Quasar at AraHealthQA Track 1 : Leveraging Zero-Shot Large Language Models for Question and Answer Categorization in Arabic Mental Health

Adiba Fairooz Chowdhury and MD Sagor Chowdhury
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004014, u2004010}@student.cuet.ac.bd

Abstract

Pre-trained language models (PLMs) show potential for advancing mental health care, yet their effectiveness in Arabic mental health contexts is underexplored. This study evaluates PLMs on two multi-label classification tasks from the AraHealthQA 2025 shared task Track 1: question categorization and answer strategy classification. We systematically evaluate several LLMs spanning Arabic-specialized, multilingual, and general-purpose architectures using zero-shot inference, with comparative analysis revealing Qwen3-14B's superior performance. Our approach combines prompt-based inference, label mapping, and strategically crafted Arabic prompts. Experiments on 350 training and 150 test samples demonstrate competitive performance, securing 4th place in both tasks (Question F1: 0.52, Answer F1: 0.76; Question Jaccard: 0.41, Answer Jaccard: 0.66). These findings reveal strengths and limitations of current PLMs for detecting complex intents in Arabic mental health contexts.

1 Introduction

Pre-Trained Language Models (PLMs) have transformed many domains, including medicine (He et al., 2023), yet research on their application to mental health remains nascent. PLMs offer promising support for patients and tools for healthcare providers, from conversational agents (Liu et al., 2023; Brocki et al., 2023) to classifying user input for therapeutic intervention (Sharma et al., 2023). However, effective mental health PLMs must grasp symptom nuances and subjectivity, a greater challenge for Arabic. Spoken by over 400 million people, Arabic's rich morphology, dialect diversity, right-to-left script, and context-sensitive character shapes complicate NLP (Guellil et al., 2021). Despite advances in other languages (Atapattu et al., 2022; Kabir et al., 2022; Sun et al., 2021), Arabic mental health NLP is underexplored, with limited prior studies (Abdulsalam et al., 2024; Aldhafer and Yakhlef, 2022; Al-Musallam and Al-Abdullatif, 2022; Al-Laith and Alenezi, 2021; El-Ramly et al., 2021).

This paper reports our submission to AraHealthQA 2025 Track 1 (Alhuzali et al., 2025), which targets Arabic mental health discourse. We assess zero-shot performance of large PLMs, particularly Qwen3-14B, on multi-label Question Categorization and Answer Strategy Classification. Ranking 4th in both subtasks, our results show zero-shot PLMs can approach fine-tuned models in low-resource, culturally specific settings. This paper's main contributions are as follows:

- First prompt-based, zero-shot classification on MentalQA 2025 without fine-tuning.
- Culturally adapted Arabic prompts for mental health classification.
- Systematic evaluation demonstrating Qwen3-14B's competitive performance.
- Analysis of PLM strengths and limitations for Arabic mental health contexts.

Implementation details are available at[1].

2 Background

2.1 Task Description

ArahealthQA Track 1 is a shared task on Arabic mental health question answering, consisting of:

- **Sub-Task 1:** Multi-label Question Categorization[2] —classifying questions into predefined categories (Table 1).
- **Sub-Task 2:** Multi-label Answer Strategy Classification[3] — categorizing answers according to predefined strategies (Table 1).

[1] https://github.com/AdibAFC/Quasar_ArahealthQA-Track1-MentalQA
[2] https://www.codabench.org/competitions/8559/
[3] https://www.codabench.org/competitions/8730/

Proceedings of The Third Arabic Natural Language Processing Conference, pages 155–163
November 8-9, 2025

#	Q-Types	#	A-Types
A	Diagnosis	1	Information
B	Treatment	2	Direct Guidance
C	Anatomy and Physiology	3	Emotional Support
D	Epidemiology		
E	Healthy Lifestyle		
F	Provider Choices		
Z	Other		

Table 1: Question (Q) and Answer (A) types.

2.2 Dataset

The shared task uses the MentalQA dataset (Alhuzali et al., 2024), containing 500 annotated Arabic Q&A posts (350 development, 150 test) specialized in mental health discourse. Table 2 illustrates input-output examples.

Subtasks	Input(Arabic)	Output
Question Categorization	عدم تركيز مع الآخرين وتوتر واختناق وعدم القدرة على النوم وساوس وزن رهيب في الدماغ والتوتر الدائم	['A', 'D', 'E']
Answer Categorization	واضح انك توتري قوى حاولى انك وانت بتتكلمى مع الآخرين أنك لا تأخذي الموضوع على أنك في وضع تقييم ثقي في نفسك و ركزي عليها اكتر من رأي الناس فيك	['1', '3']

Table 2: Sample input-output mapping with Arabic question-answer and corresponding labels

2.3 Related Work

PLM development for English has progressed rapidly with models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019). Despite Arabic being the fourth most prevalent language online with over 400 million speakers, few PLMs exist due to Arabic's linguistic complexity (Shaalan et al., 2019). Mental health NLP research has primarily focused on English, leaving Arabic question and answer classification underexplored. The recent MentalQA dataset marks important progress, with reviews emphasizing the need for specialized Arabic NLP resources in mental health (Alasmari, 2025). Recent efforts also include MedArabiQ, benchmarking large language models on Arabic medical tasks (Abu Daoud et al., 2025).
Recent developments in Arabic mental health NLP have shown promising advances (Alhuzali and Alasmari, 2025; Zahran et al., 2025), demonstrating both the effectiveness of domain-specific adaptations and the challenges of applying contemporary LLMs to Arabic mental health discourse. Practical applications have emerged (Bensalah et al., 2024), leveraging AI for multilingual mental health support. Comprehensive reviews (Alasmari, 2025) have systematically analyzed Arabic NLP applications in mental health, identifying key gaps and research directions.
This work provides novel benchmarks and insights for culturally aware, low-resource Arabic mental health NLP applications through large-scale multilingual PLMs and prompt-based adaptation.

3 System Overview

Our system evaluates multiple large language models for Arabic medical question classification using a unified zero-shot inference pipeline. We systematically compare six models, spanning Arabic-specialized, multilingual, and general-purpose architectures, to assess their effectiveness specifically in mental health discourse classification.

3.1 Model Selection Rationale

We selected models based on three criteria: (1) Arabic language capabilities, (2) architectural diversity (encoder-only vs decoder-only), and (3) computational feasibility. The Qwen family was chosen for demonstrated multilingual performance, Llama3.1 for its broad adoption and Arabic support, DeepSeek for its reasoning capabilities, and AraBERTv2 as the Arabic-specialized baseline.

3.2 Multi-Model Architecture Framework

Our evaluation framework accommodates diverse architectures, dividing them into generative (decoder-only) and classification (encoder-only) models. The generative models include Qwen3-14B[4] (14.8B parameters), Qwen2.5-7B[5], and Qwen2-7B[6], each with a 32K context length, Llama3.1-8B — Meta's instruction-tuned multilingual model[7], and DeepSeek R1-7B — a distilled model[8] optimized for reasoning tasks. On the classification side, we use AraBERTv2, an Arabic-specialized BERT variant (aubmindlab/bert-base-arabertv2[9]). To handle large models, we apply 4-bit NF4 quantization, which reduces memory usage by approximately 75% without compromising performance (Dettmers et al., 2021). Memory

[4]https://huggingface.co/Qwen/Qwen3-14B
[5]https://huggingface.co/unsloth/Qwen2.5-7B-Instruct
[6]https://huggingface.co/Qwen/Qwen2-7B
[7]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[8]https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
[9]https://huggingface.co/aubmindlab/bert-base-arabertv2

Prompt Template for Question Classification	Prompt Template for Answer Classification
أنت خبير في تصنيف الأسئلة الطبية باللغة العربية. مهمتك هي تصنيف السؤال التالي إلى فئة أو أكثر من الفئات المحدده. {category_descriptions} السؤال المراد تصنيفه: {question} تعليمات: 1. اقرأ السؤال بعناية وحلل محتواه 2. حدد الفئة أو الفئات المناسبة (يمكن أن يكون هناك أكثر من فئة واحده) 3. اشرح سبب اختيارك لكل فئة 4. في النهاية، اكتب الإجابة بالتنسيق التالي: "التصنيف النهائي: [A,B,C]" (استخدم الأحرف المناسبة مفصولة بفواصل) مثال على التنسيق: - إذا كان السؤال عن التشخيص فقط: "التصنيف النهائي: [A]" - إذا كان السؤال عن التشخيص والعلاج: "التصنيف النهائي: [A,B]"	أنت خبير في تصنيف الإجابات الطبية باللغة العربية. مهمتك هي تصنيف الإجابة التالية إلى استراتيجية أو أكثر من الاستراتيجيات المحدده. {strategy_descriptions} الإجابة المراد تصنيفها: {answer} تعليمات: 1. اقرأ الإجابة بعناية وحلل محتواها وأسلوبها 2. حدد الاستراتيجية أو الاستراتيجيات المناسبة (يمكن أن يكون هناك أكثر من استراتيجية واحده) 3. اشرح سبب اختيارك لكل استراتيجية 4. في النهاية، اكتب الإجابة بالتنسيق التالي: التصنيف النهائي: [1,2,3]" (استخدم الأرقام المناسبة مفصولة بفواصل)" مثال على التنسيق: - إذا كانت الإجابة معلوماتية فقط: "التصنيف النهائي: [1]" - إذا كانت الإجابة تحتوي على معلومات وتوجيه: "التصنيف النهائي: [1,2]" - إذا كانت الإجابة تحتوي على المعلومات والتوجيه والدعم العاطفي: "التصنيف النهائي: [1,2,3]"

Table 3: Structured prompt templates for Arabic question classification (left) and answer classification (right).

optimization is further achieved through dynamic GPU memory management, and the entire system is implemented within the unified Hugging Face ecosystem (further details are in Appendix A).

3.3 Methodology

Our approach ensures consistent evaluation protocols across all models using task-specific Arabic prompts designed for cross-model compatibility. These prompts include structured category listings and reasoning instructions, as summarized in Table 3. We apply model-specific adaptations such as enabling `thinking_mode=True` in Qwen models to facilitate structured reasoning, while other generative models use standard chat templates with equivalent reasoning prompts. For BERT-based models, classification heads are employed with prompt-based input formatting. Outputs from all models undergo a robust regex-based extraction process capable of handling multilingual responses effectively, as illustrated in Figure 1.

Extract Type	Extract Question Categories	Extract Answer Categories
Arabic patterns	"التصنيف النهائي: [A,B,C]" "الفئات: [A,B,C]" "التصنيف: [A,B,C]" "النتيجة: [A,B,C]"	"التصنيف النهائي: [1,2,3]" "الاستراتيجيات: [1,2,3]" "التصنيف: [1,2,3]" "النتيجة: [1,2,3]"
English patterns as fallback	Final Classification: [A,B] Categories: [A, C] Classification: [B]	Final Classification: [1,2] Strategies: [1] Classification: [2, 3]

Figure 1: Regex-based pattern recognition process for extracting categories from Arabic and English responses

3.4 System Pipeline and Algorithm

Our system employs a structured zero-shot classification pipeline supporting both generative and classification models under a unified framework. As illustrated in Figure 2, it uses task-specific Arabic prompts with structured reasoning and model-specific strategies like thinking mode to ensure consistent classification of medical questions. Outputs are standardized through a robust regex-based multi-pattern label extraction process, enabling direct comparison among Arabic-specialized, multilingual, and general-purpose models within the same system.

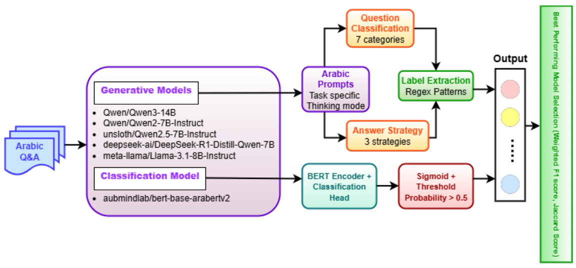


Figure 2: Zero-shot Arabic medical classification pipeline supporting multiple LLM architectures with unified prompt engineering and evaluation framework

3.5 Technical Challenges and Solutions

Achieving consistent Arabic understanding across diverse architectures was a key challenge. Our framework supports both encoder-only models like AraBERTv2 and decoder-only generative models such as Qwen, Llama, and DeepSeek, enabling direct comparison. Dynamic prompt engineering and modular regex-based output processing ensure robustness across varied response formats and languages. Memory limitations were managed with adaptive quantization—4-bit NF4 for large models and standard precision for smaller ones. Evaluation uses probabilistic thresholds and macro-averaged F1 scores for standardized, fair assessment across all models.

3.6 System Example

Detailed examples of model classifications for both questions and answers are provided in Appendix B, Figures 5 and 6.

4 Experimental Setup

4.1 Data Usage and Implementation

The model is used in a zero-shot setting without fine-tuning. Train_Dev.tsv (350 samples) was used for evaluation with gold-standard labels, while test.csv (150 samples) was used for blind inference. Arabic questions were processed without preprocessing to preserve semantic integrity. Prompts were constructed in Arabic with explicit multi-category classification instructions.

4.2 Evaluation Metrics

The model's performance on the labeled `Train_Dev.tsv` set was evaluated using the **Weighted F1-Score** and **Jaccard Similarity**.

$$F1_{\text{weighted}} = \frac{\sum_{i=1}^{n} w_i \cdot \frac{2 \cdot \text{precision}_i \cdot \text{recall}_i}{\text{precision}_i + \text{recall}_i}}{\sum_{i=1}^{n} w_i}$$

$$\text{Jaccard}(T_i, P_i) = \frac{|T_i \cap P_i|}{|T_i \cup P_i|}$$

where T_i and P_i are the ground-truth and predicted label vectors for sample i, w_i is the number of true instances of class i, and n is the total number of classes. Complete implementation details are provided in Appendix A.

5 Results

5.1 Development Set Evaluation

We report performance of various LLMs on both classification tasks using the labeled `Train_Dev.tsv` dataset in zero-shot setting.

Model	Question Class.		Answer Class.	
	F1-Score	Jaccard	F1-Score	Jaccard
Random Baseline	0.326	0.199	0.541	0.378
Majority Class	0.245	0.193	0.451	0.397
Weighted Random	0.386	0.250	0.587	0.432
Qwen3-14B	**0.507**	**0.363**	**0.767**	**0.628**
Qwen2.5-7B	0.504	0.356	0.693	0.529
Qwen2-7B	0.499	0.344	0.688	0.530
DeepSeek R1-7B	0.330	0.213	0.723	0.556
Llama3.1-8B	0.315	0.207	0.632	0.541
AraBERTv2	N/A	N/A	0.466	0.563

5.2 Official Competition Results

Our best-performing system (Qwen3-14B) achieved 4th place in both subtasks on the blind test set (150 samples):

- Question Classification: Weighted F1-Score = 0.52, Jaccard = 0.41
- Answer Classification: Weighted F1-Score = 0.76, Jaccard = 0.66

5.3 Comparative Analysis

Qwen3-14B consistently outperformed other models and baseline methods, with performance substantially exceeding random, weighted and majority class baselines. Complete baseline analysis and model comparisons are provided in Appendix C.

5.4 Error Analysis

Analysis of confusion matrices reveals key error patterns: Question classification shows frequent confusion between *Diagnosis* and *Healthy Lifestyle* (89 cases), and between *Treatment* and *Diagnosis* (111 cases). Answer classification shows significant confusion between *Information* vs. *Direct Guidance* categories. Technical issues included irregular formatting requiring robust regex post-processing and occasional model refusal to classify ambiguous content.

Technical Implementation Issues

- Irregular formatting requiring robust regex post-processing
- Inconsistent Arabic/English label mixing in model outputs
- Occasional model refusal to classify ambiguous mental health content

The confusion matrices (Figures 3 and 4) illustrate these classification patterns, with diagonal dominance indicating generally good performance despite the identified challenges. Specific examples of model output errors for both tasks are provided in Appendix D

5.5 Cross-Architecture Analysis

Our systematic evaluation reveals distinct performance patterns across model architectures:

Qwen Family Dominance: The Qwen models (Qwen3 > Qwen2.5 > Qwen2) demonstrate superior Arabic comprehension, with Qwen3-14B achieving the highest scores in both tasks. This suggests that the Qwen architecture's multilingual pretraining particularly benefits Arabic mental health discourse.

Model Size Effects: Larger models generally outperform smaller ones within the same

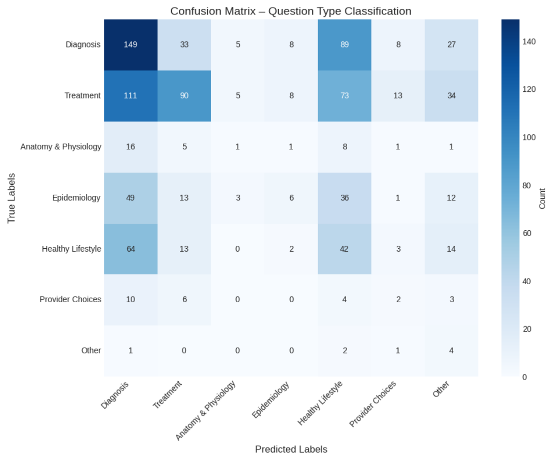


Figure 3: Question Classification Confusion Matrix

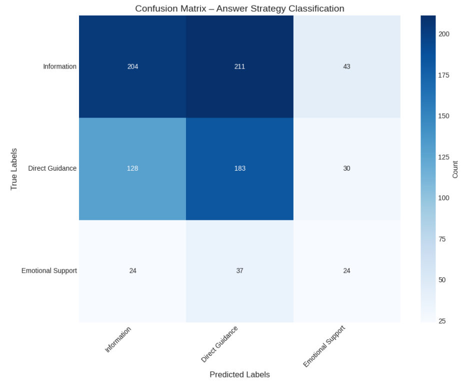


Figure 4: Answer Classification Confusion Matrix

family, with Qwen3-14B (14B) outperforming Qwen2.5-7B and Qwen2-7B in question classification, though the gap is smaller for answer classification.

Specialized vs General Models: The comparison between Arabic-specialized AraBERTv2 and multilingual generative models reveals that recent large multilingual models can match or exceed specialized models in domain-specific tasks.

6 Discussion

6.1 Model Architecture Insights

Our comparative analysis reveals several insights: (1) The Qwen family's superior performance suggests that certain multilingual pre-training strategies better capture Arabic linguistic nuances, (2) Decoder-only models generally outperform encoder-only models for these classification tasks, and (3) Model size provides diminishing returns within the same architecture family.

6.2 Arabic-Specific Challenges

The performance gap between models highlights the continued challenges in Arabic NLP, where models not specifically designed for Arabic underperform significantly (Llama3.1 vs Qwen3-14B: 0.315 vs 0.507 F1 for questions).

7 Conclusion

We presented a systematic evaluation of multiple LLM architectures for zero-shot Arabic mental health classification, with our best system (Qwen3-14B) achieving 4th place in both tasks. Our comparative analysis demonstrates that recent multilingual models can achieve competitive performance without fine-tuning, though significant performance gaps exist between model families. The Qwen architecture's superior performance suggests that specific multilingual pre-training strategies better capture Arabic linguistic nuances. Limitations include lack of domain-specific adaptation and output format variability across models. Future work includes domain-specific fine-tuning on larger Arabic medical corpora, incorporating retrieval-augmented generation for contextual understanding, evaluation across diverse Arabic dialects, investigating prompt engineering techniques for medical domains.

Acknowledgments

We thank all contributors who supported this work, especially the developers of the LLM models evaluated. We appreciate feedback from anonymous reviewers. This research did not receive specific funding.

References

Amani Abdulsalam, Abdullah Alhothali, and Saeed Al-Ghamdi. 2024. Detecting suicidality in arabic tweets using machine learning and deep learning techniques. *Arabian Journal for Science and Engineering*, 49:12729–12742.

Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar Habash, and Farah E Shamout. 2025. Medarabiq: Benchmarking large language models on arabic medical tasks. *arXiv e-prints*, pages arXiv–2505.

Afnan Al-Laith and Mansour Alenezi. 2021. Monitoring people's emotions and symptoms from arabic tweets during the covid-19 pandemic. *Information*, 12(2):86.

Nouf Al-Musallam and Majed Al-Abdullatif. 2022. Depression detection through identifying depressive arabic tweets from saudi arabia: Machine learning approach. In *2022 Fifth National Conference of Saudi Computer Colleges (NCCC)*, pages 11–18. IEEE.

Ashwag Alasmari. 2025. A scoping review of arabic natural language processing for mental health. In *Healthcare*, volume 13-9, page 963. MDPI.

Saeed H. Aldhafer and Mahdi Yakhlef. 2022. Depression detection in arabic tweets using deep learning. In *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 1–6. IEEE.

Hassan Alhuzali and Ashwag Alasmari. 2025. Pretrained language models for mental health: An empirical study on arabic qa classification. *Healthcare*, 13(9).

Hassan Alhuzali, Ashwag Alasmari, and Hamad Alsaleh. 2024. Mentalqa: An annotated arabic corpus for questions and answers of mental healthcare. *IEEE Access*, 12:101155–101165.

Hassan Alhuzali, Farah Shamout, Muhammad Abdul-Mageed, Chaimae Abouzahir, Mouath Abu-Daoud, Ashwag Alasmari, Walid Al-Eisawi, Renad Al-Monef, Ali Alqahtani, Lama Ayash, Nizar Habash, and Leen Kharouf. 2025. Arahealthqa 2025 shared task description paper. In *Proceedings of ArabicNLP 2025*.

Thilina Atapattu, Madurika Herath, Chiran Elvitigala, Prashan de Zoysa, Kasun Gunawardana, Madurika Thilakaratne, Kasun de Zoysa, and Katrina Falkner. 2022. Emoment: An emotion annotated mental health corpus from two south asian countries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6991–7001.

Nour Bensalah, Hiba Ayad, Ali Adib, and Abdelouahid Ibn El Farouk. 2024. Mindwave app: Leveraging ai for mental health support in english and arabic. In *2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC)*, pages 1–6. IEEE.

Laura Brocki, George C. Dyer, Agnieszka Gładka, and N.C. Chung. 2023. Deep learning mental health dialogue system. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 395–398. IEEE.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *Preprint*, arXiv:2110.02861. ICLR 2022 Spotlight Version.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Mahmoud El-Ramly, Hadeer Abu-Elyazid, Yomna Mo'men, Ghadah Alshaer, Nermine Adib, Khaled A. Eldeen, and Manar El-Shazly. 2021. Cairodep: Detecting depression in arabic posts using bert transformers. In *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 207–212. IEEE.

Imane Guellil, Hichem Saâdane, Faycal Azouaou, Bachir Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33:497–507.

Kuan He, Ruixiang Mao, Qian Lin, Yuxuan Ruan, Xiaodong Lan, Minlie Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: From data, technology, and applications to accountability and ethics. *arXiv e-prints*, pages arXiv–2310.05694.

Md Kamrul Kabir, Md Islam, A N M Baki Kabir, Anwarul Haque, and Md Kamrul Rhaman. 2022. Detection of depression severity using bengali social media posts on mental health: Study using natural language processing techniques. *JMIR Formative Research*, 6:e36118.

Jiaming Liu, Dan Li, Hao Cao, Tao Ren, Zhen Liao, and Jian Wu. 2023. Chatcounselor: A large language model for mental health support. *arXiv e-prints*, pages arXiv–2309.15461.

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*.

Khaled Shaalan, Sajid Siddiqui, Moustafa Alkhatib, and Ahmed Abdel Monem. 2019. Challenges in arabic natural language processing. In *Computational Linguistics, Speech and Image Processing for Arabic Language*, pages 59–83. World Scientific.

Aniket Sharma, I.W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5:46–57.

Huan Sun, Zheng Lin, Chuxu Zheng, Shasha Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Nourhan Zahran, Amr E Fouda, Rana J Hanafy, and Mostafa E Fouda. 2025. A comprehensive evaluation of large language models on mental illnesses in arabic context. *arXiv preprint arXiv:2501.06859*.

A Implementation Details

A.1 Technical Environment

- **Hardware:** NVIDIA A100 GPU (80GB VRAM), 128GB RAM
- **Software:** Python 3.10, CUDA 11.8, Transformers v4.51.0, PyTorch v2.2.0, BitsAndBytes v0.43.0

A.2 Multi-Model Configuration Parameters

Generative Models (Qwen, Llama, DeepSeek): Temperature: 0.7, max_new_tokens: 512, top_p: 0.9, repetition_penalty: 1.1. Model-specific adaptations: Qwen models use `thinking_mode=True`, DeepSeek uses temperature=0.3 for reasoning, Llama3.1 uses standard instruct templates.

Classification Model (AraBERTv2): Linear classification head (768 $\rightarrow$ num_labels), max sequence length: 512, full precision due to smaller size and architectural differences.

Quantization Strategy: 4-bit NF4 for models >10B parameters (Qwen3-14B), 8-bit or full precision for smaller models based on VRAM availability.

A.3 Architecture-Specific Implementation

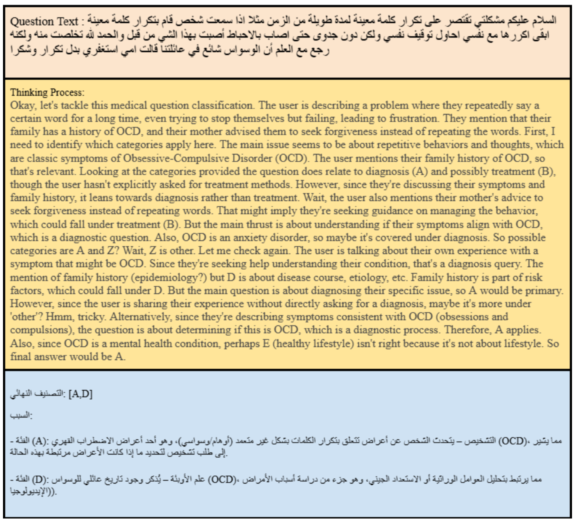
Decoder-Only Models: Unified generation pipeline with model-specific chat templates and reasoning prompts. Output processed via regex extraction and label mapping.

Encoder-Only Model (AraBERTv2): Direct classification using linear head with sigmoid activation for multi-label prediction. Compatible only with answer classification task due to formatting constraints.

A.4 Memory Management and Evaluation

Sequential model loading with dynamic quantization prevents OOM errors. Memory requirements: Qwen3-14B (4-bit): 8GB, 7B models (8-bit): 4-6GB, AraBERTv2: 1GB. All models evaluated using identical metrics (sklearn implementation) with `average='weighted'` for fair comparison. .

B System Example Figures


Question Text : السلام عليكم مشكلتي تقتصر على تكرار كلمة معينة لمدة طويلة من الزمن مثلا اذا سمعت شخص قام بتكرار كلمة معينة ابقى اكررها مع نفسي احاول توقيف نفسي ولكن دون جدوى حتى اصاب بالاحباط أصبت بهذا الشي من قبل والحمد لله تخلصت منه ولكنه رجع مع العلم أن الوسواس شائع في عائلتنا قالت امي استغفري بدل تكرار وشكرا

Thinking Process:
Okay, let's tackle this medical question classification. The user is describing a problem where they repeatedly say a certain word for a long time, even trying to stop themselves but failing, leading to frustration. They mention that their family has a history of OCD, and their mother advised them to seek forgiveness instead of repeating the words. First, I need to identify which categories apply here. The main issue seems to be about repetitive behaviors and thoughts, which are classic symptoms of Obsessive-Compulsive Disorder (OCD). The user mentions their family history of OCD, so that's relevant. Looking at the categories provided the question does relate to diagnosis (A) and possibly treatment (B), though the user hasn't explicitly asked for treatment methods. However, since they're discussing their symptoms and family history, it leans towards diagnosis rather than treatment. Wait, the user also mentions their mother's advice to seek forgiveness instead of repeating words. That might imply they're seeking guidance on managing the behavior, which could fall under treatment (B). But the main thrust is about understanding if their symptoms align with OCD, which is a diagnostic question. Also, OCD is an anxiety disorder, so maybe it's covered under diagnosis. So possible categories are A and Z? Wait, Z is other. Let me check again. The user is talking about their own experience with a symptom that might be OCD. Since they're seeking help understanding their condition, that's a diagnosis query. The mention of family history (epidemiology?) but D is about disease course, etiology, etc. Family history is part of risk factors, which could fall under D. But the main question is about diagnosing their specific issue, so A would be primary. However, since the user is sharing their experience without directly asking for a diagnosis, maybe it's more under 'other'? Hmm, tricky. Alternatively, since they're describing symptoms consistent with OCD (obsessions and compulsions), the question is about determining if this is OCD, which is a diagnostic process. Therefore, A applies. Also, since OCD is a mental health condition, perhaps E (healthy lifestyle) isn't right because it's not about lifestyle. So final answer would be A.

التصنيف النهائي: [A,D]

السبب:

- الفئة (A): التشخيص – يتحدث الشخص عن أعراض تتعلق بتكرار الكلمات بشكل غير متعمد (أوهام/وسواسي)، وهو أحد أعراض الاضطراب القهري (OCD)، مما يشير إلى طلب تشخيص لتحديد ما إذا كانت الأعراض مرتبطة بهذه الحالة.

- الفئة (D): علم الأوبئة – يُذكر وجود تاريخ عائلي للوسواس (OCD)، مما يرتبط بتحليل العوامل الوراثية أو الاستعداد الجيني، وهو جزء من دراسة أسباب الأمراض (الإبيديولوجيا)).

Figure 5: This demonstrates classification of a question about repetitive behaviors and obsessive thoughts, related to OCD. The system analyzes that the person is describing symptoms consistent with OCD and seeking understanding rather than directly asking for diagnosis. It classifies this as Category A (Diagnosis) since the question involves determining if symptoms align with a specific condition.

C Baseline Analysis and Extended Results

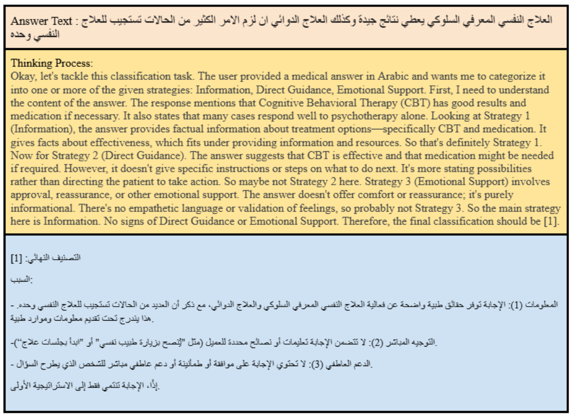
C.1 Baseline Implementation

To validate task difficulty and model performance, we implemented three baseline methods: a random baseline that assigns labels uniformly at random across categories; a majority class baseline that always predicts the most frequent label combination from the training data; and a weighted random baseline that assigns labels randomly but proportional to their frequency in the training set.

C.2 Baseline Performance Analysis

Baseline results demonstrate the inherent difficulty of both tasks:

- Question classification baselines achieve F1 scores of 0.245-0.386, indicating high task complexity with 7 possible categories
- Answer classification baselines achieve higher F1 scores of 0.451-0.587 due to fewer categories (3 vs 7)
- Our Qwen3-14B model achieves 1.6-2.1× improvement over best baselines, confirming meaningful performance gains


Answer Text : العلاج النفسي المعرفي السلوكي يعطي نتائج جيدة وكذلك العلاج الدوائي ان لزم الامر الكثير من الحالات تستجيب للعلاج النفسي وحده

Thinking Process:
Okay, let's tackle this classification task. The user provided a medical answer in Arabic and wants me to categorize it into one or more of the given strategies: Information, Direct Guidance, Emotional Support. First, I need to understand the content of the answer. The response mentions that Cognitive Behavioral Therapy (CBT) has good results and medication if necessary. It also states that many cases respond well to psychotherapy alone. Looking at Strategy 1 (Information), the answer provides factual information about treatment options—specifically CBT and medication. It gives facts about effectiveness, which fits under providing information and resources. So that's definitely Strategy 1. Now for Strategy 2 (Direct Guidance). The answer suggests that CBT is effective and that medication might be needed if required. However, it doesn't give specific instructions or steps on what to do next. It's more stating possibilities rather than directing the patient to take action. So maybe not Strategy 2 here. Strategy 3 (Emotional Support) involves approval, reassurance, or other emotional support. The answer doesn't offer comfort or reassurance; it's purely informational. There's no empathetic language or validation of feelings, so probably not Strategy 3. So the main strategy here is Information. No signs of Direct Guidance or Emotional Support. Therefore, the final classification should be [1].

التصنيف النهائي: [1]

السبب:

- المعلومات (1): الإجابة توفر حقائق طبية واضحة عن فعالية العلاج النفسي المعرفي السلوكي والعلاج الدوائي، مع ذكر أن العديد من الحالات تستجيب للعلاج النفسي وحده. هذا يندرج تحت تقديم معلومات وموارد طبية.

- التوجيه المباشر (2): لا تتضمن الإجابة تعليمات أو نصائح محددة للعميل (مثل "يُنصح بزيارة طبيب نفسي" أو "ابدأ بجلسات علاج").

- الدعم العاطفي (3): لا تحتوي الإجابة على موافقة أو طمأنينة أو دعم عاطفي مباشر للشخص الذي يطرح السؤال.

إذاً، الإجابة تنتمي فقط إلى الاستراتيجية الأولى.

Figure 6: This shows the system classifying a medical response about Cognitive Behavioral Therapy (CBT). The Arabic text discusses CBT treatment and medication options. The system's thinking process correctly identifies this as providing factual information about treatment options and classifies it as Strategy 1 (Information) since it gives facts about effectiveness without offering comfort, reassurance, or specific guidance.

C.3 Extended Model Comparison

The Qwen model family demonstrates superior Arabic understanding compared to other architectures:

- Qwen3-14B vs Qwen2.5: Marginal improvements in both tasks, suggesting architectural refinements
- Qwen vs Llama3-8B: Substantial gaps (0.507 vs 0.315 F1 for questions), highlighting multilingual pre-training advantages
- DeepSeek R1-7B: Strong answer classification (0.723 F1) but weaker question classification, indicating specialized strengths

C.4 Label Distribution Analysis

Training data shows imbalanced distributions affecting baseline performance:

- Question categories: "Diagnosis" (45.2%), "Treatment" (32.1%), "Other" (18.7%), remaining categories <5% each
- Answer strategies: "Information" (52.3%), "Direct Guidance" (31.4%), "Emotional Support" (16.3%)
- This imbalance explains why majority class baselines perform poorly despite dataset size

D Error Analysis Examples

Input Text	Actual label	Predicted label
هل حبوب ميرزاجن لها اضرار (Does Mirtazapine have any side effects?)	[A] (Diagnosis), [D] (Epidemiology), [E] (Healthy Lifestyle)	[A] (Diagnosis), [E] (Healthy Lifestyle)
هل الإحساس بقرب الأجل و الخوف من الموت و الاحلام من أعراض الاكتئاب والقلق؟! و كيف يمكنني تخطي هذه المرحلة لأن حياتي أصبحت جحيم (Are the feeling of impending doom, fear of death, and nightmares symptoms of depression and anxiety? How can I overcome this stage because my life has become hell?)	[A] (Diagnosis), [B] (Treatment), [D] (Epidemiology)	[A] (Diagnosis), [B] (Treatment)
هل يعتبر الخوف من عدم الإنجاب مستقبلاً حالة عادية خاصة لما أكون متعلقة بأطفال كثيراً وأنا على وجه جواز أنا خايفة جداً (Is the fear of not having children in the future a normal condition, especially when I am very attached to children and I am about to get married? I am very afraid.)	[A] (Diagnosis), [D] (Epidemiology)	[A] (Diagnosis), [D] (Epidemiology)

Table 4: Examples of question category classification errors showing model predictions vs. ground truth labels for Arabic mental health questions

Input Text	Actual label	Predicted label
نعم بالإضافة للكثير من الاعراض الاخرى الطبيب النفسي بعد التقييم الدقيق الشامل لكل الأعراض يصف لك مضادات الاكتئاب و مزيلات القلق مع علاج معرفي سلوكي (Yes, in addition to many other symptoms, the psychiatrist, after a comprehensive and thorough evaluation of all symptoms, will prescribe antidepressants and anti-anxiety medications, along with cognitive behavioral therapy.)	[1] (Information), [2] (Direct Guidance)	[1] (Information), [2] (Direct Guidance)
واضح انك توتري قوى حاولى انك وانت بتتكلمى مع الآخرين أنك لا تأخذ الموضوع على أنك في وضع تقييم ثقي في نفسك و ركزي عليها اكتر من رأي الناس فيك (It is clear that you are very nervous. When you are talking to others, try not to take the matter as if you are in a state of evaluation. Trust yourself and focus on yourself more than people's opinion of you.)	[1] (Information), [3] (Emotional Support)	[2] (Direct Guidance), [3] (Emotional Support)
سيتالوبرام أفضل (Citalopram is better)	[1] (Information)	[2] (Direct Guidance)

Table 5: Examples of answer strategy classification errors showing model predictions vs. ground truth labels for Arabic mental health responses