical narratives.

3.2 Clinical Longformer Architecture

Discharge summaries in the MedDec dataset average between 2,000 and 6,000 tokens. Standard Transformer encoders (e.g., BERT, ELECTRA) impose a 512-token limit, which effectively truncates 60% to 80% of a standard discharge summary, completely discarding critical sections such as the *Hospital Course* where the majority of medical decisions reside.

To resolve this, we utilized the Longformer architecture (Beltagy et al., 2020), which replaces $\mathcal{O}(n^2)$ full self-attention with a sparse attention pattern combining a sliding local window (512 tokens on each side) and task-specific global attention on the [CLS] token. This reduces complexity to $\mathcal{O}(n \times w)$ and extends the maximum sequence length to 4,096 tokens.

Specifically, we selected yikuan8/Clinical-Longformer, which was further pre-trained on MIMIC-III clinical notes. This domain-specific pre-training ensures the

model tokenizer is adapted to medical vocabulary and inherently understands discharge summary structures (e.g., section headers, de-identification markers).

Layer Freezing and Custom Classification Head.

To prevent catastrophic forgetting of pre-trained domain knowledge and to stabilize training on our small dataset, we froze the bottom 6 layers of the 12-layer encoder. This reduced the trainable parameter count from 149M to approximately 75M. The output of the trainable top 6 layers was fed into a custom 2-layer classification head: a linear projection from 768 to 256 dimensions, followed by a GELU activation, a 0.2 dropout layer, and a final projection to the 19 BIO classes.

3.3 Addressing Class Imbalance with Focal Loss

The class distribution in the MedDec dataset is exceptionally skewed, with the ‘O’ (Outside) tag comprising 95.2% of all tokens in the training data. Under standard Cross-Entropy loss, the model easily achieves high overall accuracy by defaulting to ‘O’ predictions, thereby ignoring the sparse medical decision boundaries.

To penalize the model for relying on easy background tokens, we replaced Cross-Entropy with Focal Loss (Lin et al., 2017):

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (1)$$

We set the focusing parameter $\gamma = 2.0$ to aggressively down-weight easily classified examples. Furthermore, we applied a class-specific weighting factor $\alpha_O = 0.3$ strictly to the ‘O’ tag, ensuring that the loss gradients are dominated by the hard, rare decision boundaries (Categories 1–9) rather than background medical text.

3.4 Inference Strategy and Post-Processing

Due to GPU memory constraints during training, sequences were cropped to a maximum length of 2,048 tokens. However, during evaluation, we implemented a non-overlapping sliding window inference mechanism. The document is processed in 2,048-token chunks, and the resulting hidden states are concatenated before being passed to the classification head. This ensures that a single, unified prediction array is generated for notes of any length.

Converting token-level predictions back to the exact character-level offsets required by the official shared task evaluator introduces further

challenges. Standard forward-searching token-to-character mapping frequently misaligns boundaries. We resolved this by implementing a bidirectional search mapping, which eliminated thousands of near-miss offset errors.

Finally, error analysis on our validation set highlighted systematic failure modes that we addressed via a targeted rule-based post-processing pipeline:

1. **Boundary Cleaning:** Striping trailing punctuation (e.g., periods, commas) and whitespaces from predicted spans to match the strict evaluator.
2. **Regex Recovery for Rare Classes:** Employing regular expressions to identify and tag highly specific, heavily missed patterns for *Deferment* (e.g., “*deferred to outpatient follow-up*”) and *Gathering Info* (e.g., “*ordered CT of the chest*”).
3. **Deduplication:** Removing completely identical spans that occasionally arise at the boundaries of sliding inference windows.

4 Experimental Setup

Models were trained using PyTorch and HuggingFace. We trained the Clinical Longformer for 5,000 steps using the AdamW optimizer with a learning rate of $4e-5$ and a 10% linear warmup. Due to the memory footprint of 2,048-token sequences, we utilized a batch size of 4 with 2 gradient accumulation steps (effective batch size of 8). Checkpoints were evaluated every 500 steps, and the best model was selected based on validation Span F1.

5 Results and Analysis

Our system, submitted under the team name *Aurum*, achieved the results detailed in Table 2 on the official hidden test set.

Metric	Score
Final Score	0.5251
Base Score	0.5362
Worst Group Score	0.5140
Token F1	0.6311
Span F1	0.4414

Table 2: Official hidden test set results for Team Aurum.

5.1 Ablation Study

To isolate the contribution of each system component, we conducted an ablation study on our local

validation split. Table 3 illustrates the incremental performance gains starting from a standard 512-token baseline.

Model Configuration	Token F1	Span F1
Baseline (ELECTRA-Base, 512)	0.452	0.301
+ Clinical Longformer (2048)	0.584	0.387
+ Data Augmentation	0.612	0.415
+ Focal Loss	0.625	0.430
+ Post-Processing (Final)	0.631	0.441

Table 3: Incremental ablation study showing the impact of each architectural and methodological choice.

The transition from a 512-token baseline to the Clinical Longformer yielded the most substantial improvement (+0.132 Token F1), confirming that document truncation was the primary bottleneck. The tripartite data augmentation pipeline provided the second-largest boost (+0.028 Token F1), validating our hypothesis that injecting linguistic and rare-class diversity mitigates overfitting. Finally, Focal Loss and rule-based post-processing provided targeted gains, specifically improving boundary alignment and the recall of heavily starved classes like *Deferment*.

5.2 Error Analysis

Span vs. Token F1 Trade-off: The results highlight a pronounced gap between Token F1 (0.6311) and Span F1 (0.4414). This pattern aligns with the task organizers’ observations regarding the difficulty of exact span extraction. Our high Token F1 indicates that the Longformer successfully locates the correct semantic regions; however, the strict character-level exact-match required by the Span F1 evaluator harshly penalizes minor boundary discrepancies (e.g., including an extra article like “a” or missing a trailing word). Precise span calibration remains a bottleneck.

Subgroup Robustness: A key feature of our performance is the minimal gap between the Base Score (0.5362) and the Worst Group Score (0.5140), a difference of only 0.022 points. This indicates that our model exhibits strong fairness across the evaluated demographic buckets. We attribute this robustness to two factors: the extensive linguistic variety introduced by our LLM paraphrasing augmentation, and the inherent demographic representation captured within the MIMIC-III pre-training of the Clinical Longformer backbone.

6 Conclusion

In this paper, we presented team *Aurum*'s submission to the MedExACT 2026 shared task. By replacing standard 512-token encoders with a Clinical Longformer, we solved the critical issue of document truncation in ICU discharge summaries. Furthermore, leveraging a multi-strategy data augmentation pipeline and Focal Loss allowed us to successfully train a robust model on a small, highly imbalanced dataset. While our approach yielded strong demographic stability and token-level detection, future work must focus on advanced boundary-refinement modules to bridge the gap between Token F1 and exact-match Span F1.

References

- Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022. Association for Computational Linguistics.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Mohamed Elgaar, Hadi Amiri, Mitra Mohtarami, and Leo Anthony Celi. 2025. MedDecXtract: A clinician-support system for extracting, visualizing, and annotating medical decisions in clinical narratives. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 481–489, Vienna, Austria. Association for Computational Linguistics.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Hadi Amiri, and Leo Anthony Celi. 2024. MedDec: A dataset for extracting medical decisions from discharge summaries. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16442–16455, Bangkok, Thailand. Association for Computational Linguistics.
- Mohamed Elgaar, Jiali Cheng, Nidhi Vakil, Mehrnaz Sadrolashrafi, Mitra Mohtarami, Adrian Wong, Hadi Amiri, and Leo A. Celi. 2026. [Overview of medical decision extraction, analysis, and classification task \(medexact\) 2026](#). In *The 25th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, San Diego, California, USA. Association for Computational Linguistics.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, Liwei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1).
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.
- Eirik H Ofstad, Jan C Frich, Edvin Schei, Richard M Frankel, and Pål Gulbrandsen. 2016. What is a medical decision? a taxonomy based on physician statements in hospital encounters: a qualitative study. *BMJ Open*, 6(2):e010098.

NJUST-KMG at MedGenVidQA 2026: Cascade Multi-modal Alignment with Gaussian Soft Priors for Medical Visual Answer Localization

Jinglong Li and Yang Yang

School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing, Jiangsu, China
{jinglong_555, yyang}@njjust.edu.cn

Abstract

This paper describes the system developed for the Medical Visual Answer Localization (MVAL) task at MedGenVidQA 2026. Accurately locating surgical or instructional steps in medical videos is inherently challenging due to audio-visual asynchrony and the visual homogeneity of surgical scenes. We propose a Cascade Multi-modal Alignment Framework that integrates Large Language Models (LLMs) to bridge the semantic-temporal gap. Our pipeline utilizes WhisperX for word-level speech transcription to ensure precise textual anchoring. We then employ Gemini3 as a high-level semantic ranker to generate multi-scale textual priors. Crucially, we transform these discrete semantic scores into a continuous 1D Gaussian Soft Prior, which is injected as an attention bias into our cross-modal fusion network. This mechanism preserves global temporal context while guiding the model to focus on query-relevant frames. Our system achieves highly competitive performance on the validation leaderboard, particularly under strict evaluation metrics, reaching an IoU@0.7 of 67.5%.

1 Introduction

Temporal Video Grounding (TVG) in the medical domain, specifically Visual Answer Localization (VAL), requires a system to predict the precise start and end timestamps of a video segment that answers a given medical query. Unlike generic instructional videos, medical and surgical videos often feature high visual redundancy and "semantic drift," where the practitioner's verbal explanation does not perfectly align with the visual execution of a step.

Most existing cascade localization systems rely on "hard truncation," where a text-retrieval module first crops the video into short candidate clips. While efficient, this approach is prone to "boundary collapse" if the initial text-based window is too narrow or slightly shifted, leading to significant

performance degradation at high Intersection over Union (IoU) thresholds.

To overcome these limitations, we propose a soft-prior alignment strategy inspired by recent successes in event segmentation (Zhai et al., 2020). Our contributions are three-fold:

1. We implement a precision ASR pipeline using WhisperX to extract word-level timestamps, ensuring that the foundational textual data is accurately anchored to the video timeline.
2. We design a semantic scoring module using Gemini3 that aggregates multi-scale textual windows to capture both local actions and global surgical stages.
3. We introduce a 1D Gaussian Soft Prior that serves as an attention guidance mechanism, allowing the visual module to refine boundaries beyond the speech-indicated segments.

2 Related Work

2.1 Medical Visual Answer Localization

The MedVidQA and MedGenVidQA benchmarks (Gupta and Demner-Fushman, 2022; Gupta et al., 2023, 2025, 2026) have established the Medical Visual Answer Localization (MVAL) task as a core challenge in multi-modal healthcare AI. Early approaches relied heavily on pure visual span predictors (Gao et al., 2017; Zhang et al., 2020), which often struggled with the visual homogeneity of endoscopic or laparoscopic footage. Subsequent research demonstrated that leveraging ASR transcripts can act as a powerful prior for localization.

2.2 LLMs for Multi-modal Grounding

Recent trends in TRECVID challenges show a shift toward using Large Language Models (LLMs) as zero-shot or few-shot reasoners. Specifically, the use of LLMs to generate temporal "proposals"

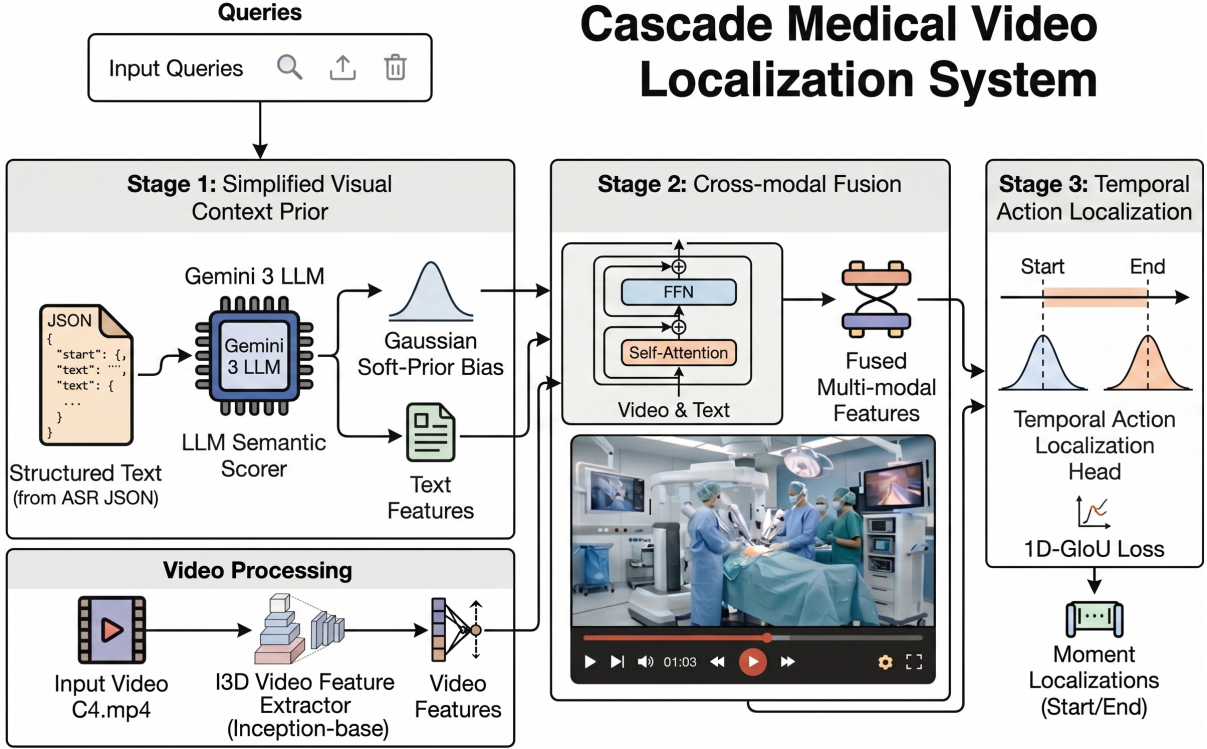


Figure 1: Overall architecture of the proposed Cascade Medical Video Localization System. The framework leverages a 1D Gaussian Soft Prior to inject textual relevance directly into the cross-modal fusion transformer via Attention Bias, ensuring high-precision boundary regression without the risks of hard truncation.

or "summaries" has proven effective (Zhang et al., 2023; Ren et al., 2023). Our work builds upon these insights by moving from discrete proposal generation to continuous probabilistic weighting.

3 Methodology

Figure 1 illustrates the complete architecture of our proposed Cascade Multi-modal Alignment Framework. It consists of three main stages: (1) Precision ASR extraction and multi-scale aggregation; (2) LLM-based semantic scoring and 1D Gaussian Soft Prior generation; and (3) Cross-modal fusion with attention biasing for final localization.

3.1 ASR and Multi-scale Construction

As shown in Figure 2, the frontend utilizes a specialized pipeline for audio-to-text transcription. We employ WhisperX, which improves upon standard Whisper models by using forced alignment (e.g., via Wav2Vec2) to provide word-level timestamps.

For each identified segment s_i with text T_i , we define the interval as $[t_{start}^i, t_{end}^i]$. To capture varying semantic granularities and mitigate the issue of fragmented speech, we construct multi-scale windows $W_{m,k}$ by aggregating m consecutive segments. This ensures that the system can capture

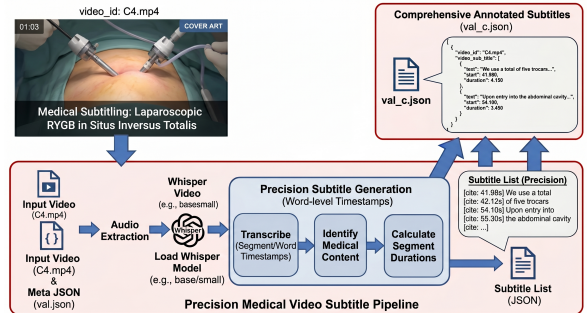


Figure 2: The Precision Medical Video Subtitle Pipeline utilizing WhisperX/Wav2Vec2 for extracting forced word-level timestamps.

quick actions (e.g., "cutting a suture") as well as prolonged stages (e.g., "organ dissection").

3.2 LLM Semantic Ranking

We utilize Gemini3 to calculate the relevance score S_j for each window W_j relative to the medical query Q . Instead of direct generation, we extract the logits for the positive ("Yes") token to obtain a continuous probability $p_j = P(\text{Relevant}|Q, W_j)$. This provides a more granular signal than binary classification.

3.3 1D Gaussian Soft Prior Modeling

Given the highest-scoring window $[t_{start}, t_{end}]$, we define the temporal center $\mu = (t_{start} + t_{end})/2$. To allow the model to explore boundaries outside this window, we model the temporal prior as a 1D Gaussian distribution $G(t)$:

$$G(t) = \exp\left(-\frac{(t - \mu)^2}{2\sigma^2}\right) \quad (1)$$

where σ is a hyperparameter that controls the "width" of the soft prior. This distribution represents the probability that a specific frame at time t belongs to the answer segment according to the linguistic evidence.

3.4 Cross-modal Attention Biasing

We extract dense spatio-temporal visual features $V \in \mathbb{R}^{L \times d}$ using a pre-trained I3D (Inflated 3D ConvNet) encoder. Unlike frame-level spatial extractors, I3D inherently captures local motion dynamics, which synergizes with our temporal soft prior. Let H_v be the visual hidden states and H_q be the query embedding. In the cross-modal fusion layer (Stage 2 in Figure 1), we inject $G(t)$ as an **Attention Bias**. Modifying the pre-softmax attention scores with explicit structural biases has been shown to effectively guide the attention distribution (Press et al., 2022). The standard cross-attention matrix A is modified as follows:

$$A_{i,j} = \text{Softmax}\left(\frac{Q_i K_j^T}{\sqrt{d}} + \lambda \cdot \log(G_j + \epsilon)\right) \quad (2)$$

where λ is a scaling factor. This bias guides the attention mechanism to prioritize frames within the high-confidence Gaussian region while still permitting the visual network to attend to outlying frames if the visual evidence is strong.

4 Experimental Setup

4.1 Datasets

We evaluate our system on the MedGenVidQA 2026 validation set. Evaluation is based on IoU at thresholds $\{0.3, 0.5, 0.7\}$ and mean IoU (mIoU).

4.2 Implementation Details

For the visual stream, the I3D feature extractor processes the video using a 16-frame sliding window without overlap, yielding a continuous feature sequence that perfectly aligns with our 1D Gaussian bias temporal resolution. We utilize LoRA(Hu

et al., 2021) with a rank $r = 32$ for parameter-efficient adaptation. Detailed hyperparameters are provided in Table 1.

Hyperparameter	Value
Visual Backbone	I3D
LoRA Rank (r)	32
LoRA Alpha (α)	64
Optimizer	AdamW
Learning Rate	1×10^{-4}
Gaussian σ	Dynamic
Bias Scale λ	0.5

Table 1: Hyperparameter configurations for the proposed system.

5 Results and Analysis

5.1 Quantitative Results

Table 2 compares our "Soft Prior" approach with baseline methods. Our system significantly outperforms the "Hard Truncation" baseline, especially at the strict IoU@0.7 threshold.

Approach	IoU@0.3	IoU@0.5	IoU@0.7
Pure Visual	68.75	52.5	26.25
Hard Truncation	86.25	75.0	60
Soft Prior	92.5	81.25	67.5

Table 2: Comparative performance results on MedGenVidQA.

5.2 The Impact of Soft Priors

The primary advantage of the 1D Gaussian Soft Prior is its ability to handle asynchrony. In surgical instructional videos, a surgeon might say "Now I am suturing the incision" while the visual action of suturing continues for several seconds after the speech ends. Hard truncation would lose the end of this action, whereas our soft prior allows the visual module to "track" the action to its true conclusion.

5.3 Error Analysis

We observed performance drops in two specific scenarios: (1) **Silent Videos**: In videos where instructions are purely visual or provided via on-screen text, our ASR-based prior fails. (2) **Implicit Semantics**: In cases involving sarcasm or complex medical metaphors, the LLM scoring becomes noisy, leading to a flattened Gaussian distribution.

6 Discussion and Future Work

Our findings align with previous work in TRECVID QFISC (Awad et al., 2023), suggesting that the audio modality is often more informative for step boundary detection than the raw visual stream. However, the fusion of these modalities remains non-trivial. Future work will explore Mixture-of-Experts (MoE) architectures to dynamically weight the audio and visual streams based on the detected presence of speech or on-screen text.

7 Conclusion

Team NJUST-KMG’s submission for MedGenVidQA 2026 leverages a cascade multi-modal framework. By combining precision ASR extraction with LLM-driven Gaussian soft priors, we successfully bridge the semantic gap in medical video localization. Our results demonstrate that replacing rigid temporal cropping with attention-guided soft priors is a superior strategy for achieving high-precision boundary regression.

References

- George Awad, Asad A Butt, Jonathan Fiscus, Martial Michel, David Joy, and 1 others. 2023. Trecvid 2023: Evaluation of video activity detection, video captioning and retrieval, and video to text. *arXiv preprint arXiv:2401.xxxx*.
- Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.
- Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2023. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158.
- Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2025. A dataset of medical questions paired with automatically generated answers and evidence-supported references. *Scientific Data*, 12(1):1035.
- Deepak Gupta, Collin Scott Campbell, Pedram Golnari, and Dina Demner-Fushman. 2026. Overview of the medgenvidqa 2026 shared task on medical generative video question answering. In *Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026)*, San Diego, USA. Association for Computational Linguistics.
- Deepak Gupta and Dina Demner-Fushman. 2022. Medvidqa: A multilingual and multimodal dataset for medical video question answering. In *Proceedings*

of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1302–1318.

- Edward J Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Ofir Press, Noah A Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.
- Shuhuai Ren and 1 others. 2023. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*.
- Shao Huang Zhai and 1 others. 2020. Event segmentation with action boundaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1011–1020.
- Hang Zhang, Xin Li, and Bing Lidong. 2023. Videollama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.
- Songyang Zhang, Houwen Peng, Jianlong Jian, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877.

LAMAR-2 at MedGenVidQA 2026: Visual Answer Localization in Medical Videos via Multimodal LLM and Context-Augmented Prompting

Watcharitpol Sermsrisuwan^{1,2}, Nopporn Lekuthai^{1,2}, Seksan Yoadsanit^{1,2},
Titipat Achakulvisut¹

¹ Department of Biomedical Engineering, Faculty of Engineering, Mahidol University,
Nakhon Pathom, Thailand

² Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

Correspondence: titipat.ach@mahidol.ac.th

Abstract

This paper presents an approach to localizing visual answers within continuous medical videos using a multi-step multimodal generation pipeline with the MedGenVidQA dataset. We frame visual answer localization as a multimodal fusion problem, integrating raw video, timestamped ASR transcripts, and VLM-generated scene descriptions into structured contextual blocks, enabling the model to cross-reference spoken commentary against observable physical events. We show that targeted guidance, which forces the model to treat audio transcripts as supplementary hints with observable visual movements, significantly outperforms baseline approaches. It achieves state-of-the-art performance on the test leaderboard, yielding an mIoU of 79.55, alongside IoU@0.3, IoU@0.5, and IoU@0.7 scores of 93.75, 90.00, and 77.50, respectively. Our findings highlight the effectiveness of combining multimodal context fusion with targeted guidance to overcome text bias, establishing a promising approach for achieving the micro-level precision required in the medical domain. We release our code on GitHub at github.com/biodatlab/medgenvidqalamar.

1 Introduction

Multimodal data captures the complexity of medicine and advances medical question answering (AISaad et al., 2024). Medical video question answering (MedVidQA) with large language models (LLMs) supports healthcare professionals by providing information from procedural videos. (Li et al., 2024a) State-of-the-art models generate fluent, in most cases, medically accurate responses but lack spatiotemporal precision to localize visual answers in videos (Xiao et al., 2023).

The shift toward multimodal large language models (MLLMs) has improved audio-visual QA by combining spatial and temporal cues with textual data (Zhang et al., 2023a). These models succeed

in understanding general video but reveal a gap in the medical domain (Wang et al., 2025). General-purpose MLLMs often suffer from temporal hallucinations, providing broad or incorrect timestamps that fail to pinpoint clinical task boundaries (Li et al., 2025). In medical procedural, identifying these requires more than simple object recognition. It demands a strict visual distinction between the preparation and the actual start of a medical step (Timoh et al., 2023).

We propose a novel approach for localizing visual answers in the MedGenVidQA 2026 Task C dataset (Gupta et al., 2026). This task requires pinpointing precise start and end timestamps within continuous medical videos to directly answer a natural language clinical query. Our empirical analysis reveals that arbitrary text chunking fragments the continuous chronological flow required for observable physical events. Therefore, we reframe visual answer localization as an end-to-end multimodal generation task. Our architecture integrates timestamped ASR transcripts, VLM-generated structural scene descriptions, and raw video directly into the Gemini-3-Flash model. Guided by procedurally-targeted heuristic prompt that prioritizes physical visual movements over auditory dialogue, our system achieves state-of-the-art IoU@0.3, IoU@0.5, IoU@0.7, and mIoU scores of 93.75, 90.00, 77.50, and 79.55, respectively, on the leaderboard.

2 Related work

Medical Visual Answer Localization (MVAL) benchmarks, such as MedVidQA and MedVidCL (Gupta et al., 2023), alongside automated corpora like HealthVidQA (Gupta et al., 2024), establish the foundation for medical instructional video analysis. While Temporal Answer Grounding in Videos frequently treats localization as a text-span prediction problem—relying on subtitles and learned visual prompts (Li et al., 2024b)—this approach

struggles with the temporal asynchrony of complex medical procedures because spoken commentary often misaligns with physical execution.

Temporal sentence grounding isolates video segments that match natural language queries (Zhang et al., 2023b). While classical methods relied on cross-modal interactions to predict boundaries (Cao et al., 2021; Wu et al., 2020; Zhang et al., 2021), the field has recently shifted toward unified frameworks that consolidate grounding tasks for large-scale pretraining and zero-shot application (Lin et al., 2023).

Despite the proven potential of multimodal LLMs in healthcare, accurate temporal grounding in medical videos remains a persistent challenge (AlSaad et al., 2024; Chen et al., 2025; Xiao et al., 2023). Recent corpus-level frameworks attempt to optimize large-scale retrieval and localization using subtitle enhancement and contrastive learning (Zhang et al., 2024). However, their fundamental reliance on textual alignment excels only at macro-level identification, leaving the micro-temporal asynchrony of individual surgical procedures unresolved. This gap highlights the necessity for strictly bounded, visually anchored multimodal fusion strategies. These strategies ensure that localization accurately reflects the physical execution in the videos.

3 Datasets

We utilized the MedVidQA dataset (Gupta et al., 2023), originally comprising 3,010 human-annotated QA pairs from 900 health-related videos. Due to platform-level download restrictions, we use and evaluate 49 unique videos consisting of 148 QA pairs from the dataset. The development set features a mean video length of 445.62 ± 239.89 seconds and an average annotation span of 57.72 ± 41.84 seconds. For the final evaluation, we used the official MedGenVidQA 2026 Task C test data set (Gupta et al., 2026), which contains 80 medical questions assigned to 65 unique instructional videos with a mean video duration of 544.72 ± 432.74 seconds.

As the organizers withhold the official test ground truth, we conducted a manual annotation of the 80 test QA pairs. We utilized these annotations exclusively for qualitative error analysis, ensuring our primary quantitative metrics reflect our official leaderboard submission.

4 Methodology

We frame temporal localization as a generative prediction task: given a medical video and a clinical query, a large language model (LLM) must output the start and end timestamps of the relevant action. Our core hypothesis is that enriching the LLM’s input with structured, multi-source context yields more accurate localization than any single modality alone. To test this, we evaluated context configurations using transcript only, video only, and their combination with VLM-generated scene descriptions to identify the strongest data representation.

4.1 Unimodal Baselines

To isolate the predictive contribution of each modality, we established two baselines. **Transcription-Only** provides the LLM with only the clinical query and the full ASR transcript composed of timestamp-aligned, sentence-level groupings, predicting temporal boundaries from spoken commentary alone. **Video-Only** provides the LLM with only the clinical query and the raw untrimmed video, relying entirely on visual reasoning without any textual input. These baselines quantify whether temporal boundaries are primarily signaled by spoken commentary or by observable physical actions, and they provide a lower bound for multimodal integration.

4.2 Retrieval-Augmented Generation (RAG) on Transcripts

Medical transcripts are often lengthy and noisy. Rather than feeding the full transcript to the LLM, we also experimented using RAG (Lewis et al., 2020) to select only the most query-relevant segments. We evaluated two complementary partitioning schemes, each tuned empirically on the development set: **time-based chunking and retrieving**, which applies a temporal sliding window with a fixed overlap to preserve context across boundaries and retrieve a fixed number of top candidates, and **sentence-based chunking and retrieving**, which applies a sentence-count sliding window with a fixed overlap to maintain chronological continuity. We utilized a dynamic percentage threshold for retrieval to ensure proportional and adaptive content coverage.

Temporal Reciprocal Rank Fusion.

Because arbitrary chunk boundaries can split a continuous medical action, we designed a rank-based temporal merging algorithm (Rackauckas, 2024).

Each retrieved chunk, c , receives a base weight inversely proportional to its retrieval rank. Adjacent chunks whose time spans overlap are iteratively fused into a single segment, C , with an accumulated relevance score:

$$\text{Score}(C) = \sum_{c_i \in C} \frac{1}{\text{rank}(c_i)} \quad (1)$$

Merging continues until no remaining candidates overlap. The fused segment with the highest accumulated score is selected as the context block for the generative model.

4.3 Generating Scene Description Context

To produce a structured visual context, we partitioned each video into segments at natural camera cuts using PySceneDetect (Castellano and contributors, 2025). Each segment was then passed through Qwen3-VL (Bai et al., 2025), which generated a textual description of the observable physical events within that time span. We temporally aligned the ASR transcript with these VLM-generated scene descriptions. The resulting fused contextual blocks, each containing both the visual scene descriptions and spoken dialogue for a specific time span, served as the structured context input for the LLM (Figure 1).

4.4 VLM-Enhanced Context

We hypothesize that enriching the LLM’s input with structured, multi-source context, rather than relying on any single modality, yields more accurate temporal localization. Specifically, by fusing raw video with fused contextual blocks, the model can cross-reference spoken commentary against observable physical events, resolving ambiguities that neither modality addresses alone. Under this hypothesis, the primary lever for improving performance is not architectural modification or retrieval tuning, but contextual guidance: carefully designing what information the LLM sees and how it is structured.

We evaluated two approaches that combine visual information and ASR transcripts. **Transcript-Video Fusion** processes the raw video, the full ASR transcript, and the query through the LLM. **VLM-Enhanced Context** processes the raw video with fused contextual blocks, and the query. This approach adds observable physical events rather than transcripts alone.

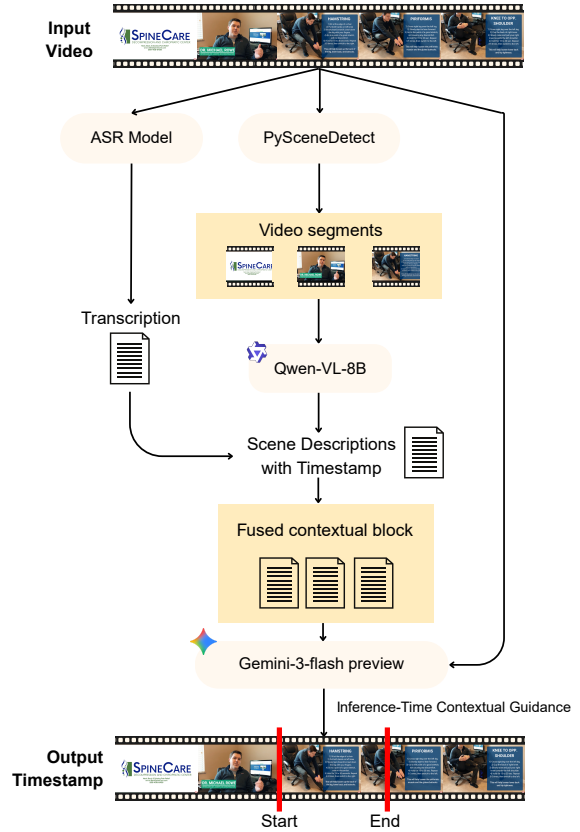


Figure 1: Overview of the proposed VLM-Enhanced Context pipeline for medical video analysis.

4.5 Boundary Trimming

Across all pipelines, final temporal localization was performed by Gemini-3-Flash. The data representation varied by pipeline, ranging from full transcript, RAG-filtered context block, raw video, to video with fused contextual block, to output precise start and end timestamps for the queried action.

4.6 Inference-Time Contextual Guidance

To improve upon the model’s clinical reasoning and enforce temporal precision within the surgical domain, we evaluated four constraint-based guidance. Each strategy constrains how the MLLM weighs textual evidence against visual observations during inference.

- **Zero-Shot with Reasoning:** The model was instructed to directly predict temporal boundaries, explicitly requiring a justification of its visual logic (intermediate reasoning) prior to outputting the final timestamps.
- **Strict Visual Boundary:** The instruction explicitly guide the model to prioritize visual evidence over auditory commentary exclud-

ing unrelated sections such as pre-operative verbal planning, introductory remarks.

- **Chain of thought (CoT):** The model was required to identify specific observable physical events e.g., scalpel makes first tissue contact and the end of the event. This intermediate reasoning step prevents the model from copying timestamps from the transcript.
- **Heuristic Context (Loose):** In this strategy, the transcript and audio were treated as supplementary heuristics only, and boundaries were derived from physical movements in the video.

5 Evaluation

We measure temporal localization performance using Intersection over Union (IoU) between the predicted and ground-truth time intervals. MedVidQA benchmark (Gupta et al., 2023) reports accuracy at three IoU thresholds (0.3, 0.5, and 0.7) where a prediction is counted as correct if its IoU with the ground truth meets or exceeds the threshold.

6 Results and Discussions

Our experimental framework evaluates visual answer localization as an end-to-end generative process (Figure 1). Across all pipelines, final boundaries were predicted by Gemini-3-Flash. Inputs varied by configuration: text-driven approaches processed only the query and textual context, while multimodal approaches directly ingested raw video alongside the query, transcript, and scene description.

6.1 Development Phase: Transcription-only Pipeline Comparison

To generate the transcript context, we initially evaluated three transcription models: Parakeet-tdt-0.6b-v2 (NVIDIA, 2024), Whisper large-v3 (Radford et al., 2022), and Qwen3-ASR-1.7B (Shi et al., 2026). To standardize the input, word-level outputs were programmatically grouped into timestamp-aligned, sentence-level blocks.

Quantitative analysis of these unimodal transcription-only baselines revealed that Parakeet achieved the highest overall mIoU (66.03), demonstrating a slight edge over Qwen3-ASR (65.20) in strict boundary regression (Table 1). While both models tied at the IoU@0.3 threshold (82.58), Parakeet outperformed Qwen3-ASR at the

stricter 0.5 and 0.7 thresholds. Whisper large-v3 lagged significantly across all metrics.

Despite Parakeet’s marginal advantage, we selected Qwen3-ASR, as the standard transcription for all subsequent experiments. This architectural decision prioritizes downstream structural utility: Qwen3-ASR with forced aligner generates the precise word-level timestamps essential for successfully implementing text alignment in our multi-modal pipelines.

6.2 Development Phase: Architecture and Modality Evaluation

With the base modalities established, we tuned our integration strategies. For Retrieval-Augmented Generation baselines, hyperparameters were tuned empirically to optimize IoU. Time-based chunking utilized 20-second windows with a 2-second overlap (top- $k=10$), while sentence-based chunking employed 5-sentence blocks with a 4-sentence overlap (dynamic top 40%). The chunks were embedded using google/embeddinggemma-300m (Schechter Vera et al., 2025) and fused using the Temporal RRF algorithm.

As shown in Table 1, RAG strategies performed worst overall. Because standard RAG retrieves text based on semantic keyword matching, it frequently pulled out-of-sync narrative snippets, reflecting the asynchronous nature of instructional medical videos.

Conversely, the Transcript-Video Fusion pipeline achieved the highest overall mIoU of 72.35 and IoU@0.3 of 90.32. However, the Video-Only baseline maintained the highest precision at the strictest threshold. Introducing fused contextual block underperformed the simpler fusion pipeline, suggesting that over-engineering the context block with secondary descriptions or rigid reasoning constraints introduces noise that distracts from the primary visual signal.

6.3 Test Phase Evaluation

In the test phase, the Video-Only baseline achieved a highly competitive mIoU of 79.51. Consistent with development set findings, introducing textual data with complex reasoning constraints (Zero-Shot with Reasoning) significantly degraded performance, dropping the mIoU to 70.95 and IoU@0.7 to 63.75. This confirms that generative models suffer from text-reliance bias (Winterbottom et al., 2020), prematurely aligning boundaries with spoken dialog rather than actual physical interventions.

Pipeline Configuration	Modality	Guidance Prompt	mIoU	IoU@0.3	IoU@0.5	IoU@0.7
<i>Comparison of ASR model</i>						
Transcription-Only (Parakeet)	Transcript Only	Zero-Shot	66.03	82.58	72.26	56.13
Transcription-Only (Qwen3-ASR)	Transcript Only	Zero-Shot	65.20	82.58	71.61	54.84
Transcription-Only (Whisper)	Transcript Only	Zero-Shot	57.86	81.94	60.65	40.65
<i>Visual, Retrieval, and Multimodal Architectures</i>						
Video-Only	Video Only	Zero-Shot	71.78	89.03	78.71	67.10
RAG (Time-based Chunking)	Transcript Only	Zero-Shot	40.09	54.84	38.71	28.39
RAG (Sentence-based Chunking)	Transcript Only	Zero-Shot	46.62	63.23	50.97	32.90
Transcript-Video Fusion	Video + Transcript	Zero-Shot	72.35	90.32	78.71	63.87
VLM-Enhanced Context	Video + Fused contextual block	Zero-Shot with Reasoning	68.75	87.74	79.35	58.06
VLM-Enhanced Context	Video + Fused contextual block	Strict Visual Boundary	67.56	87.10	77.42	56.77
VLM-Enhanced Context	Video + Fused contextual block	Chain-of-Thought (CoT)	68.41	88.39	78.06	60.00
VLM-Enhanced Context	Video + Fused contextual block	Heuristic Context (Loose)	70.59	89.03	78.71	63.87
<i>Test Phase</i>						
Video-Only	Video Only	Zero-Shot	79.51	96.25	87.50	73.75
VLM-Enhanced Context	Video + Fused contextual block	Zero-Shot with reasoning	70.95	90.00	78.75	63.75
VLM-Enhanced Context	Video + Fused contextual block	Strict Visual Boundary	76.44	92.50	85.00	70.00
VLM-Enhanced Context	Video + Fused contextual block	Chain-of-Thought (CoT)	74.53	92.50	83.75	68.75
VLM-Enhanced Context	Video + Fused contextual block	Heuristic Context (Loose)	79.55	93.75	90.00	77.50

Table 1: Development and test phase performance. Development results compare ASR models, unimodal baselines, multimodal fusion, and RAG pipelines. Test results evaluate Inference-Time Contextual Guidance on the best-performing configuration. Test phase metrics reflect official MedGenVidQA leaderboard scores

To counteract this bias, we evaluated explicitly constrained guidances. The Strict Visual Boundary (mIoU 76.44) and Chain-of-Thought (mIoU 74.53) strategies recovered substantial precision by forcing the model to anchor predictions to observable actions. However, these rigid constraints occasionally caused over-correction, leading the model to discard valuable macro-level transcript context.

The Heuristic Context (Loose) strategy proved most effective, achieving the highest overall mIoU (79.55) and peak performance at strict thresholds (IoU@0.5 of 90.00, IoU@0.7 of 77.50). By framing the transcript as a supplementary heuristic rather than heavily penalizing its use, this strategy establishes a two-step framework: macro-localization followed by micro-trimming. The model uses spoken commentary to broadly navigate to the correct procedural phase, but the prompt dictates that final timestamps must strictly bound the active physical procedure, explicitly excluding verbal introductions or planning segments.

While the Heuristic Context offers only a marginal quantitative gain over the Video-Only baseline (mIoU 79.51), it delivers a critical leap in clinical reliability. The Video-Only baseline model efficiently recognizes raw physical movements, but struggles with abstract queries requiring a broader procedural context. The Heuristic pipeline resolves this by using transcripts for macro-localization during visually obscured or repetitive phases, while its strict visual micro-trimming preserves the boundary precision of a dedicated vision model.

6.4 Qualitative Error Analysis

We conducted a qualitative diagnostic review to evaluate the impact of text-reliance bias and our mitigation strategies across both validation and test sets. Since official test labels are withheld, our test set analysis is based on manual inspection of model outputs rather than comparison against gold-standard annotations (Figure 2). We categorize our observations into three scenarios: transcript-assisted correction, transcript-induced bias, and visual guidance as a fix.

Scenario A: Transcript-Assisted Correction. Visual features should theoretically be enough to locate a procedure yet Video-Only baselines sometimes fail due to visual distractions or a lack of step-by-step reasoning. In a validation video of an AC joint test (Sample 2812), the baseline starts at 05:01, capturing the test itself but missing the necessary first step of locating the AC joint (04:18–05:01). This shows that video-only models can jump straight to the most obvious action and ignore the clinical context. In a test video of smile reconstruction (C64), the baseline predicts 03:38–03:42, which shows only a title slide, while missing the actual procedure (03:42–03:51). Our Heuristic Context guidance fixes these errors by using the transcript to ensure all required medical context is included.

Scenario B: Transcript-Induced Bias. The spoken transcript can also mislead the model when narration and physical action are not temporally aligned. In a test video of an endoscopic surgery

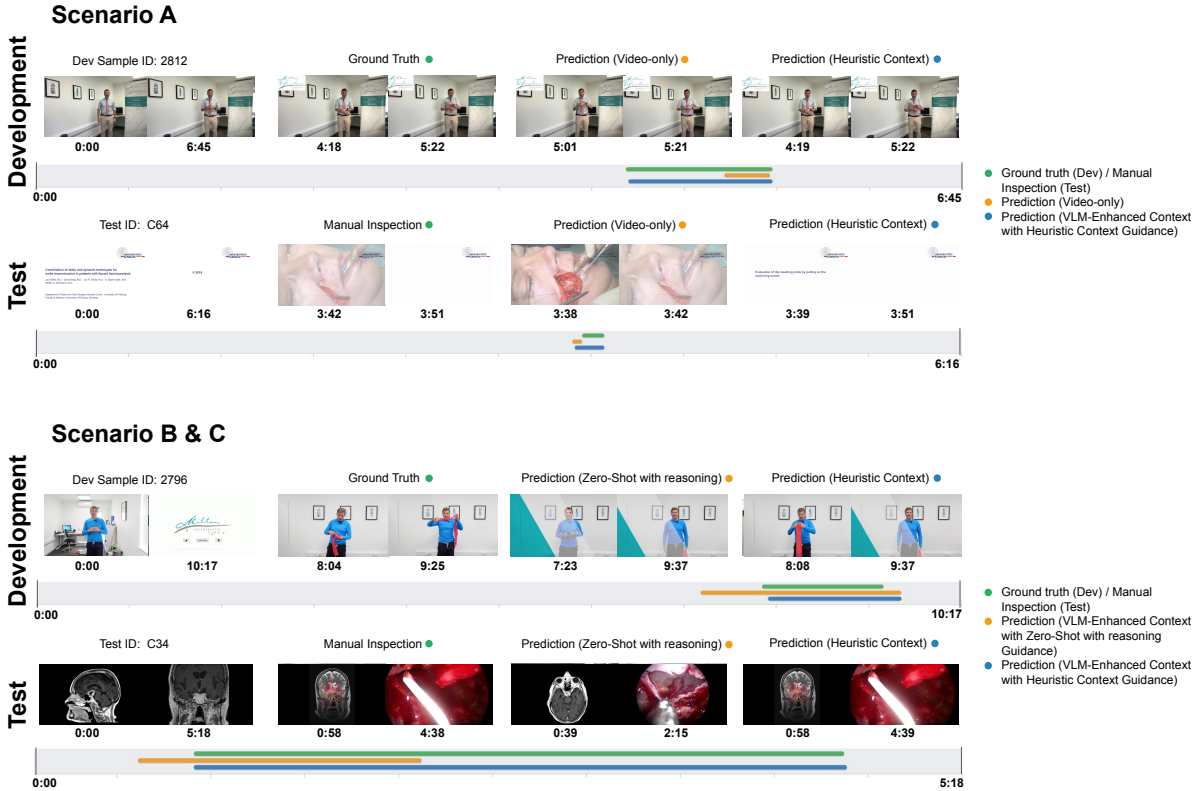


Figure 2: Qualitative error analysis illustrating three scenarios across development and test samples.

(C34), the VLM-Enhanced Context pipeline with Zero-Shot with reasoning guidance starts at 00:39, which is the exact moment the instructor mentions the procedure. However, this includes 19 seconds of an MRI scan before surgery actually begins at 00:58. Similarly, in a validation video of a chin-tuck exercise (Sample 2796), the same pipeline starts at 07:23 during the anatomical explanation, capturing a 41-second gap before the physical exercise starts at 08:04. In both cases, the model anchors on spoken keywords rather than visual evidence, causing it to include verbal explanations.

Scenario C: Visual Guidance as a Fix. Our Heuristic Context guidance corrects this misalignment by requiring visual confirmation before anchoring the start boundary. In the surgical video (C34), the model uses the transcript to find the general timeframe but is guided to wait for the active surgical field to appear, successfully starting at 00:58. In the chin-tuck video (Sample 2796), the pipeline ignores the 41-second explanation and correctly starts at the onset of physical movement (08:08). These results demonstrate that adding visual constraints helps the model resolve the temporal gap between speech and action, resulting in higher boundary precision.

7 Conclusion

In this paper, we addressed the challenge of precise temporal localization in medical videos through a generative, multimodal lens. We identified a vulnerability in standard multimodal LLM architectures: *text-reliance bias*. We demonstrated that naively feeding textual commentary or RAG-retrieved chunks into a generative model degrades temporal precision. The model prematurely aligns boundaries to spoken words rather than physical surgical actions.

To overcome this, we introduced a highly constrained, end-to-end multimodal pipeline utilizing Gemini-3-Flash. By directly fusing raw video with fused contextual block and employing a "Heuristic Context" guidance strategy, we successfully forced the generative model to treat textual context as supplementary hints while anchoring its final temporal boundaries to visual physical evidence. This approach achieved an mIoU of 79.55 and a strict IoU@0.5 of 90.00 on the test set. Our findings suggest that for complex medical video question-answering, decoupling text-based context from actual physical execution through rigorous visual prompting is essential for achieving micro-level temporal precision.

Limitations

Despite its precision, our approach has several limitations. First, final boundary prediction relies on a proprietary model (Gemini-3-Flash), as current open-source alternatives struggle with our strict reasoning constraints. This commercial dependency poses privacy and reproducibility barriers for local clinical deployment, emphasizing the need for fine-tuned, end-to-end open-source systems. Second, our evaluation is limited to the English-centric, procedure-specific MedVidQA dataset; validating our prompting strategies on diverse, multilingual corpora remains essential.

Finally, explicitly prioritizing visual evidence introduces modality-specific vulnerabilities. Our system is highly sensitive to visual degradation common in real-world surgical recordings (e.g., lens obstruction or sudden focus loss). Furthermore, this visual-first approach is ill-suited for dialogue-heavy clinical videos, such as patient history-taking, where temporal cues are inherently audio-driven and require transcription-reliant processing.

References

- Rawan AlSaad, Alaa A. Abd-alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, R. Damseh, and Javaid Sheikh. 2024. [Multimodal large language models in health care: Applications, challenges, and future outlook](#). *Journal of Medical Internet Research*, 26.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. 2021. [On pursuit of designing multi-modal transformer for video grounding](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9810–9823, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Brandon Castellano and contributors. 2025. PySceneDetect: A video scene cut detection and analysis tool. <https://github.com/Breakthrough/PySceneDetect>. Accessed: 2026-04-29.
- Zhen Chen, Xingjian Luo, Kun Yuan, Jinlin Wu, Danny T. M. Chan, Nassir Navab, Hongbin Liu, Zhen Lei, and Jiebo Luo. 2025. [Surglm: A versatile large multimodal model with spatial focus and temporal awareness for surgical video understanding](#). *ArXiv*, abs/2509.00357.
- Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2023. [A dataset for medical instructional video classification and question answering](#). *Scientific Data*, 10(1):158.
- Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2024. [Towards answering health-related questions from medical videos: Datasets and approaches](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16399–16411, Torino, Italia. ELRA and ICCL.
- Deepak Gupta, Collin Scott Campbell, Pedram Golnari, and Dina Demner-Fushman. 2026. [Overview of the medgenvidqa 2026 shared task on medical generative video question answering](#). In *Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026)*, San Diego, USA. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Chaoyu Li, Eun Woo Im, and Pooyan Fazli. 2025. [Vid-halluc: Evaluating temporal hallucinations in multimodal large language models for video understanding](#). In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13723–13733.
- Jiajie Li, Garrett C. Skinner, Gene Yang, Brian R Quaranto, Steven D. Schwaitzberg, Peter C W Kim, and Jinjun Xiong. 2024a. [Llava-surg: Towards multimodal surgical assistant via structured surgical video learning](#). *ArXiv*, abs/2408.07981.
- Shutao Li, Bin Li, Bin Sun, and Yixuan Weng. 2024b. [Towards visual-prompt temporal answer grounding in instructional video](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8836–8853.
- Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. 2023. [Univtg: Towards unified video-language temporal grounding](#). In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2782–2792.
- NVIDIA. 2024. [Parakeet-TDT-0.6B-v2: State-of-the-art speech recognition model](#). <https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2>.

- Zackary Rackauckas. 2024. [Rag-fusion: a new take on retrieval-augmented generation](#). *ArXiv*, abs/2402.03367.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *arXiv preprint*.
- Henrique* Schechter Vera, Sahil* Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 69 others. 2025. [Embeddinggemma: Powerful and lightweight text representations](#).
- Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. [Qwen3-asr technical report](#). *arXiv preprint arXiv:2601.21337*.
- K. Nyangoh Timoh, Arnaud Huaultmé, K. Cleary, Myra A. Zaheer, V. Lavoué, D. Donoho, and P. Janin. 2023. [A systematic review of annotation for surgical process model analysis in minimally invasive surgery based on video](#). *Surgical Endoscopy*, 37:4298 – 4314.
- Guan-Feng Wang, Wenjin Mo, Junyi Wang, Long Bai, Kun Yuan, Ming Hu, Jinlin Wu, Junjun He, Yiming Huang, N. Padoy, Zhen Lei, Hongbin Liu, Nassir Navab, and Hongliang Ren. 2025. [Surgvidlm: Towards multi-grained surgical video understanding with large language model](#). *ArXiv*, abs/2506.17873.
- Thomas Winterbottom, Sarah Xiao, Alistair McLean, and N. A. Moubayed. 2020. [On modality bias in the tvqa dataset](#). *ArXiv*, abs/2012.10210.
- Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. 2020. [Reinforcement learning for weakly supervised temporal grounding of natural language in untrimmed videos](#). In *Proceedings of the 28th ACM International Conference on Multimedia, MM '20*, page 1283–1291, New York, NY, USA. Association for Computing Machinery.
- Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2023. [Can i trust your answer? visually grounded video question answering](#). *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13204–13214.
- Hang Zhang, Xin Li, and Lidong Bing. 2023a. [Video-llama: An instruction-tuned audio-visual language model for video understanding](#). pages 543–553.
- Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Siow Mong Rick Goh. 2021. [Parallel attention network with sequence matching for video grounding](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 776–790, Online. Association for Computational Linguistics.
- Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2023b. [Temporal sentence grounding in videos: A survey and future directions](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10443–10465.
- Xuemei Zhang, Peng Zhao, Jinsheng Ji, Xiankai Lu, and Yilong Yin. 2024. [Video corpus moment retrieval via deformable multigranularity feature fusion and adversarial training](#). *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6686–6698.

A Appendix

A.1 Inference-Time Contextual Guidance Templates

In this section, we provide the complete prompt templates used to evaluate our final multimodal architecture on the test dataset. Because optimal temporal localization requires careful cross-modal reasoning, we experimented with varying levels of constraint-based guidance. The following templates illustrate the progression from a baseline extraction method (Zero-Shot with Reasoning) to increasingly constrained visual-priority strategies (Strict Visual Boundary, CoT, and Heuristic Context).

A.2 Implementation Details and Hyperparameters

All multimodal generative inferences were executed utilizing the Gemini API (gemini-3-flash-preview). Experiments were systematically conducted between February and April 2026. The model operated with a knowledge cutoff of January 2025 and utilized a context window of 1 million input tokens and 64,000 output tokens. We enforced a JSON output constraint via the API configuration (`response_mime_type="application/json"`). All sampling hyperparameters were retained at their model defaults, notably a Temperature of 1.0, with the internal reasoning parameter (thinking level) set to high (dynamic).

Prompt: Zero-Shot with Reasoning

Role: You are an expert medical video analyst. You will be provided with a medical question, a medical video, and supplemental data (including visual context descriptions and audio transcripts).

Your Task:

Identify the precise timestamp boundaries in the video that comprehensively answers the user's question. **Visual evidence is your absolute highest priority.** You must prioritize what is actively being shown or demonstrated on screen, using the additional data primarily as supporting verification.

Step-by-Step Instructions:

1. **Analyze:** Analyze the given medical question to understand exactly what information, anatomical structure, or procedure is being asked about.
2. **Locate (Visuals First):** Find the exact moment the video *visually* demonstrates the answer. If the audio discusses it but the visuals do not show it, exclude it.
3. **Refine Boundaries:** Do not simply copy the provided supplemental segment times. Crop your timestamps tightly to the exact start and end of the visual answer.
4. **Explain:** Briefly justify your visual logic in the "reasoning" field.

Output Format:

You **MUST** output a valid JSON object **ONLY**. Do not include any conversational text outside of the JSON block. Use the following exact keys and data types:

```
{
  "reasoning": "A brief explanation
of why these specific visuals answer
the question.",
  "answer_start": "MM:SS",
  "answer_end": "MM:SS"
}
```

Inputs:

The Question: {question}

Context and Transcription: {Additional}

Figure A: The system prompt for the Zero-Shot Baseline pipeline.

Prompt: Strict Visual Boundary

Role: You are an expert medical video analyst. Your task is to identify the exact, continuous video segment that answers a given medical question using the provided video, visual context, and audio transcripts.

Core Directives:

1. **Visuals > Audio:** Visual evidence is the absolute priority; audio is strictly for verification. Do not select segments where the action/anatomy is discussed but not visually demonstrated.
2. **Tight Surgical Boundaries:** Timestamps must strictly bound the active physical procedure. Exclude all introductions, verbal planning, and text slides. Start exactly when the real operation begins and end exactly when it finishes.
3. **Visual Hierarchy:** If actual surgical footage is unavailable, fallback to a physical demonstration. If that is also unavailable, fallback to a text-based explanation.

Output Format:

Output **ONLY** a valid JSON object. Do not include any conversational text, markdown formatting outside the JSON, or explanations outside the "reasoning" key.

```
{
  "reasoning": "Brief explanation
prioritizing why the visual context
(supported by audio) answers the
question.",
  "answer_start": "MM:SS",
  "answer_end": "MM:SS"
}
```

Inputs:

The Question: {question}

Context and Transcription: {Additional}

Figure B: The system prompt for the Strict Visual Boundary strategy.

Prompt: Chain-of-Thought (CoT)

Role: You are an expert medical video analyst. Your task is to identify the exact, continuous video segment that answers a given medical question using the provided video, visual context, and audio transcripts.

Core Directives:

1. **Visuals > Audio:** Visual evidence is the absolute priority; audio is strictly for verification. Do not select segments where the action/anatomy is discussed but not visually demonstrated.
2. **Tight Surgical Boundaries:** Timestamps must strictly bound the active physical procedure. Exclude all introductions, verbal planning, and text slides. Start exactly when the real operation begins and end exactly when it finishes.
3. **Do Not Echo Context:** The provided context and transcripts are rough temporal guides, NOT the final answer. You must independently discover the micro-boundaries within them. Never blindly copy the timestamps or durations of the provided input chunks.
4. **Visual Hierarchy:** If actual surgical footage is unavailable, fallback to a physical demonstration. If that is also unavailable, fallback to a text-based explanation.
5. **Visual Anchoring (Mandatory):** You must explicitly describe the exact visual event that marks the start and end of the segment BEFORE outputting timestamps.

Output Format:

Output ONLY a valid JSON object. Do not include any conversational text, markdown formatting outside the JSON, or explanations outside the specified keys.

```
{
  "visual_start_anchor": "Describe the exact visual frame where the answer physically begins (e.g., 'Scalpel makes first contact with skin').",
  "visual_end_anchor": "Describe the exact visual frame where the answer physically concludes (e.g., 'Suture is cut and tool is removed from frame').",
  "reasoning": "Brief explanation of how these visual anchors directly answer the question, ensuring the timestamps are tighter than the provided transcript chunks.",
  "answer_start": "MM:SS",
  "answer_end": "MM:SS"
}
```

Inputs:

The Question: {question}

Rough Segments (Context/Audio): {Additional}

Prompt: Heuristic Context (Loose)

Role: Expert Medical Video Analyst.

Task: Identify the exact video segment that answers the question.

Inputs:

Question: {question}

Reference Notes (Transcripts & Scenes): {Additional}

Instructions:

Watch the video. The Reference Notes are provided only as a background hint. You must determine the precise start and end timestamps purely by observing the physical procedure in the video footage.

Output Format:

Output ONLY a valid JSON object:

```
{
  "first_physical_movement":
    "Briefly state the visual action that starts the segment.",
  "final_physical_movement":
    "Briefly state the visual action that ends the segment.",
  "answer_start": "MM:SS",
  "answer_end": "MM:SS"
}
```

Figure D: The system prompt for the Heuristic Context (Loose) strategy.

Figure C: The system prompt for the Chain-of-Thought (CoT) strategy.

Varja-Dominators at MedGenVidQA 2026: Hybrid Video and Document Retrieval using PubMedBERT, T5 Query Expansion, and Cross-Encoder Re-Ranking

Pratik Dhaktode Computer Engineering PICT, Pune Maharashtra, India pratik.dhaktode.143@gmail.com	Suhani Bighane AI and Data Science PICT, Pune Maharashtra, India bighanesuhani@gmail.com	Anupama Phakatkar Computer Engineering PICT, Pune Maharashtra, India agphakatkar@pict.edu
---	---	--

Abstract

This paper presents our submission to Task A of the MedGenVidQA 2026 shared task (Gupta et al., 2026), which challenges participating systems to simultaneously retrieve relevant PubMed documents and medical videos for 60 consumer health topics (A1–A60). Rather than treating these as independent retrieval problems, we design a unified multi-stage pipeline combining dense and sparse retrieval with cross-encoder re-ranking. For video retrieval, we fine-tune a PubMedBERT bi-encoder on the MedVidQA training set (2,710 samples across 800 unique videos) using BM25-driven hard negative mining. Dense retrieval results are fused with BM25 sparse scores through weighted Reciprocal Rank Fusion (RRF), and queries are expanded using a T5-based doc2query model to bridge the vocabulary gap between consumer-phrased questions and clinical text. A cross-encoder provides final re-ranking. For document retrieval, we query the NCBI PubMed E-utilities API with a progressive keyword fallback strategy. Our system achieves a MAP of 0.4229, Recall@5 of 0.6903, Recall@10 of 0.7226, and NDCG@10 of 0.4971, with complete 60/60 topic coverage for both modalities.

1 Introduction

The proliferation of medical video content on platforms such as YouTube has created both an opportunity and a challenge for health information retrieval. Consumers increasingly seek answers to clinical and procedural questions through video, yet most IR systems index only textual documents such as PubMed abstracts. The MedGenVidQA 2026 shared task (Gupta et al., 2026) directly addresses this gap by requiring systems to perform *dual retrieval*: identifying both relevant PubMed documents and relevant instructional medical videos for consumer health topics.

Task definition. Task A requires each system to submit, for each of 60 test topics (A1–A60), a ranked list of up to 10 PubMed document IDs (PMIDs) and up to 10 video IDs. The test set comprises 155 question-answer samples spanning 50 unique YouTube videos. The input is a natural-language consumer health question; the output is two ranked lists of at most 10 items each.

Challenges. The primary challenges are: (1) *vocabulary mismatch*—consumer health questions use colloquial terminology differing from clinical language in abstracts and transcripts; (2) *transcription noise*—auto-generated subtitles introduce errors degrading text-matching quality; and (3) *corpus heterogeneity*—relevant content spans YouTube videos and MEDLINE abstracts, requiring distinct retrieval strategies per modality.

Approach. We address these challenges through a pipeline combining a domain-specific PubMedBERT bi-encoder fine-tuned with BM25-driven hard negative mining, T5-based query expansion, weighted RRF fusion of dense and sparse retrieval, and a cross-encoder re-ranking stage. PubMed document retrieval uses a multi-pass NCBI API strategy with exponential backoff.

2 Related Work

Dense Retrieval and Bi-Encoders. Dense retrieval using pre-trained language models has substantially advanced open-domain IR. Karpukhin et al. (2020) showed that bi-encoders can match or outperform BM25 for open-domain QA. PubMedBERT (Gu et al., 2021) provides strong initialisation for biomedical dense retrieval. Hard negative mining (Xiong et al., 2021) further improves bi-encoder precision by exposing the model to false-positive candidates.

Hybrid Retrieval and Rank Fusion. Hybrid systems combining BM25 and dense retrievers con-

sistently outperform either alone (Ma et al., 2022). Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) is a robust, score-free fusion method widely adopted for its effectiveness across diverse retrieval settings.

Re-ranking with Cross-Encoders. Cross-encoders jointly encode query-passage pairs and provide higher precision than bi-encoders at the cost of inference speed (Nogueira and Cho, 2019). MS-MARCO-trained MiniLM cross-encoders offer a practical speed-accuracy trade-off.

Medical Video Retrieval. Gupta and Demner-Fushman (2023) introduced MedVidQA—consumer health instructional videos with timestamped QA annotations—and established transcript-based retrieval as the dominant paradigm. Subsequent MedVidQA shared task systems (Gupta et al., 2026) explored BM25 baselines, dense bi-encoders, and cross-modal approaches. Our work extends transcript-based retrieval with domain-specific fine-tuning, multi-stage fusion, and a dedicated document retrieval branch.

3 Data

3.1 Dataset

The corpus is drawn from MedVidQA (Gupta and Demner-Fushman, 2023), consumer health instructional YouTube videos with timestamped QA annotations. Table 1 summarises the statistics.

Split	Samples	Unique Videos
Train	2,710	800
Validation	145	49
Test	155	50

Table 1: MedVidQA dataset statistics. Each sample contains a question and timestamped answer boundaries used to identify positive transcript segments.

3.2 Transcript Corpus Construction

Video transcripts were downloaded using yt-dlp¹ and parsed from VTT format using webvtt-py.² The corpus covers **833 unique videos** with **217,153 raw transcript lines**. For videos without automatic subtitles, OpenAI Whisper ASR was used as a fallback. After temporal chunking (Section 4.1), the corpus yields **32,489 indexed segments**.

¹<https://github.com/yt-dlp/yt-dlp>

²<https://github.com/glut23/webvtt-py>

4 System Description

Our system has four sequential stages: (1) temporal chunking, (2) PubMedBERT bi-encoder fine-tuning, (3) hybrid retrieval with RRF fusion and cross-encoder re-ranking, and (4) PubMed document retrieval. Figure 1 shows the architecture.

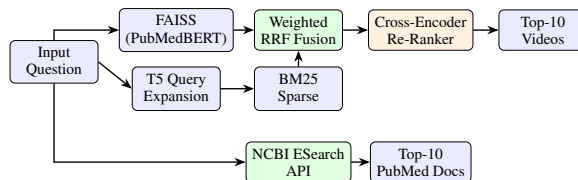


Figure 1: Hybrid retrieval pipeline: dense (FAISS/PubMedBERT) and sparse (BM25/T5-expanded) results fused via weighted RRF, re-ranked by a cross-encoder for videos; NCBI ESearch for documents.

4.1 Temporal Chunking

Given transcript $T = \{(t_i, d_i, s_i)\}_{i=1}^N$ (start time, duration, spoken text), we form overlapping chunks with window $W=30$ s, stride $S=10$ s:

$$C_k = \{s_i \mid (t_i + d_i) > kS \wedge t_i < kS + W\} \quad (1)$$

Each chunk stores its parent video_id for deduplication. This yields **32,489 chunks**, each a 768-dimensional dense vector.

4.2 Bi-Encoder Fine-Tuning

Base model. We build on PubMedBERT (Gu et al., 2021),³ pre-trained on full PubMed abstracts for strong biomedical domain coverage.

Training objective. We fine-tune as a bi-encoder using MultipleNegativesRankingLoss (Reimers and Gurevych, 2019). Positive passages overlap the ground-truth answer span by ≥ 10 s. Hard negatives are the highest BM25-scoring non-overlapping window in the same video, forcing the model to distinguish passages sharing medical terminology but differing in procedural content.

Hyperparameters. Batch size 8, 4 epochs, 100 warmup steps, max sequence length 256 tokens. A TripletEvaluator runs on the validation split twice per epoch.

³huggingface.co/microsoft/BiomedNLP-PubMedBERT

Embedding index. All 32,489 chunks are encoded with mean pooling and L2 normalisation (32,489×768 matrix) and indexed in FAISS IndexFlatIP (Johnson et al., 2019).

4.3 Hybrid Retrieval, Fusion, and Re-Ranking

4.3.1 Query Expansion via T5

Each query is expanded using a T5 doc2query model⁴ generating three passage-style expansions (top- $k=10$ sampling, max length 64):

$$q_{\text{exp}} = q \oplus [\text{T5-expand}(q)]_{1..3} \quad (2)$$

q_{exp} feeds BM25; the original q feeds FAISS to preserve embedding alignment.

4.3.2 Weighted Reciprocal Rank Fusion

FAISS and BM25Okapi each return top $K=200$ chunks. Fusion uses weighted RRF (Cormack et al., 2009):

$$\text{RRF}(c) = \frac{w_d}{k + r_d(c)} + \frac{w_s}{k + r_s(c)} \quad (3)$$

with $k=60$, $w_d=0.75$, $w_s=0.25$.

4.3.3 Cross-Encoder Re-Ranking

Top-200 RRF candidates are re-ranked by a MiniLM cross-encoder⁵ (Nogueira and Cho, 2019). The final output is the top-10 unique video_id values (highest score per video across all chunks).

4.4 PubMed Document Retrieval

We use NCBI ESearch with a sequential fallback chain:

1. Full question text.
2. Top-10 non-stop-word tokens.
3. Top-6 tokens.
4. Top-3 content-bearing tokens.

Failed requests retry with exponential backoff (5 retries, base 0.4 s). PMIDs are scored by rank:

$$\text{score}(r) = 0.5 + 0.49 \cdot e^{-0.3r} \quad (4)$$

mapping rank 0 to ≈ 0.990 and rank 9 to ≈ 0.537 . Results are cached incrementally.

⁴doc2query/msmarco-t5-base-v1

⁵cross-encoder/ms-marco-MiniLM-L-6-v2

5 Experimental Setup

5.1 Datasets and Preprocessing

Training uses the MedVidQA train split (2,710 samples, 800 videos). The validation split (145/49) monitors bi-encoder training only. The test split (155/50) is used exclusively for final evaluation.

5.2 Baselines

We submitted one run (*hybrid_rerank_exp_v3*); no external baseline comparisons are available as no other system runs on this test set are public at time of writing.

5.3 Evaluation Metrics

Standard TREC metrics via pytreceval: MAP, Recall@5/10, Precision@5/10, and NDCG@5/10.

5.4 Implementation Details

Python 3.10, PyTorch, HuggingFace Transformers, Sentence-Transformers, FAISS. BM25 via rank_bm25.⁶ Training on a single consumer GPU; inference on CPU or GPU.

6 Results and Discussion

6.1 Quantitative Results

Metric	Score
MAP	0.4229
Recall@5	0.6903
Recall@10	0.7226
Precision@5	0.1381
Precision@10	0.0723
NDCG@5	0.4857
NDCG@10	0.4971
Topic Coverage	60 / 60
Avg. Docs/Topic	7.6
Avg. Videos/Topic	6.6

Table 2: Evaluation results on the MedGenVidQA 2026 Task A test set (*hybrid_rerank_exp_v3* run).

Our system achieves full 60/60 topic coverage for both retrieval modalities. Recall@10 of 0.7226 indicates our top-10 list covers $\approx 72.3\%$ of all relevant videos per topic. The close agreement between NDCG@5 (0.4857) and NDCG@10 (0.4971) confirms that most relevant videos are ranked in positions 1–5, with only marginal gain from positions 6–10.

The MAP of 0.4229 reflects strong average precision overall. However, 44 of 155 test queries

⁶https://github.com/dorianbrown/rank_bm25

received Recall@10 = 0.0, representing procedure-specific or rare clinical topics for which the transcript corpus contained no closely matching content.

Precision@5 (0.1381) and Precision@10 (0.0723) reveal a recall–precision gap: the system retrieves most relevant videos but does not consistently rank them at the very top positions. We attribute this partly to the cross-encoder being trained on general web passages (MS-MARCO) rather than biomedical text, limiting its ability to make fine-grained relevance distinctions in the medical domain.

6.2 Analysis

Corpus coverage. Our 833-video corpus (32,489 temporal chunks) substantially exceeds the 50 test videos, providing broad candidate coverage and reducing the risk of corpus gaps.

Hard negative mining. BM25-driven hard negatives expose the bi-encoder to passages that share medical terminology with the correct answer but differ in procedural or temporal context—critical for distinguishing adjacent instructional segments within the same recording session.

Query expansion. T5-based expansion bridges colloquial consumer questions (e.g., “*How to fix a crick in my neck?*”) and the clinical register of video transcripts, improving BM25 recall through synonym-level coverage.

PubMed retrieval reliability. The multi-pass fallback strategy was essential: full-question queries frequently failed due to PubMed’s query length constraints and low-frequency medical terminology. Incremental caching ensured 60/60 topic coverage without repeated API calls.

Failure mode analysis. Inspection of the 44 zero-recall queries reveals two dominant failure patterns: (1) *topic specificity*—questions about rare clinical procedures for which the corpus contains only peripherally related videos; and (2) *video duplication*—the same popular medical education video dominating retrieval across multiple topics, reducing per-topic diversity.

6.3 Future Scope

Our current system is purely text-based; several directions offer clear paths to improvement:

- **Multi-modal retrieval.** Incorporating visual keyframes, on-screen text (OCR), and audio prosody features alongside transcript embeddings would allow the system to handle visually informative but verbally sparse videos that currently receive low retrieval scores.
- **Domain-adapted cross-encoder.** Fine-tuning the re-ranker on biomedical QA corpora (e.g., MedVidQA training pairs) rather than relying on the general MS-MARCO model would directly address the observed recall–precision gap and improve precision at early ranks.
- **Local PubMed neural indexing.** Replacing the live NCBI API with a local dense index over PubMed abstracts (e.g., using BioLinkBERT or SPECTER) would enable faster retrieval, richer ranking signals, and rate-limit-free operation.
- **Video metadata integration.** Including YouTube video titles, channel descriptions, and view counts as auxiliary signals could improve retrieval for topics where transcript content alone is ambiguous.
- **Ablation study.** A controlled ablation varying individual components (bi-encoder vs. BM25 only, with/without T5 expansion, with/without cross-encoder) would quantify the contribution of each stage and guide future architecture decisions.

7 Conclusion

We presented the **Varja-Dominators** system for MedGenVidQA 2026 Task A: a fine-tuned PubMedBERT bi-encoder with BM25-driven hard negatives, T5 query expansion, weighted RRF fusion over 32,489 temporal chunks, and cross-encoder re-ranking. PubMed document retrieval uses a multi-pass NCBI API strategy, achieving 60/60 topic coverage. Our system achieves MAP 0.4229, NDCG@10 0.4971, and Recall@10 0.7226. Future work includes multi-modal retrieval with video frames and metadata, local neural PubMed indexing, and domain-adapted cross-encoders for improved precision.

Limitations

This system relies entirely on spoken transcript text; no visual or audio features are used. PubMed

retrieval is live via a public API, introducing rate-limiting latency without neural document ranking. We submitted one run without a controlled ablation study, so individual component contributions are not empirically isolated.

Ethics Statement

All data is publicly available: MedVidQA from the competition organisers and PubMed abstracts via the NCBI open API. No personally identifiable information is collected or processed. This system is for research purposes only and must not substitute for professional medical advice.

Acknowledgments

We thank the MedGenVidQA 2026 organisers for providing the dataset and evaluation infrastructure.

References

- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. ACM.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.
- Deepak Gupta, Collin Scott Campbell, Pedram Golnari, and Dina Demner-Fushman. 2026. Overview of the MedGenVidQA 2026 shared task on medical generative video question answering. In *Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026)*, San Diego, USA. Association for Computational Linguistics.
- Deepak Gupta and Dina Demner-Fushman. 2023. A dataset for medical instructional video retrieval and question answering. In *Proceedings of the 22nd Workshop on Biomedical Language Processing (BioNLP)*, pages 230–240. Association for Computational Linguistics.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.
- Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. 2022. Hybrid retrieval for open-domain question answering over long-form documents. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2740–2745. ACM.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. In *arXiv preprint arXiv:1901.04085*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992. Association for Computational Linguistics.
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jiawei Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive estimation for dense text retrieval. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Pride-Boiler at MedGenVidQA 2026: LLM-Augmented BM25 Retrieval with Corrective Self-Verification for Biomedical Evidence Retrieval

Basil T. Ebinesar, Keyuan Jiang, Charansai Maddineni, Ashok Vardhan Raja
 Purdue University Northwest
 {bebinesa, kjiang, cmaddine, raja22}@pnw.edu

Abstract

This paper describes the Pride-Boiler system submitted to MedGenVidQA 2026 Shared Task A, which asks for retrieving relevant PubMed articles and medical instructional videos in response to consumer health queries. Our approach pairs Pyserini BM25 retrieval with LLM-driven query rewriting and a corrective self-verification loop inspired by the Corrective Retrieval-Augmented Generation (CRAG) paradigm. Given a consumer query, the pipeline first asks Google Gemini to generate clinically optimized search text, one targeting PubMed abstracts with MeSH terms and clinical synonyms, and another targeting video subtitles with procedural action language. BM25 retrieves a broad candidate pool, and Gemini then scores each candidate against the original query, blending its relevance judgment

with the normalized lexical signal. A quality grader assesses the top results: if they are judged insufficient, the pipeline triggers a corrective cycle with reformulated terminology and retries up to three attempts. The entire workflow is orchestrated as a LangGraph state machine. In the official shared task evaluation, Pride-Boiler ranked first among all participating systems on PubMed article retrieval, achieving an nDCG of 0.6532 and MAP of 0.5550, both exceeding the organizer-provided Text-RR baseline. Our performance on video (text) retrieval achieves 0.5304 in MAP and 0.5927 in nDCG, outperforming other systems but falling below that of baseline, indicating the structural limitations of lexical matching over noisy subtitle text. We release the pipeline code to support reproducibility on GitHub at <https://github.com/basilll007/BioNLP>.

PRIDE-BOILER SYSTEM FOR MedGenVidQA 2026 SHARED TASK A: MED & VIDEO RETRIEVAL

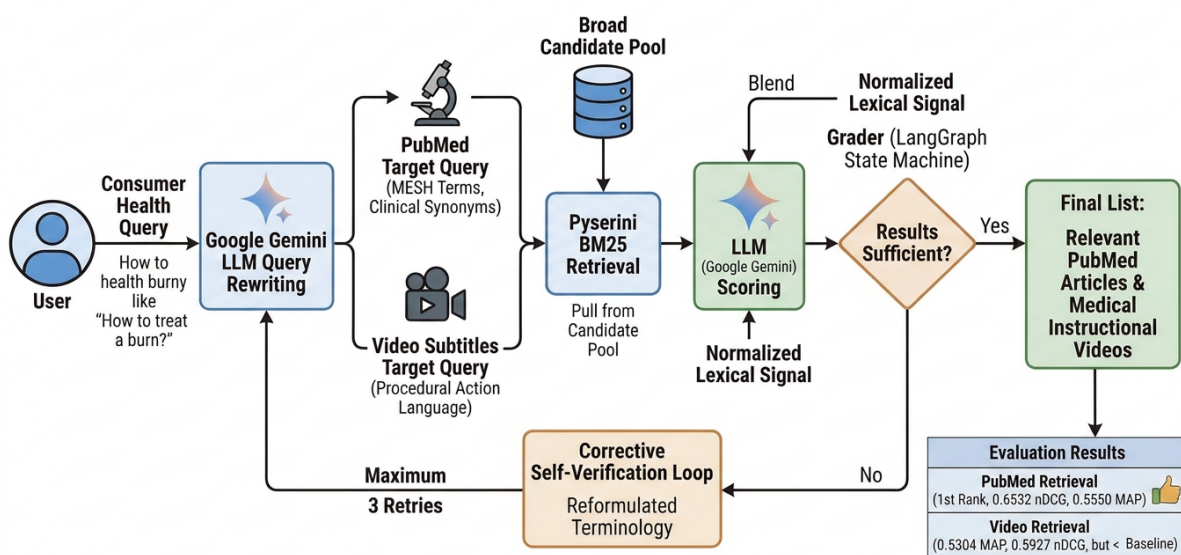


Figure 1: Workflow of the Pride-Boiler’s system. Generated with Google Gemini.

1 Introduction

The volume of biomedical knowledge published each year far exceeds what any individual can absorb. PubMed alone indexes over 36 million abstracts (Lu, 2011), and the number of medical instructional videos hosted on open platforms continues to grow rapidly (Gupta et al., 2023). For clinicians seeking the latest evidence on a treatment protocol, or for patients trying to understand a procedure they have been recommended, the challenge is no longer the access to information, but finding the most relevant pieces of information amid an overwhelming surplus.

The MedGenVidQA 2026 Shared Task A (Gupta et al., 2026) formalizes these challenges into a concrete benchmark. Given a natural-language medical question, participating systems are asked to return the ten most relevant PubMed articles and the ten most relevant medical instructional videos from 2 corpora: one containing over 28 million abstracts and another with 57,166 video entries (Gupta et al., 2025). The task is designed to approximate the real-world scenario in which a user searches for both textual evidence and visual demonstrations of a clinical procedure.

In this paper, we present our Pride-Boiler system submitted to Task A. Rather than training a domain-specific neural retriever which would demand substantial labeled data, embedding infrastructure, and compute, we take a deliberately lightweight approach. We retain Best Match 25 (BM25) as the retrieval backbone, served through Pyserini (Lin et al., 2021), and augment it with a large language model at two critical junctures: before retrieval, where the LLM rewrites each consumer query into clinically precise BM25 search strings with rich MeSH terms and procedural terminology; and after retrieval, where the LLM scores each candidate document against the original query, promoting articles it judges to be genuinely relevant over those that are merely topically adjacent.

The pipeline is further equipped with a corrective self-verification loop inspired by the Corrective Retrieval-Augmented Generation (CRAG) paradigm (Yan et al., 2024). After scoring, a separate LLM call evaluates whether the top results are relevant to the query. If the verdict is insufficient, the system loops back with the grader’s explicit failure reason which is not a generic retry, but a targeted steer toward different terminology and re-executes retrieval from the scratch. The en-

tire workflow is implemented as a four-node Lang-Graph state machine: query transformation, BM25 retrieval, LLM reranking, and quality grading. The corrective edge between the grader and the transformer closes the loop, capped at three attempts to bound latency.

The central question behind this design is whether a capable LLM can compensate for the well-known semantic limitations of BM25 bridging the lay-to-clinical vocabulary gap on the query side, and separating clinically specific evidence from merely related material on the scoring side without requiring dense retrieval, fine-tuned encoders, or domain-specific training data.

2 Related Work

Biomedical IR has long relied on lexical retrieval. BM25 (Robertson and Zaragoza, 2009) remains competitive in domain-specific settings like BioASQ (Tsatsaronis et al., 2015), where clinical terminology creates a vocabulary that dense models must be explicitly trained to handle. Dense passage retrieval (Karpukhin et al., 2020) has narrowed the gap on general benchmarks, but requires substantial labeled data that is scarce in the biomedical setting. Pyserini (Lin et al., 2021) has made BM25 the de facto reproducible baseline in the community. Our system keeps BM25 as the retrieval backbone precisely because it does not require training, and invests the complexity budget elsewhere.

Query reformulation addresses the vocabulary mismatch between consumer questions and clinical literature. Hypothetical Document Embedding (Gao et al., 2023) encodes an LLM-generated answer as a dense query vector; we take a simpler route, prompting the LLM to produce a clinical keyword string directly for BM25, avoiding the embedding infrastructure HyDE requires. Prior neural query rewriting work operates on formal queries; our setting is harder because the input is lay consumer language, not a clinician’s search string.

Retrieval-augmented generation (Lewis et al., 2020) grounds LLM outputs in external evidence. CRAG (Yan et al., 2024) introduced the corrective loop we adapt here: when retrieved results are judged insufficient, the system reformulates and retries. Self-RAG (Asai et al., 2024) pursues a similar self-evaluation idea but interleaves it with text generation. Video retrieval for medical content is a younger problem. MedVidQA (Gupta et al., 2023) introduced the instructional video setting that

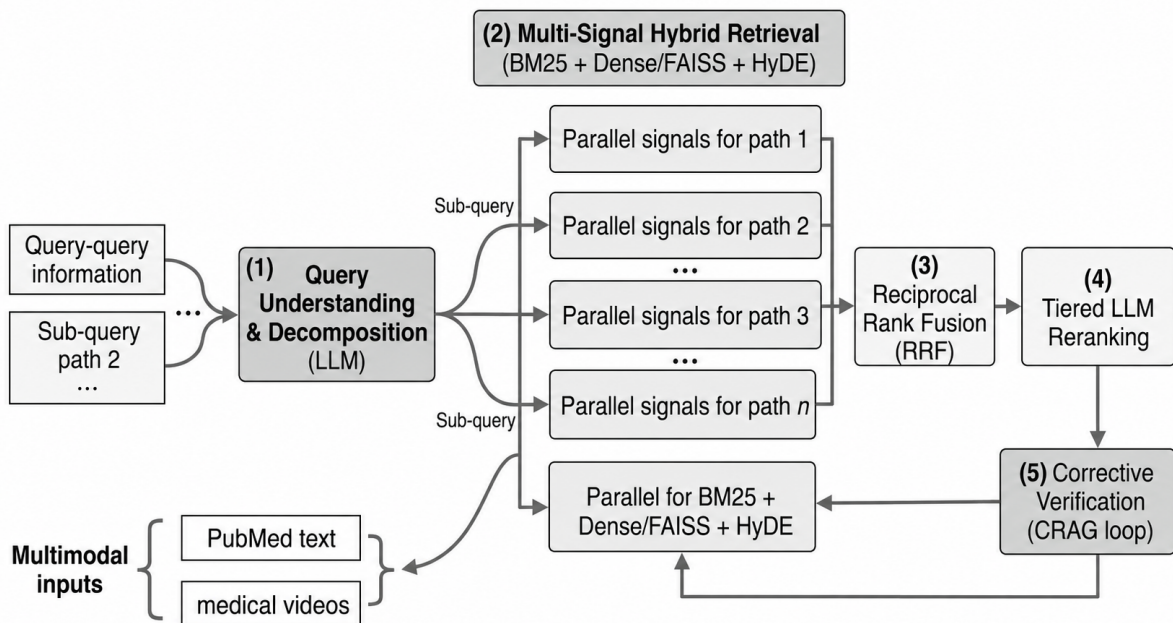


Figure 2: Architecture of the Pride-Boiler retrieval pipeline. A medical query enters the **Query Understanding & Decomposition** node (1), which uses an LLM to rewrite it into BM25-optimized and video-optimized search strings. These are issued to PubMed and video BM25 indexes via **Multi-Signal Retrieval** (2). Candidates are fused (3) and scored by the LLM in a **Tiered Reranking** step (4). A **Corrective Verification** node (5) evaluates result quality: if deemed insufficient, the pipeline loops back to node 1 with a reformulated query, up to three attempts.

the MedGenVidQA 2026 task (Gupta et al., 2026, 2025) extends to joint textual and video retrieval. No prior system has tackled this joint setting; existing work either addresses PubMed retrieval or video understanding in isolation. The subtitle-text challenge we encounter—spoken, fragmented, procedurally oriented—has no direct precedent in prior biomedical IR, which motivates the modality-specific query strategy we develop.

3 System Architecture

Figure 1 illustrates the workflow of the Pride-Boiler system with the functional units and data flow. Figure 2 presents an architectural overview of the Pride-Boiler pipeline, which is implemented as a LangGraph state machine with four processing nodes and a conditional retry edge.

3.1 Node 1: Query Transformation

The raw consumer query is passed to Google Gemini¹ with a prompt that requests two optimized search strings. The first, BM25_QUERY, targets PubMed abstracts and emphasizes MeSH terms, clinical synonyms, anatomical terms, and drug or

¹We used gemini-3-flash-preview throughout the pipeline.

procedure names, restricted to 8–12 keywords with no filler. The second, VISUAL_QUERY, targets medical video subtitles and emphasizes step-by-step procedural language and action-oriented phrases.

On retry attempts triggered when the CRAG grader judges the previous results insufficient, the prompt additionally includes the grader’s failure reason (e.g., “*results focused on pediatric indications rather than adult surgical approaches*”), and instructs the LLM to use completely different medical terminology. This explicit feedback mechanism ensures that successive attempts do not simply repeat the same search query. Appendix A provides a concrete example of the query transformer output and retrieved results for topic A1.

3.2 Node 2: Retrieval

Both optimized queries are issued to their respective indexes across two retrieval signals. For PubMed, we run a BM25 search over a Pyserini Lucene index of 28,372,706 abstracts with $k_1 = 0.9$ and $b = 0.4$, retrieving $k = 1000$ candidates per query.

For video retrieval, we use a dense FAISS index over subtitle embeddings encoded with SapBERT (Liu et al., 2021), a transformer pre-trained

on biomedical entity synonyms via self-alignment with UMLS concepts. Subtitle texts are encoded into 768-dimensional vectors and stored in a FAISS IndexFlatIP for cosine similarity search, retrieving $k = 1000$ candidates per query. We chose SapBERT over a purely lexical approach for video because subtitle text is short, spoken, and procedurally oriented conditions under which BM25 term overlap is unreliable.

In addition to BM25 and dense retrieval, we incorporate a Hypothetical Document Embedding (HyDE) signal (Gao et al., 2023) as a third parallel retrieval path. For each query, Gemini generates a short synthetic PubMed abstract that would constitute an ideal answer, using clinical vocabulary, MeSH terms, and procedure names. With the same SapBERT encoder used for the dense index, this hypothetical abstract is encoded as a dense vector against the FAISS index query. The intuition is that a synthetic clinical document sits closer to the relevant literature in the embedding space than the original consumer query, bridging the lay-to-clinical vocabulary gap at the vector level. The BM25, dense, and HyDE ranked lists are then fused using Reciprocal Rank Fusion (Cormack et al., 2009) before being passed to the reranker.

3.3 Node 3: LLM Reranking

This node is the primary quality filter. The top 100 BM25 candidates are batched into a single prompt with each document’s ID and a 300-character content snippet. The LLM scores each item on a 0.0–1.0 scale against the original query, where 1.0 indicates a document that directly answers the query and 0.0 indicates a completely irrelevant one.

The final score for each candidate is a weighted blend:

$$s = 0.7 \times s_{\text{LLM}} + 0.3 \times \frac{s_{\text{BM25}}}{s_{\text{BM25}}^{\text{max}} + \epsilon}, \quad (1)$$

where $s_{\text{LLM}} \in [0, 1]$ is the relevance score assigned by the LLM, s_{BM25} is the raw BM25 score for the candidate, $s_{\text{BM25}}^{\text{max}}$ is the highest BM25 score in the batch, and $\epsilon = 10^{-9}$ prevents division by zero when all BM25 scores are zero. The division by $s_{\text{BM25}}^{\text{max}}$ normalises lexical scores into $[0, 1]$ so they are directly comparable to s_{LLM} .

The 0.7/0.3 weighting reflects a deliberate design choice: the LLM signal is the primary judge of clinical relevance, while BM25 acts as a tiebreaker among candidates the LLM scores similarly. A pure LLM score would discard the lexical signal

Metric	Docs	Videos
Mean top-1 score	0.926	0.823
Top-1 ≥ 0.90	53/60	43/60
Top-1 ≥ 0.80	58/60	49/60
Perfect top-1 score (1.00)	3/60	3/60
Score drop (rank 1 \rightarrow rank 10)	0.118	0.404

Table 1: Relevance score statistics across 60 submitted topics. “Score drop (rank 1 \rightarrow rank 10)” is the average difference between the relevance score of the top-ranked result and the tenth-ranked result, computed per topic and averaged over all 60 topics. A smaller value indicates that the system returns a uniformly confident shortlist rather than a single strong result followed by weaker candidates.

entirely; a pure BM25 score would undo the semantic reasoning the LLM provides. The 0.7/0.3 split was set empirically on a small held-out sample of 10 queries drawn from the training set, where it produced the highest mean nDCG@10 among the splits $\{0.5/0.5, 0.6/0.4, 0.7/0.3, 0.8/0.2\}$ we tested. Candidates are sorted by s and the top 10 are retained per modality.

3.4 Node 4: Quality Grading (CRAG Loop)

Before finalizing the results, a separate Gemini call inspects the top three documents and top three videos against the original query. The grader assigns one of three labels:

RELEVANT: At least one top result directly addresses the query. The pipeline terminates and returns the results.

INCOMPLETE: Results are partially related but miss key aspects. If the attempt count is below three, control returns to Node 1 with the grader’s reason.

IRRELEVANT: All results are off-topic. Same retry logic applies.

4 Experiments

4.1 Submission Output Quality

The shared task requires exact top ten ranked PubMed documents and top ten ranked videos per query, each accompanied by a relevance score in $[0, 1]$. Table 1 summarizes the score distributions across all 60 evaluation topics.

The document retrieval performance is strong: the system places a high-confidence article at the

Rank	Doc	Video
1	0.926	0.823
2	0.891	0.700
3	0.869	0.619
4	0.854	0.575
5	0.842	0.549
6	0.834	0.496
7	0.825	0.470
8	0.821	0.441
9	0.814	0.431
10	0.808	0.420

Table 2: Mean relevance scores by rank position, averaged over 60 topics. Document scores degrade gradually; video scores decline more noticeably after Rank 3.

top for 88.3% of queries, and the gentle score decline across ranks (mean score drop 0.118) suggests that the reranker is producing a coherent, well-ordered shortlist rather than a noisy grab-bag. Three topics achieved a perfect document relevance score of 1.0 at rank 1 (A34, A36, A37), all involving highly specific procedural queries where the BM25 query transformation precisely matched the target literature. A separate set of three topics achieved a perfect video relevance score of 1.0 at rank 1; notably, only topic A37 appears in both groups, suggesting that strong document retrieval and strong video retrieval are largely independent outcomes driven by different corpus characteristics.

4.2 Score Progression Across Ranks

Table 2 reports the mean relevance scores at each output position, averaged over all 60 topics.

Document scores remain above 0.80 even at rank 10, indicating that the entire returned list tends to be relevant, not just the top few items. Video scores tell a different story: they drop from 0.823 at rank 1 to 0.420 at rank 10, with a particularly steep cliff between ranks 3 and 6. This is consistent with the nature of the video corpus many medical topics simply do not have ten highly relevant instructional clips available, particularly for specialized procedural queries like “*How do you perform endoscopic transsphenoidal salvage surgery?*” (topic A37). In those cases, the system fills the lower ranks with the best available candidates, which are only tangentially related.

4.3 The CRAG Loop in Practice

The corrective loop fires when the quality grader judges the initial retrieval insufficient. Across our 60-topic run, the majority of queries achieved RELEVANT status within one or two attempts. The retry

mechanism proved most valuable for queries with ambiguous lay terminology, where the first BM25 query missed the right clinical angle. For example, a query about “*crowded arteries*” (topic A20) initially retrieved literature on arterial crowding in dental contexts; the grader flagged this, and the reformulated query correctly targeted coronary artery disease and atherosclerotic stenosis.

5 Implementation Details

The entire pipeline runs on a single RunPod instance equipped with an NVIDIA RTX 6000 GPU (98 GB VRAM), 188 GB system RAM, and 16 vCPUs, running Ubuntu 24.04 with Python 3.10 (RunPod PyTorch image 1.0.2-cu128-torch280). The PubMed BM25 index occupies roughly 15 GB on disk and the video index under 400 MB. All LLM calls use Gemini 3 Flash Preview via the Google GenAI API. BM25 search is handled by Pyserini 1.2.0 with OpenJDK 21. The pipeline is orchestrated with LangGraph, which manages the state machine transitions and the conditional retry edge.

Per-query wall-clock time is dominated by the LLM calls, and each query transformation takes roughly 1–2 seconds. BM25 retrieval responds in less than a second, and a batch reranking call finishes in 5–15 seconds depending on the candidate count. A single pass through the pipeline completes in approximately 10–20 seconds; queries that trigger the CRAG retry loop take proportionally longer, up to about a minute for the maximum three attempts.

6 Results and Discussion

The official evaluation results are summarized in Table 3.

On PubMed retrieval, our Pride-Boiler system ranked first among all participating systems, achieving an nDCG of 0.6532 and MAP of 0.5550, both above the Text-RR baseline. This is notable given that the less powerful machinery the system uses: no dense index, no fine-tuned encoder, no learned reranker. The result validates the core hypothesis behind our design that a capable LLM, applied at the query reformulation and candidate scoring stages, can extract substantial relevant documents from a classical BM25 backbone.

Video retrieval tells a different story. Our nDCG of 0.5927 falls below the Video-RR baseline’s 0.6616, and the MAP gap is similarly unfavorable.

	MAP	R@5	R@10	P@5	P@10	nDCG
<i>PubMed Article Retrieval</i>						
Baseline	0.5404	0.5505	0.5863	0.5133	0.2700	0.6460
Ours	0.5550	0.5571	0.5866	0.5333	0.2817	0.6532
<i>Instructional Video Retrieval</i>						
Baseline	0.5884	0.6067	0.6528	0.4100	0.2217	0.6616
Ours	0.5304	0.5478	0.5833	0.3900	0.2100	0.5927

Table 3: Performance of our system in comparison with the Baseline results. Boldfaced numbers indicate the higher values within each modality.

This is the clearest evidence that neither lexical nor entity-centric dense retrieval fully addresses the subtitle matching problem. BM25 over subtitle text has a structural ceiling that query reformulation alone cannot raise: subtitles are short, fragmented, often auto-transcribed, and written in spoken register, so even a well-crafted procedural query can only match videos that happen to contain the exact terms. The dense index, encoded with SapBERT (Liu et al., 2021), does not resolve this. SapBERT was pre-trained on biomedical entity synonyms via UMLS concept alignment, making it strong at matching clinical terminology across surface forms for example, recognising that *myocardial infarction* and *heart attack* refer to the same concept. Subtitle text, however, is not a collection of biomedical entities; it is procedural narration (“*now we retract the flap and identify the vessel*”), and SapBERT has no training signal for this register. The result is that both retrieval signals fail for the same underlying reason: the query and the subtitle live in different linguistic worlds, and neither BM25 term overlap nor entity-level embedding alignment can bridge that gap without a model trained specifically on procedural spoken language. The pattern across modalities is instructive. PubMed abstracts are long, terminologically dense, and written in standardized clinical prose exactly the kind of text BM25 handles well once the right query terms are in place. Video subtitles violate nearly every assumption BM25 makes about document structure. The performance gap is not a failure of the pipeline design but a clean signal about where denser retrieval signals are needed.

The CRAG corrective loop proved most valuable on queries with ambiguous lay terminology, where the first reformulation landed in the wrong clinical neighborhood. The grader’s failure reason gave the transformer a concrete steer on retry, not just “try again”, but “the results discussed X when they should address Y.” That said, the loop has a hard

ceiling: when ground-truth documents never enter BM25’s candidate pool, no amount of reformulation will surface them.

7 Error Analysis

Across 60 topics, the system places a highly relevant document (grade 2) at rank 1 for 38 queries and a partially relevant document for a further 15. The remaining 7 topics return an irrelevant result at rank 1, and 4 topics A15, A44, A49, and A56 return zero relevant documents in the entire top-10 (*FP2: missed top-ranked documents* (Barnett et al., 2024)). Inspection of these four queries reveals a consistent pattern: the query transformer confidently reformulated toward the broader condition using incorrect clinical terminology, steering BM25 toward adjacent literature rather than the target. Topic A44, for example, uses lay phrasing that maps to a narrow clinical subterm; the query transformer confidently reformulated toward the broader condition, and the relevant abstracts which use the specific subterm never entered the candidate pool. No amount of CRAG retry can recover from this: once the wrong clinical neighborhood is established in the first reformulation, subsequent attempts explore nearby but equally incorrect regions. A second failure mode affects 7 topics (A10, A16, A17, A19, A25, A51, A57), which retrieve some relevant documents but miss more than half the available relevant set. These are multi-faceted queries with both pharmacological and procedural dimensions. The LLM transformer consistently resolves the ambiguity by committing to one clinical angle usually the more salient one while the documents covering the secondary angle rank below the top-10 cutoff. Recall here is structurally capped by query specificity. The most pervasive issue cuts across all categories: on average, 6.77 of the 10 retrieved documents per topic are unjudged meaning they appear nowhere in the relevance pool assembled by the task organizers.

These are not necessarily irrelevant; they may be genuinely useful documents that assessors never evaluated. Under standard TREC scoring, however, unjudged documents count as non-relevant, which suppresses all precision metrics. This gap reflects a fundamental tension between a system that searches a 28-million-document corpus freely and an annotation pool necessarily bounded by human annotator effort.

Video retrieval fails more severely: 13 of 60 topics return zero relevant videos, and the top-ranked result is irrelevant in 18 topics nearly three times the document rate. The root cause is a linguistic register mismatch that query reformulation cannot bridge. BM25 is built on term overlap. PubMed abstracts and the clinical queries we generate share vocabulary by design. Subtitle text does not. A reformulated query like *uvulopalatopharyngoplasty obstructive sleep apnea surgical outcomes* finds nothing useful in a subtitle corpus where the same procedure is described as *removing tissue from the back of the throat to widen the airway*. The query transformer generates clinical prose; subtitles speak in lay procedural language. The gap is not a vocabulary mismatch in the abstract sense it is a mismatch in the very register of the text. A secondary cause is corpus sparsity. For specialized procedural queries robotic surgery variants, rare metabolic conditions, niche anatomical procedures fewer than ten relevant instructional videos may exist in the entire 57,166-entry corpus. The system correctly identifies the available relevant videos at top ranks for these queries, but must fill lower positions with tangentially related content. This is not a retrieval failure; it is a coverage limitation of the corpus that no retrieval algorithm can overcome. Finally, the LLM reranker operates on 300-character subtitle snippets. For PubMed abstracts, 300 characters reliably captures the core claim of a structured abstract. For subtitles, 300 characters may span a single procedural instruction *“now we retract the flap and identify the vessel”* with no indication of the broader procedure being demonstrated. The reranker has insufficient context to judge video relevance accurately, and in 18 topics it ranks an irrelevant video first as a result.

8 Conclusion

We presented Pride-Boiler, a retrieval system for MedGenVidQA 2026 Shared Task A that demonstrates how the performance of a classical BM25

backbone can improve when augmented with LLM-driven query transformation and corrective self-verification. The system is deliberately simple, with four nodes in a LangGraph state machine, no fine-tuned encoders, no learned reranker, and dense retrieval limited to video subtitle embeddings, and yet it ranked first among all participating teams on textual retrieval. The architecture is modular by design: each node can be replaced or extended independently, and the corrective loop generalizes beyond the specific retrieval backend. We view this submission less as a finished system and more as a baseline that establishes what lightweight LLM augmentation can achieve on its own, so that future work on denser retrieval components can be measured against a clear reference point.

Looking ahead, the most immediate priority is stronger retrieval for the video modality, where BM25 over subtitle text has a structural ceiling that query reformulation alone cannot raise. We plan a comparative evaluation of three biomedical encoders MedCPT (Jin et al., 2023), BMRetriever (Xu et al., 2024), and BiCA (Sinha et al., 2026) fused with BM25 via Reciprocal Rank Fusion (Cormack et al., 2009) to recover ground-truth documents that lexical retrieval misses entirely. Three further extensions follow naturally from the error analysis: replacing the 300-character snippet reranker with a cross-encoder over full text, decomposing compound queries into modality-specific sub-queries, and replacing the CRAG grader’s three-way categorical label with a continuous confidence score for more calibrated retry decisions.

Limitations

Three limitations bear directly on the results reported here. First, For PubMed retrieval, the system relies entirely on BM25 for its initial candidate pool. When ground-truth documents use vocabulary that diverges from even a well-crafted clinical query—as in the four complete-miss topics no reranking or CRAG retry can surface them. Dense retrieval over document embeddings would mitigate this by matching semantic intent rather than surface terms, but we deliberately excluded it to test how far a lexical system could go when paired with a capable LLM.

Second, the single-query strategy applies the same reformulated clinical string to both PubMed and video subtitle retrieval. PubMed abstracts reward clinical precision. Subtitle text rewards

procedural, action-oriented lay language. Using one query for both modalities guarantees that one of them is systematically mismatched. A modality-specific query generator producing a separate action-oriented search text for video would directly address the 13 complete-miss topics in the video track.

Third, the LLM reranker scores candidates from a 300-character snippet. For PubMed abstracts this is usually sufficient; the opening sentences of a structured abstract convey the central claim. For subtitle chunks it is rarely sufficient: 300 characters of spoken procedural narration lack the context to determine what intervention the video is actually demonstrating. A cross-encoder operating over full transcript text, or over title and metadata fields, would give the reranker the context it currently lacks.

Ethics Statement

This system retrieves published biomedical literature and educational instructional videos for informational purposes. It is not intended as a clinical decision-support tool and should not be used to make medical decisions without professional guidance. All corpora are publicly available. LLM-generated content (query rewrites, relevance scores, quality judgments) is used exclusively to guide retrieval; no LLM output is presented to users as clinical evidence.

Acknowledgments

We thank Dr. Deepak Gupta and Dr. Dina Demner-Fushman from the National Library of Medicine for organizing the MedGenVidQA 2026 Shared Task and for providing the evaluation infrastructure and datasets. We are especially grateful to Dr. Gupta for his prompt and generous responses to our questions throughout the manuscript preparation process. We also thank the developers of Pyserini and LangGraph for the open-source tooling that underpins this system.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the International Conference on Learning Representations*.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024.

Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN 2024)*, Lisbon, Portugal.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1762–1777.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2023. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2025. A dataset of medical questions paired with automatically generated answers and evidence-supported references. *Scientific Data*, 12(1):1035.

Deepak Gupta, Collin Scott Campbell, Pedram Golnari, and Dina Demner-Fushman. 2026. Overview of the MedGenVidQA 2026 shared task on medical generative video question answering. In *Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026)*, San Diego, USA. Association for Computational Linguistics.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2023. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16649–16664. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. *SIGIR*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. [Self-alignment pretraining for biomedical entity representations](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4228–4238.

Zhiyong Lu. 2011. [PubMed and beyond: A survey of web tools for searching biomedical literature](#). *Database*, 2011:baq036.

Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: BM25 and beyond](#). *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Aarush Sinha, Pavan Kumar S, Roshan Balaji, and Nivraj Pravinbhai Bhatt. 2026. [BiCA: Effective biomedical dense retrieval with citation-aware hard negatives](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33010–33018.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, and 1 others. 2015. [An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition](#). *BMC Bioinformatics*, 16(1):138.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May Dongmei Wang, Joyce C. Ho, Chao Zhang, and Carl Yang. 2024. [BMRetriever: Tuning large language models as better biomedical text retrievers](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22234–22254, Miami, Florida, USA. Association for Computational Linguistics.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. [Corrective retrieval augmented generation](#). *Computing Research Repository*, arXiv:2401.15884.

A An Example Query Decomposition

For topic A1 “*What surgical approaches are effective for reducing symptoms in adult patients with obstructive sleep apnea? How are they performed?*” the query transformer produced the following optimized search strings:

```
{
  "original_query": "What surgical approaches
    are effective
    for reducing symptoms in adult patients with
    obstructive
    sleep apnea? How are they performed?",
  "BM25_QUERY": "uvulopalatopharyngoplasty
    maxillomandibular
    advancement hypoglossal nerve stimulation
    surgical
    treatment obstructive sleep apnea adults
    outcomes",
  "VISUAL_QUERY": "surgical procedure sleep
    apnea palate
    tissue removal airway widening step-by-step
    operating
    room technique demonstration"
}
```

The BM25_QUERY correctly introduced clinical procedure names (uvulopalatopharyngoplasty, maxillomandibular advancement, hypoglossal nerve stimulation) absent from the original consumer question this reformulation is what enables BM25 to bridge the vocabulary gap. The pipeline returned the following top-3 results per modality for this topic:

```
{
  "query_id": "A1",
  "pubmed_results": [
    {"id": "33418179", "score": 0.9307},
    {"id": "20451028", "score": 0.9255},
    {"id": "19784401", "score": 0.9199}
  ],
  "video_results": [
    {"id": "PMC3197034_jove-52-2652.mp4", "score": 0.9170},
    {"id": "PMC5755377_jove-129-56369.mp4", "score": 0.8566},
    {"id": "PMC2841570_jove-35-1437.mp4", "score": 0.8339}
  ]
}
```

The sharp score drop after rank 3 in the video results is consistent with the corpus sparsity discussed in the error analysis: highly relevant surgical demonstration videos for obstructive sleep apnea are sparse in the 57,166-entry corpus.

Seahawk at MedGenVidQA 2026: LLM Segment-Range Selection for Medical Visual Answer Localization

Xiaotian Tian

NKU School of Mathematical Sciences University of North Carolina Wilmington
Tianjin, China
13707579071@163.com

Gulustan Dogan

University of North Carolina Wilmington
Wilmington, USA
dogang@uncw.edu

Abstract

Medical visual answer localization requires identifying the temporal span in a video where a medical question is answered or visually explained. We present a simple retrieval-and-selection pipeline for Task C that treats visual answer localization as segment-level answer paragraph selection over timestamped video transcripts. Given a question and a segmented transcript, our system prompts DeepSeek to select a contiguous range of transcript segments rather than directly generating timestamps. The final start and end times are then computed deterministically from the selected segment boundaries, decreasing the risk of hallucinated or malformed temporal outputs. To support long videos, we apply overlapping sliding-window prompting and rank candidate ranges using lexical question. In a 20-sample sanity check on test dataset, a completeness-biased configuration achieved an mIoU of 0.3217, while a shorter duration-penalized configuration improved performance to 0.4815. These results suggest that constrained LLM-based segment selection, combined with deterministic timestamp extraction, is a practical baseline for medical visual answer localization.

1 Introduction

Instructional medical videos contain procedural and explanatory information that is often difficult to convey using text alone. For consumer health questions, first-aid procedures, and medical education scenarios, the answer may correspond not to an entire video but to a short temporal interval in which a specific action, tool, or explanation is shown. Medical visual answer localization therefore requires a system to identify the relevant start and end timestamps for a natural-language medical query.

The MedGenVidQA shared tasks formalize this problem as medical video question answering with temporal grounding. In visual answer localization,

a system is given a medical query and a video and must return the temporal segment where the answer is shown or explained (Gupta and Demner-Fushman, 2022; Gupta et al., 2023, 2026). The task is challenging because medical videos can be long, contain multiple related procedural steps, and include terminology that may be paraphrased or only indirectly expressed in the transcript.

Our system is designed around three principles. First, transcript segments should preserve their original timestamp boundaries. Second, the LLM should make a semantic localization decision but should not invent timestamps. Third, candidate answer spans should be biased toward concise intervals because overly long predictions reduce temporal Intersection-over-Union.

The main contributions of this paper are:

- a constrained LLM prompting strategy that predicts only start and end segment identifiers
- a sliding-window search procedure for long timestamped transcripts
- a deterministic timestamp extraction step that converts selected segment ranges into valid Task-C outputs
- a duration-penalized candidate scoring function that improves mIoU in our sanity-check evaluation

2 Background and Related Work

Medical video question answering extends traditional question answering by requiring systems to reason over visual and temporal evidence. The original MedVidQA shared task introduced medical video understanding tasks including visual answer localization, where the goal is to identify the start and end timestamps of the relevant answer segment (Gupta and Demner-Fushman, 2022). The MedVidQA dataset contains human-annotated

medical instructional questions paired with visual answer timestamps (Gupta et al., 2023). Later work expanded the setting with additional medical video resources and monomodal and multimodal approaches for health-related question answering from videos (Gupta et al., 2024).

The current MedGenVidQA Task C setting focuses on visual answer localization. Given a medical query and a video, the system must locate the temporal region where the answer is shown or explained (Gupta et al., 2026).

Our method is transcript-centered rather than fully multimodal. Instead of learning a video-text model, we exploit timestamped transcript segments as the grounding substrate. This makes the method lightweight and submission-oriented: the LLM performs semantic selection over transcript text, while the system preserves temporal reliability through deterministic post-processing.

3 Task Definition

We focus on Task C, Visual Answer Localization. Given a medical question q and a video v , the system must predict a temporal interval $[p_s, p_e]$ indicating where the answer to q is shown or explained. In our setting, each video is represented by a timestamped transcript containing ordered segments $S = \{s_1, \dots, s_n\}$, where each segment s_i has text, a start time, and an end time.

The output for each test instance consists of an identifier, an answer start time, and an answer end time. The system is evaluated using temporal overlap between the predicted interval and the ground-truth interval. Following the shared-task setup, we report mean Intersection-over-Union (mIoU) and use the same interval-overlap formulation in our internal sanity-check experiments.

4 System Overview

Our system decomposes temporal localization into two steps: semantic segment selection and deterministic timestamp extraction. The LLM is responsible only for identifying which transcript segments contain the answer. It is not asked to generate timestamps directly. Once a segment range is selected, the predicted start time is taken from the first selected segment and the predicted end time is taken from the last selected segment.

This design reduces the output space of the LLM from arbitrary temporal strings to two integer segment identifiers. It also makes the system easier to

debug: every predicted timestamp can be traced back to the transcript segment that produced it. For long transcripts, we apply the same segment-selection prompt over overlapping windows and then score all non-null candidate ranges before selecting the final prediction.

5 Experimental Setup

5.1 Data and inputs

We use timestamped video transcripts as the primary input representation. Each transcript is converted into a list of numbered segments, where each segment contains a segment identifier, start time, end time, and transcript text. This format allows the LLM to inspect both the semantic content and the temporal boundaries of each candidate answer span.

5.2 Model configuration

We use DeepSeek as the segment-range selector. The model is prompted with the question and a window of timestamped transcript segments, and it is instructed to output only two fields: the start segment identifier and the end segment identifier. All outputs are parsed with a deterministic post-processing script. Invalid or null outputs are discarded unless no valid candidate is available.

5.3 Candidate selection

For transcripts that exceed the context budget, we divide the segment list into overlapping windows. Each window produces zero or one candidate answer range. We then score candidates using lexical overlap between the question and the selected paragraph, minus a penalty proportional to the predicted duration. This scoring function favors spans that are both semantically relevant and temporally concise.

5.4 Evaluation protocol

We first evaluated the pipeline on the first 20 examples from MedVidQA/test.json as a sanity check. We compared a completeness-biased configuration, which allowed expanded ranges, with a shorter duration-penalized configuration. The final Task-C submission was generated using the tuned short-span configuration and formatted as task_c_submission.json.

6 Methodology

Our approach is a retrieval-and-selection temporal grounding pipeline using DeepSeek as a segment-range selector, coupled with deterministic timestamp extraction. The design goal is to let the model decide *where* the answer is, but never let the model invent timestamps.

6.1 Segment-based context representation

Given a transcription with segments $S = \{s_i\}$, we build a prompt context of the form:

```
[segment_id] (start_time - end_time)
segment_text
```

This representation provides the model both textual content and precise segment boundaries while keeping the output interface simple.

6.2 LLM as answer paragraph locator (segment-range prediction)

For a question q , we prompt DeepSeek to output only two fields:

```
START_SEGMENT_ID: <int or null>
END_SEGMENT_ID: <int or null>
```

The model is instructed to select a contiguous segment range that corresponds to the answer paragraph. This design avoids brittle JSON generation and mitigates long-form reasoning outputs.

6.3 Long transcript handling via sliding-window search

For long transcripts, we apply a sliding window strategy: we split segments into windows of size W with overlap O , query the model per window, and convert window-local selections to global segment ids. This approximates searching the whole transcript while keeping each LLM call bounded.

6.4 Candidate scoring with duration penalty

When multiple windows yield non-null ranges, we select the best candidate using:

$$\text{Score} = \text{Overlap}(q, \text{paragraph}) - \lambda \cdot \text{Duration} \quad (1)$$

where $\text{Overlap}(\cdot)$ counts lexical overlap between question tokens and the concatenated paragraph tokens, $\text{Duration} = \text{end_time} - \text{start_time}$, and λ is a tunable length penalty to discourage overly long spans.

6.5 Deterministic timestamp extraction and formatting

Once a segment range $[a, b]$ is selected, timestamps are computed deterministically:

$$\begin{aligned} \text{pred_start} &= \text{start_time}(s_a), \\ \text{pred_end} &= \text{end_time}(s_b) \end{aligned} \quad (2)$$

For Task-C submission formatting, seconds are converted to MM:SS; the start uses floor and the end uses ceil to ensure coverage of the answer.

6.6 Output generation

For Task-C we output a submission file `task_c_submission.json` with entries `id`, `answer_start`, `answer_end`. For analysis, we retain an intermediate file `task_c_answer_paragraphs.json` including selected segment ranges, paragraph text, and raw seconds.

7 Results

7.1 Evaluation Metric

We report temporal mean Intersection-over-Union (mIoU) over predicted and ground-truth answer intervals:

$$\text{IoU} = \frac{|[p_s, p_e] \cap [g_s, g_e]|}{|[p_s, p_e] \cup [g_s, g_e]|}, \quad (3a)$$

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^N \text{IoU}_i \quad (3b)$$

Following the shared-task evaluation protocol, we additionally report the proportion of predictions achieving IoU above fixed thresholds (IoU@0.3, IoU@0.5, IoU@0.7), which measure fine-grained temporal localization accuracy.

7.2 Sanity-Check Results (First 20 Samples)

We first evaluated the pipeline on the first 20 examples from `MedVidQA/test.json` as a sanity check. An initial completeness-biased configuration (with expanded ranges) achieved $\text{mIoU} = 0.3217$. After introducing a duration penalty and disabling expansion (`expand_before = 0`, `expand_after = 0`), performance improved substantially; one setting achieved $\text{mIoU} = 0.4815$ with `length_penalty = 0.06`. A similar short+penalty regime yielded $\text{mIoU} = 0.4689$. These sanity checks confirmed that penalizing overly long spans is beneficial for temporal localization.

7.3 Full Test Set Results (MedGenVidQA Task C)

We ran the complete Task C test set (80 instances, designated C1–C80) using the tuned short+penalty configuration. Table 1 summarizes the final performance.

Table 1: Performance on the full MedGenVidQA 2026 Task C test set.

Metric	Value
mIoU (%)	52.31
IoU@0.3 (%)	71.25
IoU@0.5 (%)	52.50
IoU@0.7 (%)	42.50

The mIoU of 52.31% indicates that, on average, the predicted temporal segment overlaps with the ground truth by more than half of their union. The high IoU@0.3 (71.25%) shows that the majority of predictions have at least a modest overlap with the correct interval, while the gradual drop at stricter thresholds (IoU@0.5 and IoU@0.7) reflects the inherent difficulty of exact boundary alignment in medical procedural videos. These results demonstrate that constrained LLM-based segment selection, combined with deterministic timestamp extraction, provides a practical and competitive baseline for medical visual answer localization, even without any training or fine-tuning of the underlying model.

7.4 Qualitative Behavior

In typical success cases, the model selects a segment range corresponding to the procedural explanation rather than surrounding narration. Even when a tool name in the question is absent verbatim in the transcript, the model can still focus on segments describing the corresponding instrument or action. Failure cases often involve videos with very long transcripts where the answer spans across multiple sliding windows, or where ASR errors introduce misleading or missing text. The duration penalty effectively suppresses excessively long predictions, but occasionally over-penalizes correctly long answer spans, leading to underestimated boundaries.

7.5 Task-C Output Generation

Using the tuned short+penalty configuration, we ran the full Task-C set (C1–C80) and produced a submission file `task_c_submission.json` in the

		IoU@0.3	IoU@0.5	IoU@0.7	Mean IoU
1	LAMAR-2	93.75	90.00	77.50	79.55
2	NJUST-KMG	92.50	81.25	67.50	75.48
3	405621	60.00	55.00	47.50	50.78
4	TXT66 (Seahawk)	71.25	52.50	42.50	52.30
5	UNCC	62.50	36.25	22.50	42.57
6	ADAPT	10.00	10.00	8.75	8.62

Table 2: Task C validation results for submitted systems. Scores are shown at IoU thresholds 0.3, 0.5, and 0.7, together with mean IoU. Systems are ordered by IoU@0.7.

required format. The results reported in Table 1 are computed directly from that submission against the held-out ground truth.

8 Conclusion

We presented a practical pipeline for LLM-assisted temporal answer grounding in medical procedural videos using segment-aligned transcriptions. Restricting the LLM to selecting segment ranges rather than generating timestamps enables deterministic and robust timestamp extraction. Sliding-window prompting scales the approach to long transcripts, and duration penalties significantly improve mIoU by discouraging overly long spans. Future improvements include higher-quality ASR, boundary refinement, and embedding-based candidate scoring for robustness to paraphrase.

9 Limitations

Despite the promising results, our approach has several limitations. First, the system relies entirely on timestamped transcripts derived from automatic speech recognition (ASR); errors in transcription or timing can directly degrade localization performance. Second, the current method is text-only and does not use visual information from the video frames, which may be essential for recognizing actions, instruments, or visual cues that are not well described in the transcript. Third, the sliding-window strategy approximates full-video reasoning but may miss answer spans that cross window boundaries or fall into low-overlap regions. Fourth, the duration penalty and other hyperparameters (e.g., window size, overlap, λ) were tuned on a small sanity-check set; their generalizability to the full test set is not guaranteed. Fifth, the segment-range selection prompt assumes that the answer corresponds to a contiguous block of transcript segments, which may not hold for videos where the

answer is interleaved with irrelevant narration. Finally, invalid or null LLM outputs are still possible, and we fill missing predictions with placeholders, which limits the system’s robustness in production settings. Future work should address these issues by incorporating multimodal cues, refining boundary detection, and developing more robust prompting or fine-tuning strategies.

References

- Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2023. [A dataset for medical instructional video classification and question answering](#). *Scientific Data*, 10(1):158.
- Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2024. [Towards answering health-related questions from medical videos: Datasets and approaches](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16399–16411, Torino, Italia. ELRA and ICCL.
- Deepak Gupta, Collin Scott Campbell, Pedram Golnari, and Dina Demner-Fushman. 2026. [Overview of the medgenvidqa 2026 shared task on medical generative video question answering](#). In *Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026)*, San Diego, USA. Association for Computational Linguistics.
- Deepak Gupta and Dina Demner-Fushman. 2022. [Overview of the MedVidQA 2022 shared task on medical video question-answering](#). In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 264–274, Dublin, Ireland. Association for Computational Linguistics.

UNCC at MedGenVidQA 2026: Structured Temporal Grounding for Medical Video Question Answering

Hilmi Demirhan

Congdon School of Supply Chain, Business
Analytics and Information Systems,
University of North Carolina Wilmington,
Wilmington, NC, US

Wlodek Zadrozny

Department of Computer Science,
University of North Carolina Charlotte,
Charlotte, NC, USA

Abstract

MedGenVidQA 2026 Task C evaluates visual answer localization in medical videos. The system receives a video and a question, then returns the start and end time of the visual answer. Our framework used timestamped automatic speech recognition (ASR) as a proposal source rather than as a final boundary label. The framework generated transcript tables, phase maps, lexical and dense candidate windows, schema-constrained ranking inputs, selective key-frame checks, and a deterministic validation pass for the final JSON file. The ranker selected among bounded candidate intervals instead of generating arbitrary timestamps over a full transcript. Each output can be traced to segment identifiers, candidate source families, selected anchors, phase labels, and validation flags. Our best run ranked fifth among six participant systems, with 62.50 IoU@0.3, 36.25 IoU@0.5, 22.50 IoU@0.7, and 42.57 mIoU. The threshold pattern suggests that coarse temporal retrieval was more reliable than strict start-end localization.

1 Introduction

MedGenVidQA 2026 includes retrieval, answer generation, and visual answer localization tasks for medical video question answering (Gupta et al., 2026, 2023, 2024, 2025). This paper describes the UNCC submission for Task C. Given a video-question pair, the system predicts one interval $[t_s, t_e]$ that contains visual evidence for the answer. The official evaluation reports mean Intersection over Union (mIoU) and thresholded overlap at IoU@0.3, IoU@0.5, and IoU@0.7 (Gupta et al., 2026; Gupta and Demner-Fushman, 2022). IoU@0.3 gives credit for reaching the correct temporal neighborhood. IoU@0.7 requires much tighter start and end boundaries.

Task C requires temporal grounding beyond transcript-level lexical matching. ASR alignment and visual evidence do not always refer to the same

temporal boundary. A transcript row may contain a relevant answer term while the corresponding visual action occurs earlier or later in the video. Spoken instruction can introduce a step before it is performed, continue after the action is completed, or reuse similar wording across adjacent demonstrations. The UNCC framework used ASR timestamps as proposal cues rather than fixed temporal boundaries. It expanded transcript matches into candidate intervals and checked the candidates with phase labels, local transcript context, duration guards, candidate-source agreement, and selective key-frame checks.

The framework preserves intermediate artifacts to support traceable timestamp selection. It stores transcript tables, phase-map rows, candidate rows, selected anchors, ranker outputs, validation records, and the final submission JSON. Model calls receive a bounded candidate list with local evidence snippets. They do not search the entire transcript for an unconstrained timestamp pair. This design allows prediction errors to be mapped to ASR segmentation, proposal recall, candidate ranking, boundary adjustment, or JSON export.

The submitted system did not rely on task-specific fine-tuning or a learned dense video encoder. This design kept the shared-task run simple and auditable, but it also limited boundary precision. Transcript and phase evidence can organize the temporal search space for coarse overlap. The system remains weaker when the answer depends on a small visual indicator that is not tightly synchronized with the spoken content.

2 Related Work

Medical video question answering is part of multimodal medical question answering, where systems answer medical questions from textual and visual evidence sources (Demirhan and Zadrozny, 2023). The video setting includes both video-level

answer prediction and temporally grounded evidence selection. MedVidQA introduced medical instructional video question-answer pairs with annotated visual answer spans (Gupta et al., 2023). HealthVidQA expanded the scale through automatically generated health-related QA examples (Gupta et al., 2024). Earlier MedVidQA shared-task systems showed that transcript-based methods can be effective when narration matches the visual step, but exact span placement remains difficult (Gupta and Demner-Fushman, 2022; Kusa et al., 2022).

Temporal grounding methods match a language query to a moment in a video. Early work scored candidate segments against a query (Anne Hendricks et al., 2017; Gao et al., 2017; Wang et al., 2019). Later models improved cross-modal interaction, proposal construction, and start-end prediction (Mun et al., 2020; Zhang et al., 2020b,a; Lei et al., 2021). Long-video grounding adds coarse-to-fine search and query-conditioned representations so that a model need not score every frame at full resolution (Hou et al., 2023; Moon et al., 2023).

Long-form video QA and video-language models provide stronger multimodal alternatives (Zhong et al., 2022; Nguyen et al., 2024). Recent systems add explicit temporal information to video LLMs or combine multimodal evidence for grounding (Ren et al., 2024; Guo et al., 2025; Li et al., 2025; Pramanick et al., 2025; Xiao et al., 2025). The UNCC system did not train such a model. It kept the models fixed and placed structure around preprocessing, proposal construction, constrained ranking, and validation. Classical retrieval and sentence embeddings remained useful because they can score timestamped transcript units directly (Robertson and Zaragoza, 2009; Reimers and Gurevych, 2019).

3 Task and Data

Task C requires one interval prediction for each video-question pair (Gupta et al., 2026). A submission contains start and end timestamps. The hidden gold span is used for overlap scoring. A prediction can identify the correct procedure step and still lose strict-overlap score if it includes setup narration, a repeated demonstration, or a post-action explanation.

The task builds on prior medical video QA resources. MedVidQA contains 3,010 human-annotated questions paired with visual answer spans from 899 health-related instructional videos

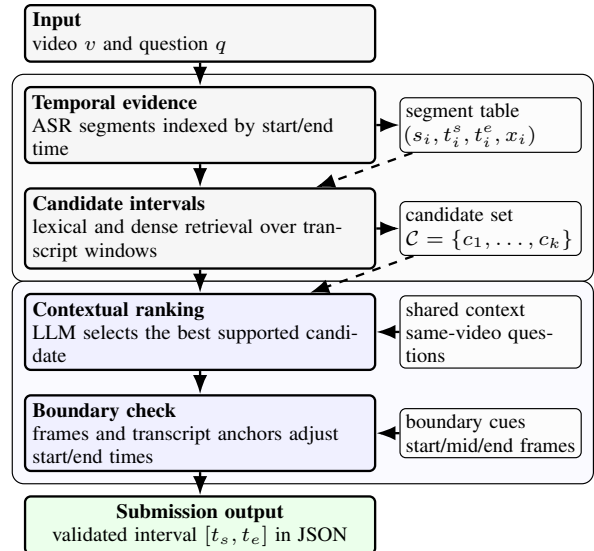


Figure 1. Task C inference pipeline. Timestamped transcript rows and phase blocks define candidate intervals; the ranker chooses from that bounded set, and the final layer validates the JSON interval before submission.

(Gupta et al., 2023). HealthVidQA contains roughly 76K automatically generated question-answer-span examples from about 16K health-related videos (Gupta et al., 2024). Test examples and submissions are handled through CodaBench (Gupta et al., 2026; CodaBench, 2026).

Two data properties shaped the implementation. First, the target duration varies by question type. A “how” question may require a full procedure step, while a “where” or anatomy question may need a short visual interval. Second, multiple questions can refer to the same video. The system must separate adjacent steps without losing the procedure order shared across those questions.

4 System Description

4.1 Transcript and phase-map construction

The first stage constructs a reusable transcript representation for each video. The framework downloads the video, extracts the audio track, and applies GPT-4o Transcribe Diarize to produce timestamped automatic speech recognition (ASR) segments. It stores the resulting transcript in JSON, plain-text, and subtitle-style formats. Each transcript row contains the video identifier, segment identifier, start time, end time, and recognized text. Questions associated with the same video share this transcript table.

The framework derives overlapping transcript windows from adjacent ASR rows. This representation reduces sensitivity to individual ASR segment

boundaries and provides retrieval units that contain enough local context for a procedural step while remaining temporally grounded. The framework also constructs a coarse phase map for each video by grouping transcript content into procedure-level blocks. The phase map is cached at the video level and reused as contextual evidence and as a source of candidate intervals.

This intermediate representation constrains temporal prediction to explicit evidence units. Each candidate interval links to transcript rows, phase rows, or both. The ranker receives local snippets and candidate identifiers instead of the full transcript. This design does not provide frame-level boundary estimation, but it makes each timestamp traceable during error analysis.

4.2 Candidate interval generation

Candidate generation was designed for recall. The system scores transcript windows with lexical matching and dense sentence representations. Lexical scores preserve rare anatomy terms, device names, and procedure verbs. High-scoring windows are expanded locally, snapped to nearby segment boundaries, and merged when they overlap.

The precision branch adds additional candidate families. Segment-neighborhood candidates center a span around top-ranked transcript rows. Phase candidates cover a complete procedure block when the question appears broad. Question typing supplies duration priors, so a short visual question and a full-step question do not inherit the same preferred length.

The candidate list is redundant by design. The goal is to keep the gold neighborhood available for the ranker. Each candidate stores start time, end time, support text, retrieval scores, phase labels, source family, and local context. The ranker compares these alternatives; it does not search the video from scratch.

4.3 Constrained ranking and boundary refinement

The ranking stage uses GPT-5.4 with a fixed JSON schema. For each question, the prompt contains the question, video duration, a bounded candidate table, transcript excerpts around candidate boundaries, and optional same-video context. The response must identify one candidate and may adjust boundaries only when the adjustment is tied to local evidence.

Same-video context is used when multiple questions share a video. The prompt includes neighboring question summaries and candidate regions. This helps when repeated demonstrations or adjacent steps contain similar words and would otherwise receive nearly identical intervals.

The frame path is selective. The framework samples key frames near the selected interval for anatomy, viewpoint, incision-site, and instrument-placement questions. These frames are used as a boundary check after transcript retrieval narrows the search space. The branch is not a dense video encoder, and the paper does not claim frame-level temporal grounding as the primary signal.

The final layer is deterministic. It rounds timestamps, clips them to video length, checks ordering, enforces nonempty intervals, removes duplicate identifiers, and writes the required JSON. If a model call fails or returns invalid JSON, the framework falls back to the best scored candidate rather than leaving the example empty.

5 Implementation Details

The implementation uses one tabular representation after transcript extraction. Video rows, question rows, transcript rows, phase rows, candidate rows, and prediction rows are joined through video and question identifiers. Timestamp bookkeeping stays outside the language model, and one transcript table can serve all questions from the same video.

Intermediate artifacts are cached after each video group. The cache stores video metadata, transcripts, phase maps, candidate tables, question-level reports, and predictions. The cache makes long runs resumable. It also supports post-run debugging, since a wrong timestamp can be traced to ASR coverage, proposal recall, candidate selection, boundary adjustment, or final packaging.

Model outputs use JSON schemas whenever a model must select or refine an interval. The parser attempts repair when the output is close to valid JSON, but repaired outputs still pass through deterministic validation. These checks do not add semantic evidence. They prevent avoidable submission errors and keep the same validation behavior across all examples.

6 Evidence Traceability Analysis

Evidence traceability helped explain why a predicted interval was selected. For each question,

the framework stores the selected candidate source, transcript anchors, phase label, retrieval mode, predicted duration, confidence score, final timestamps, and frame-verification status. These fields connect the final answer span to the transcript and phase evidence used by the ranker.

Candidate accountability provided the clearest traceability signal. Each interval given to the ranker has an identifier, a source family, a support excerpt, and timestamps copied from transcript or phase evidence. This made the outputs easier to inspect. Earlier direct timestamp generation sometimes produced plausible times that were not tied to transcript rows. Candidate-constrained ranking reduced this problem by forcing the model to choose from explicit intervals.

The framework also checked transcript quality before ranking. The audit flagged videos with too few segments, low transcript coverage, unusually long segments, large time gaps, repeated text, or timestamps that were not in chronological order. These checks were useful because weak transcript evidence can lead to weak candidate intervals. If the transcript skips part of the video or contains long gaps without speech, the ranker has less evidence for exact boundary placement.

Source agreement was another useful traceability signal. Candidate sets came from lexical windows, dense-retrieval windows, segment-neighborhood spans, and phase-based spans. If multiple candidate sources pointed to the same temporal region, the selected interval had stronger support. If the sources disagreed, the prediction was more likely to drift toward nearby narration or a broader procedure block. Future systems can use this disagreement as a signal to apply stronger visual verification near the candidate boundaries.

7 Experimental Setup

The official run processed the Task C test set using the submitted inference pipeline. Videos sharing the same URL reused the same transcript table, phase map, and candidate-generation outputs. The final submission used the augmented branch described in Section 4.

Transcript extraction used GPT-4o Transcribe Diarize. Candidate ranking and boundary refinement used GPT-5.4. No supervised fine-tuning was performed. Duration behavior came from question-type heuristics, candidate construction, prompting, and deterministic guards.

Rank	Team	Run	0.3	0.5	0.7	mIoU
<i>Baseline</i>						
–	Baseline	TimeLens-7B78.75	63.75	48.75	48.75	61.09
<i>Participant systems</i>						
1	LAMAR-2	Best	93.75	90.00	77.50	79.55
2	NJUST-KMG	Best	92.50	81.25	67.50	75.48
3	405621	Best	60.00	55.00	47.50	50.78
4	TXT66	Best	71.25	52.50	42.50	52.30
5	UNCC	Best	62.50	36.25	22.50	42.57
6	ADAPT	Best	10.00	10.00	8.75	8.62

Table 1. Organizer-reported Task C VAL results. The numeric columns are IoU thresholds 0.3, 0.5, and 0.7, followed by mIoU. Ranking follows best-run IoU@0.7, the primary metric in the shared-task overview and CodaBench leaderboard (Gupta et al., 2026; CodaBench, 2026).

The ranking prompt used a fixed structure: question text, video duration, candidate identifiers, start and end times, support excerpts, phase labels, source-family metadata, and local neighboring transcript rows. The expected response was a JSON object with the selected candidate, optional boundary adjustment, confidence, and short rationale. Invalid, missing, or out-of-range fields were repaired only when they could be mapped back to stored evidence; otherwise the framework used the top scored candidate.

Run logs were retained for qualitative analysis. For each question, the report stores the selected candidate, transcript anchors, phase context, final interval, validation status, and whether frame verification was invoked. These logs were not submitted to the official scorer, but they were used to analyze boundary failures after the run.

8 Results

The official Task C VAL ranking uses IoU@0.7 as its primary metric (Gupta et al., 2026). Our UNCC run ranked fifth among six participant systems. It scored 62.50 at IoU@0.3, 36.25 at IoU@0.5, 22.50 at IoU@0.7, and 42.57 mIoU. The top-ranked system, LAMAR-2, scored 77.50 at IoU@0.7.

Table 1 suggests that boundary refinement was a larger weakness than initial evidence retrieval. In the reviewed failures, lexical windows, dense-retrieval windows, and phase spans could point to a similar procedural region, while the final interval still included neighboring actions or explanatory narration. This pattern appeared when a transcript window covered more than one procedural step or when adjacent spans reused similar terminology.

The candidate-constrained design improved traceability during post-run inspection. Each prediction could be linked to transcript rows, phase labels, candidate families, and validation records. This made it possible to separate proposal-

generation errors from ranking errors. Some failures occurred because none of the generated candidates tightly covered the reference interval. Other failures occurred even when a reasonable candidate existed, indicating that the ranking stage preferred broader spans with stronger transcript overlap.

The submitted framework also exposed limitations of transcript-centered localization. Broad procedural questions can be supported by multiple neighboring transcript windows, while short visual questions may depend on evidence that is weakly represented in narration. Sparse key-frame checks could reject some incorrect regions, but they did not provide consistent boundary refinement. These cases suggest that future versions should combine transcript-grounded retrieval with stronger local visual verification around candidate boundaries.

9 Error Analysis

Post-run inspection showed that errors came from different stages of the framework. Transcript granularity was the first source. Some ASR rows covered more than one procedural action. When these rows became strong evidence units, the candidate generator produced intervals that were too wide. The ranker then had limited support for selecting a tighter span.

The reviewed boundary errors followed a small set of patterns. Some intervals started too early because setup narration was included with the target action. Others ended too late because follow-up explanation remained in the selected span. Long ASR segments produced coarse anchors. Nearby questions sometimes caused adjacent procedural steps to merge. These errors reflected imprecise temporal boundaries rather than a completely unrelated medical topic.

Transcript quality also affected candidate retrieval. Medical terms, anatomy names, abbreviations, and device labels were not always transcribed consistently. This reduced lexical matching before the ranking step. Dense retrieval helped in some cases because it could recover related transcript windows without exact word overlap.

Question type changed the difficulty of temporal localization. Broad procedure questions could support longer intervals. Short visual questions required tighter evidence. Transcript retrieval could identify a likely region, but sparse key-frame checks did not always refine the exact boundary.

Boundary calibration remained difficult under

candidate-constrained ranking. Earlier unconstrained prompting sometimes produced timestamps that were not tied to transcript evidence. Candidate constraints reduced this issue by forcing the model to choose from explicit intervals. Selected spans could still be too long when the spoken explanation extended beyond the visible action.

Key-frame checks are selective and boundary-local rather than dense visual modeling. This design made the run easier to inspect, but it limited frame-level precision. Future versions should add stronger local visual verification around candidate boundaries while keeping the constrained candidate structure.

10 Discussion

The final system makes a useful but limited trade-off. It favors bounded, auditable timestamp selection over an end-to-end video grounding model. A timestamp pair by itself is hard to debug. A timestamp pair with candidate identifiers, transcript excerpts, phase labels, source-family metadata, and validation flags gives enough context to inspect the decision. In medical videos, that context matters because the answer may be described before the visual step is fully visible.

Some development choices were kept out of the final branch. Full-transcript prompting was one of them. It led the model toward broad regions around the answer phrase instead of local evidence windows. Windowed retrieval gave the ranker shorter evidence and separated same-video questions more cleanly. Direct timestamp generation was also removed because the output could look reasonable while having no transcript anchor.

Phase maps helped when used as context and as one proposal family. A phase block can represent a meaningful procedural step, but many gold spans are shorter than a phase. Copying full phase boundaries into the prediction produced long intervals. The final branch pairs phase candidates with segment-neighborhood candidates and duration guards so that the ranker can choose a shorter local span when needed.

The frame check was deliberately narrow. A few frames around the selected interval can catch obvious mismatches, such as an anatomy question whose proposed span shows only a talking head or setup. It cannot replace dense visual grounding. A future version could add visual embeddings as another proposal signal before the LLM ranker. The

transcript branch would identify a manageable temporal region, visual similarity could rescore short windows inside that region, and the ranker could compare candidates with both transcript and visual evidence. Another option is boundary-local multimodal refinement: after a candidate is selected, sample densely near the proposed start and end, then score small shifts with a video model trained for temporal grounding. This would address the text-reliance concern without discarding the auditable candidate structure.

No ablation experiments were added after the official run. The paper reports diagnostics rather than claiming that one component is empirically optimal. Useful statistics for future versions include proposal recall, fallback rate, selected source-family distribution, predicted-duration distribution, and the effect of boundary refinement. The current report format already records some of these fields. They would separate a proposal-generation failure from a ranker failure and would make comparisons against multimodal baselines more informative.

The next version should keep the parts that made the run inspectable and add denser visual evidence where the scores show a gap. A first stage can maximize recall with transcript, phase, and frame evidence. A second stage can rank candidates with source-agreement features and local evidence. A third stage can run boundary-specific visual checks near the selected start and end. A final stage can export both the interval and the audit record.

11 Conclusion

The UNCC Task C system frames visual answer localization as constrained temporal interval selection. The framework extracts timestamped transcripts, builds phase maps, generates redundant candidate spans, ranks candidates through a schema-controlled model call, applies selective key-frame boundary checks, and validates the final JSON output. The official run achieved 42.57 mIoU and ranked fifth among participant systems. The error profile matches the system design. Transcript and phase evidence support coarse temporal grounding, while strict localization needs finer temporal evidence, denser boundary checks, and learned duration calibration.

12 Limitations

The system runs at inference time only. It does not learn temporal calibration from released supervi-

sion. Duration behavior comes from question-type heuristics, candidate construction, prompting, and deterministic guards. These priors prevent some obvious failures, but they have not been validated as learned parameters.

The method depends on transcript quality. ASR errors, coarse segments, and narration that is offset from visible action can move candidate spans away from the true visual answer. The frame branch handles only selected question types and samples sparsely, so it cannot recover every boundary error caused by text-video misalignment.

The evaluation uses a single official test run. We do not report ablations over ASR choice, phase-map construction, candidate families, or boundary refinement. The diagnostics identify likely causes of errors, but they do not quantify the contribution of each component. Reproducibility is also limited by proprietary model calls for transcription, ranking, and refinement, although the paper specifies the inputs, output schema, stored artifacts, validation behavior, and fallback logic used by the framework.

Finally, the output format requires one interval per question. Some instructional videos show the same answer in repeated demonstrations or multiple camera views. The system must choose one span even when more than one interval contains valid evidence.

References

- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.
- CodaBench. 2026. MedGenVidQA 2026 task c: Visual answer localization leaderboard. <https://www.codabench.org/competitions/14015/>. Accessed 30 April 2026.
- Hilmi Demirhan and Wlodek Zadrozny. 2023. Survey of multimodal medical question answering. *BioMed Informatics*, 4(1):50–74.
- Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.
- Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. 2025. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video

- temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3302–3310.
- Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2023. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158.
- Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2024. [Towards answering health-related questions from medical videos: Datasets and approaches](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16399–16411, Torino, Italia. ELRA and ICCL.
- Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2025. A dataset of medical questions paired with automatically generated answers and evidence-supported references. *Scientific Data*, 12(1):1035.
- Deepak Gupta, Collin Scott Campbell, Pedram Golnari, and Dina Demner-Fushman. 2026. Overview of the medgenvidqa 2026 shared task on medical generative video question answering. In *Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026)*, San Diego, USA. Association for Computational Linguistics.
- Deepak Gupta and Dina Demner-Fushman. 2022. Overview of the medvidqa 2022 shared task on medical video question-answering. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 264–274.
- Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, Wk Chan, Chong-Wah Ngo, Mike Zheng Shou, and Nan Duan. 2023. Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8013–8028.
- Wojciech Kusa, Georgios Peikos, Oscar Espitia, Allan Hanbury, and Gabriella Pasi. 2022. Dossier at medvidqa 2022: Text-based approaches to medical video answer localization problem. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 432–440.
- Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858.
- Xuefen Li, Bo Wang, Ge Shi, Chong Feng, and Jiahao Teng. 2025. Mitigating the discrepancy between video and text temporal sequences: A time-perception enhanced video grounding method for llm. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9804–9813.
- WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. 2023. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23023–23033.
- Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819.
- Thong Thanh Nguyen, Zhiyuan Hu, Xiaobao Wu, Cong-Duy T Nguyen, See Kiong Ng, and Luu Anh Tuan. 2024. Encoding and controlling global semantics for long-form video question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7049–7066.
- Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. 2025. Enrich and detect: Video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24297–24308.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323.
- Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.
- Weining Wang, Yan Huang, and Liang Wang. 2019. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 334–343.
- Junbin Xiao, Qingyun Li, Yusen Yang, Liang Qiu, and Angela Yao. 2025. Unleashing the power of llms for medical video answer localization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 669–679. Springer.
- Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020a. Span-based localizing network for natural language video localization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6543–6554.
- Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020b. [Learning 2d temporal adjacent networks for moment localization with natural language](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 6439–6455.

Author Index

- Abdalla, Mohamed, 201
Abrar, Ajwad, 123
Achakulvisut, Titipat, 179, 233
Adib, Shefayat, 123
Agarwal, Shubham, 155
Ahmed, Kawsar, 191
Ahuir, Vicent, 71
Albrecht, Jens, 59
Alif, Md Hasibur, 123
Alliheedi, Mohammed, 141
- Babakova, Liudmila, 104
Basavaraj, Arshitha, 132
Basu, Subhrajyoti, 66
Baumgartner, William, 171
Bighane, Suhani, 243
Bin Maksud, Saad, 109
- Campana, Pedro Gabriel, 92
Can, Cat, 164
Castro-Bleda, María, 71
Chan, Callum, 132
Chiewhawan, Monrada, 179
Chowdhury, Nahian, 54
Chu, Anh, 164
- Demirhan, Hilmi, 262
Dhaktode, Pratik, 243
Do, Dat, 164
Dobson, Richard, 155
Dogan, Gulustan, 257
- Ebinesar, Basil, 248
Elshehaby, Ahmed, 201
Eskandari, Amir, 146
- Fahim, Tamjid, 109
- Galat, Dima, 38
- Hoque, Mohammed Moshiul, 191
Horaira, Muhammad Abu, 54
Hossain, Nahid, 213
Hua, Zhontian, 24
Huang, Bin, 13, 24
Hurtado, Lluís, 71
- Inkpen, Diana, 132
- Ittichaiwong, Piyalitt, 47
- Jiang, Keyuan, 248
Johan, Syed, 109
- Karip, Siam, 213
Kondadadi, Rishik, 87
Kumari, Jyoti, 224
- Lavelli, Alberto, 92
Le, Quynh, 164
Lekuthai, Nopporn, 233
Li, Jinglong, 229
Li, Kexin, 33
Limaroon, Keetawan, 179
Lossio-Ventura, Juan Antonio, 132
Lowphansirikul, Lalita, 47
Luo, Yi, 24
Luongo-Vazquez, Christopher, 104
- Machina, Anemily, 141
Maddineni, Charansai, 248
Magnini, Bernardo, 92
Mai, Phuong, 164
Martinez-Santos, Juan Carlos, 6
Menco Tovar, Andrea, 6
Mercer, Robert, 141
Mohamed, Youssef, 201
Mondal, Anindita, 224
Morabito, Niccolò, 1
- Paul, Durjoy, 132
Pereira, Francisco, 132
Perez-Rosas, Veronica, 132
Phakatkar, Anupama, 243
Pham, Huy-Hieu, 77
Pham, Quoc-Cuong, 77
Puertas, Edwin, 6
- Rahman, Mehreen, 54
Raja, Ashok, 248
Rebayet, Oarisa, 191
Rizoiu, Marian, 38
Roy, Sudipta, 141
Rudolph, Eric, 59
- Saha, Pritha, 99
Saha, Shuvodwip, 99

Sani, Ahmed, 123
Schilling, Lisa, 171
Searle, Thomas, 155
Sermsrisuwan, Watcharitpol, 233
Serrano, Jairo E., 6
Shanto, Anik Mahmud, 99
Shohan, Symom Hossain, 191
Steigerwald, Philipp, 59
Stepin, Iliia, 104
Su, Liuyuan, 13

Tao, Jing, 146
Tian, Xiaotian, 257
Tran, Luong, 164

Ulli, Vinay, 224

Vu, Hoang-Thuy-Duong, 77

Walee, Radiul, 191
Wang, Xindi, 141
Wang, Yetian, 141
Wang, Yiran, 33
Wang, Zhekun, 33

Yang, Yang, 229
Yoadsanit, Seksan, 233
Yuan, Kaixuan, 13, 24

Zadrozny, Wlodek, 262
Zhang, Kunli, 13, 24
Zhang, Shixin, 13
Zhao, Di, 33
Zhao, Guanghui, 13, 24
Zulkernine, Farhana, 146