You Never Know a Person, You Only Know Their Defenses: Detecting Levels of Psychological Defense Mechanisms in Supportive Conversations

Hongbin Na[1,*], **Zimu Wang**[2,*], **Zhaoming Chen**[3,*],
Peilin Zhou[4], **Yining Hua**[5] **Grace Ziqi Zhou**[6], **Haiyang Zhang**[2], **Tao Shen**[1],
Wei Wang[2], **John Torous**[5], **Shaoxiong Ji**[7,8], **Ling Chen**[1]
[1]University of Technology Sydney [2]Xi'an Jiaotong-Liverpool University [3]University of Utah
[4]New York University Abu Dhabi [5]Harvard University [6]The University of Sydney
[7]ELLIS Institute Finland [8]University of Turku
🤗 PsyDefConv DMRS Co-Pilot

Abstract

Psychological defenses are strategies, often automatic, that people use to manage distress. Rigid use or overuse of defenses is negatively linked to mental health and shapes what speakers disclose and how they accept or resist help. However, defenses are complex and difficult to reliably measure, particularly in clinical dialogues. We introduce PsyDefConv, a dialogue corpus with help seeker utterances labeled for defense level, and DMRS Co-Pilot, a four-stage pipeline that provides evidence-based pre-annotations. The corpus contains 200 dialogues and 4,709 utterances, including 2,336 help seeker turns, with double-blind labeling reaching Cohen's κ of 0.639. In a counterbalanced study, the co-pilot reduced average annotation time by 24.0%. In expert review, it averaged 4.62 for evidence supportiveness, 4.44 for clinical plausibility, and 4.40 for insight on a seven-point scale. Benchmarks with strong large language models (LLMs) in zero-shot and fine-tuning settings demonstrate clear headroom, with the best macro F1-score around 30% and a tendency to overpredict mature defenses. Corpus analyses confirm that mature defenses are most common and reveal emotion-specific deviations. We release the corpus, annotations, code, and prompts to support research on defensive functioning in language.

1 Introduction

> *"The False Self, if successful in its function, hides the True Self."*
> —Winnicott, 2018

Psychological defenses are among the oldest concepts in psychological science, and are defined as strategies people use to protect themselves from psychic pain, which in turn guide what they disclose, how they reframe difficulties, and how they accept or resist help (Freud, 1936). Recent work has advanced affect modeling (Wang et al., 2024a, 2025; Ma et al., 2025; Zhao et al., 2025), empathy (Hua et al., 2025b; Cai et al., 2024a; Sorin et al., 2024), and strategy selection (Kang et al., 2024; Hua et al., 2025a; Na et al., 2025) for emotional support dialogue (ESC). While psychological defenses are important components of emotionally supportive conversations (Di Giuseppe et al., 2024), the defensive function of utterances remains largely unmodeled in current ESC systems.

The Defense Mechanism Rating Scales (DMRS) is recognized as a gold-standard taxonomy of defensive functioning, comprising three categories, seven levels, and approximately thirty mechanisms validated by 150 descriptive items (Perry and Henry, 2004; Vaillant, 2012). Traditionally, DMRS judgments are tailored for clinical case formulations that synthesize complex behavioral patterns across time and varying situations. In contrast, conversational corpora typically consist of brief dialogue exchanges that lack such longitudinal depth. Rather than viewing this as a limitation, we align the unit of analysis with the unit of evidence by focusing on the seven DMRS levels instead of the individual mechanisms. This theoretically grounded aggregation allows us to operationalize defensive functioning from local conversational cues, significantly enhancing identifiability, inter-annotator agreement, and reproducibility while remaining theoretically faithful to the DMRS hierarchy.

Despite the central role of defenses in clinical theory and practice, to our knowledge, there is no publicly available conversational dataset that annotates DMRS-based defensive functioning at any granularity. The absence of such a resource prevents a reproducible study of how defenses appear in language and blocks systematic evaluation of models on this construct. To address this gap, we introduce PsyDefConv, a dialogue resource that labels defense levels for help seeker utterances. Building on ESConv (Liu et al., 2021), we draw

[*]Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2026, pages 14428–14448

a representative subset of 200 dialogues through stratified sampling over the joint distribution of problem types and emotions. The corpus contains 4,709 utterances in total with 2,336 seeker turns. To cover conversational phenomena, we augment the seven DMRS levels with two labels: *No Defense* to mark phatic or functional turns, and *Needs More Information* to mark cases where context is insufficient. Two trained annotators perform independent double blind labeling and reach substantial agreement with Cohen's $\kappa = 0.639$. Disagreements are adjudicated to form a gold standard.

To support efficiency and consistency, we introduce DMRS CO-PILOT, a four-stage pipeline that delivers level recommendations as pre-annotations. As shown in Figure 3, the system contextualizes the target text, screens candidate items, validates evidence, and synthesizes ranked conclusions. It is refined with annotator feedback and engineered for robust outputs under a strict schema. In a counterbalanced study, it reduces average time per task by 24.0%. In expert review, it averages 4.62 for evidence supportiveness, 4.44 for plausibility, and 4.40 for insight.

We conduct diverse experiments on a range of large language models (LLMs) in both zero-shot and fine-tuning settings. Analyses show that mature defenses dominate overall. Shame displays a higher share of low-level defenses. Anger shows more utterances without defenses. Models tend to over-predict the high adaptive level, and the best macro F1-score is around 30%, which indicates substantial headroom and the need for theory-aware supervision.

Our resource and method align clinical theory with what is observable in dialogue and address two practical challenges: the lack of a conversational dataset for defensive functioning and the under-specification of mechanism labels in text-only episodes. By focusing on DMRS levels and tooling the workflow, we provide a reproducible basis for study and evaluation. Overall, our main contributions are as follows: (1) *Resource*: to the best of our knowledge, our released PSYDEFCONV is the first conversational dataset annotated with DMRS defense levels, covering 2,336 help seeker utterances with double-blind labeling and substantial agreement ($\kappa = 0.639$); (2) *Tooling*: we present DMRS CO-PILOT, a four-stage pipeline that contextualizes the text, screens and validates candidate items, and synthesizes ranked level recommendations as pre-annotations to support consistency and efficiency; (3) *Benchmark and analysis*: we evaluate strong language models in zero-shot and fine-tuning settings, and report error structures and distributional findings, with best macro F1-score around 30%, demonstrating substantial headroom and guiding future work.

2 Background and Related Work

2.1 Psychological Defense Mechanisms

Psychological defense mechanisms are a cornerstone of psychodynamic theory. They refer to the unconscious processes through which the ego manages conflicts among the id, the superego, and external reality (Freud, 1936). Defense mechanisms function by distorting or denying aspects of reality to reduce anxiety and protect the sense of self. While these processes are indispensable for psychological stability, their rigid or excessive use can hinder adaptation and contribute to psychological distress or interpersonal difficulties. In supportive conversations, they often surface in language, as avoidance, rationalization, resistance, or constructive coping, shaping how the dialogue unfolds. Recognizing the defensive function of a speaker's words can offer valuable insight into their inner state and guide more attuned emotional support.

Defense Mechanism Rating Scales (DMRS, Perry et al., 1993) is one of the most empirically grounded instruments for assessing defenses. It arranges these mechanisms into a seven-level hierarchy, from Level 1 (Action Defenses) to Level 7 (High-Adaptive Defenses), providing a structured way to evaluate overall defensive functioning based on psychotherapy transcripts or interview material (Perry and Henry, 2004; Di Giuseppe and Perry, 2021). However, applying these concepts to natural language in an automated way remains an emerging challenge. The subtle, context-dependent nature of defensive expression poses significant difficulties for current NLP models. Bridging this gap will require theory-informed datasets and computational approaches sensitive to the nuances of defensive communication.

2.2 Emotional Support Conversations

Research on ESC aims to build dialogue agents that alleviate users' stress through multi-turn, strategy-grounded interaction. The foundational ESConv corpus (Liu et al., 2021) established this direction with 1,053 dialogues and eight annotated support strategies. The Extended ESConv and ESConv-

Level 0: No Defenses	
Mechanisms	N/A
Definition	Purely functional utterances that serve to maintain conversational flow, express social niceties, or exchange non-emotional information. They do not engage with psychological conflict.
Example	Hello, thank you for your time, or a simple Okay.
Level 1: Action Defenses	
Mechanisms	Passive Aggression, Help-Rejecting Complaining, Acting Out
Definition	Dealing with internal conflicts or external stressors by acting on the environment. The individual's distress is channeled into behavior, often impulsively and without reflection, as a way to release tension, gratify wishes, or avoid fears and painful feelings.
Example	After being criticized at work, instead of discussing the issue, a person goes home and starts a fight with their partner over something minor.
Level 2: Major Image-Distorting Defenses	
Mechanisms	Splitting (of self-image and others' image), Projective Identification
Definition	Coping with intolerable anxiety by grossly distorting the image of oneself or others. This is achieved by splitting representations into polar opposites (all-good or all-bad), which simplifies reality and protects the individual from the anxiety of dealing with ambivalence.
Example	My new boss is a genius and will solve everything, or My boss is completely incompetent and is ruining the company.
Level 3: Disavowal Defenses	
Mechanisms	Denial, Rationalization, Projection, Autistic Fantasy
Definition	Dealing with stressors by refusing to acknowledge unacceptable aspects of reality or one's own experience. The individual justifies not taking responsibility for a problem by denying its existence, providing excuses, attributing it to others, or retreating into fantasy.
Example	I didn't get the promotion because my boss is biased against me, not because my performance was lacking.
Level 4: Minor Image-Distorting Defenses	
Mechanisms	Devaluation (of self-image and others' image), Idealization (of self-image and others' image), Omnipotence
Definition	Protecting self-esteem from threats like failure or criticism by distorting one's image in a less severe manner than Level 2. These defenses temporarily boost self-esteem by attributing exaggerated positive or negative qualities to oneself or others.
Example	Sure, they succeeded, but it was just luck. Anyone could have done it.
Level 5: Neurotic Defenses	
Mechanisms	Repression, Dissociation, Reaction Formation, Displacement
Definition	Managing emotional conflict by keeping unacceptable wishes, thoughts, or motives out of conscious awareness. The individual may experience the feelings associated with a conflict while the idea is blocked (or vice versa), leading to indirect or displaced expressions.
Example	Feeling unexplainably irritable and anxious all day, only to later realize it's the anniversary of a forgotten painful event.
Level 6: Obsessional Defenses	
Mechanisms	Isolation of Affect, Intellectualization, Undoing
Definition	Managing threatening feelings by separating them from the thoughts or events that caused them. The individual remains aware of the cognitive details but avoids the emotional impact by using excessive logic, abstract thinking, or symbolic acts to maintain control.
Example	Describing a traumatic car accident with precise, technical detail but showing no emotion, as if reading a report.
Level 7: High-Adaptive Defenses	
Mechanisms	Affiliation, Altruism, Anticipation, Humor, Self-Assertion, Self-Observation, Sublimation, Suppression
Definition	Representing the most adaptive and constructive ways of handling stressors. The individual faces conflict by consciously integrating feelings with thoughts, anticipating challenges, seeking support, and channeling emotions into productive outcomes.
Example	I'm feeling overwhelmed by this project. I'm going to call a colleague to talk through some strategies and get a fresh perspective.
Level 8: Needs More Information	
Mechanisms	N/A
Definition	Used when an utterance is too ambiguous or the context is insufficient.
Example	When a user's reply is a vague "maybe" in response to a deep emotional question, the intent cannot be determined.

Table 1: Overview of defense mechanisms and their corresponding levels, including examples and descriptions. Levels 0 and 8 are auxiliary labels added to the seven DMRS levels.

SRA datasets (Madani and Srihari, 2025) crop long conversations and generate strategy-conditioned continuations, enabling analysis of how LLMs maintain coherence and strategy consistency with enhanced turns and strategy types. Multi-Strategy ESConv (Dv1/Dv2) (Bai et al., 2025a) emphasizes multiple strategies within a single turn and reconstructs ESConv toward 258 distinct strategy sequences, showing that LLMs outperform supervised models in producing multi-strategy replies.

Another line of work involves creating synthetic conversations to address data scarcity. Zheng et al. (2023a) presents AugESC, which fine-tunes GPT-J 6B on ESConv and uses it to complete dialogue threads from EmpatheticDialogues. Zheng et al. (2023b) creates ExTES via ChatGPT's in-context generation, which is of comparable or slightly higher quality than human-collected benchmarks. Zhang et al. (2024) introduces the ESD-CoT dataset, extending ESConv with explicit reasoning to improve interpretability. ServeForEmo (Ye et al., 2025) is constructed using a role-playing framework with three LLM agents: the help-seeker, the strategy-advisor, and the supporter.

3 PsyDefConv

3.1 Data Source

We build upon ESConv (Liu et al., 2021), a large-scale, publicly available English corpus of emotional support dialogues, and follow its role defini-

tions of seeker and supporter. To obtain a representative 200-dialogue subset, we perform stratified sampling on the joint distribution of problem and emotion types observed in ESConv. We allocate quotas to each combination of the five major problem types and six core emotions in proportion to their ESConv frequencies, thereby preserving the source joint distribution and supporting a broad coverage of dialogue types.

3.2 Annotation Scheme

Each instance corresponds to a single seeker utterance. We focus solely on annotating seeker turns to capture the defensive responses of the distressed individual, rather than the strategies employed by the supporter. Annotators have access only to the dialogue context preceding and including the current seeker utterance. They do not see subsequent turns and annotate only the current utterance. The annotators are two fluent English speakers with expertise in both psychology and natural language processing.

Our scheme adapts the Defense Mechanism Rating Scales (DMRS) (Perry and Henry, 2004; Di Giuseppe and Perry, 2021), an empirically validated rating scale that operationalizes defenses into a hierarchy comprising three categories, seven defense levels, and approximately thirty mechanisms, defined via 150 descriptive items. Although DMRS targets mechanism-level assessment in clinical settings, applying mechanisms to dialogue transcripts is difficult due to the lack of longitudinal, cross-situational evidence. We therefore label *defense levels*, which provide functionally coherent, semantically stable groupings that are more reliably identifiable from dialogue text, supporting higher inter-annotator agreement and reproducibility.

To accommodate conversational phenomena, we add two labels: utterances without defensive function (e.g., greetings, phatic expressions) are labeled *No Defense*, and utterances that are unclassifiable due to insufficient context are labeled *Needs More Information*. These extensions increase coverage of non-defensive and ambiguous cases commonly observed in natural dialogues and reduce annotation noise. For an overview of the different defense levels and their corresponding mechanisms, refer to Table 1.

3.3 Annotator Training

Before full-scale annotation, we ran a structured annotator training program to validate the labeling scheme and iteratively refine operational guidelines. The program comprised four iterations, each covering 10 dialogues, for a total of 461 seeker utterances. Two annotators independently labeled every utterance using our nine-category scheme; after each iteration, we held calibration meetings to review disagreements and update the codebook.

The agreement between annotators progressed over time. Iteration 1 yielded moderate reliability ($\kappa = 0.496$). After we introduced a simplified evaluation protocol, reliability dipped in Iteration 2 ($\kappa = 0.426$) but then rose sharply: Iteration 3 reached $\kappa = 0.642$, and Iteration 4 $\kappa = 0.667$, indicating a shift from moderate to substantial agreement. Overall, reliability improved by 34.5% from the first to the final iteration.

3.4 Formal Annotation

The formal annotation of all 200 dialogues, consisting of 2,336 seeker utterances, was conducted by two trained annotators using the Label Studio (Tkachenko et al., 2020-2025) platform developed by HumanSignal. Analysis results from the *DMRS Co-Pilot* system were imported as pre-annotation suggestions to support efficiency. Independent double-blind annotation yielded a Cohen's κ of 0.639, indicating substantial agreement. To construct the final gold-standard corpus, disagreements were reviewed jointly and resolved by consensus with reference to the annotation handbook.

3.5 Dataset Statistics

As shown in Table 2, PSYDEFCONV consists of 200 dialogues and 4,709 utterances approximately balanced between seekers and supporters. Each dialogue contains an average of 23.5 turns, with a mean utterance length of 19.8 tokens. Detailed statistical distributions of the PSYDEFCONV dataset, including seekers' presenting problems, expressed emotions, and defense mechanism categories, are provided in Appendix B.

Distribution of Psychological Attributes. Table 5 presents the distribution of psychological factors annotated at the seeker level, including presenting problems, expressed emotions, and defense mechanisms. The most common problems include *Ongoing Depression* and *Job Crisis*, while *Anxiety*, *Depression*, and *Sadness* emerge as the dominant emotional expressions. In terms of defense mechanisms, mature defense strategies are the most frequently observed (over 50%), while neurotic

Category	Total	Supporter	Seeker
# Dialogues	200	–	–
# Utterances	4,709	2,373	2,336
Avg. Turns per Dialogue	23.5 ± 6.6	11.9 ± 3.4	11.7 ± 3.3
Avg. Length of Utterances	19.8 ± 16.5	20.9 ± 17.0	18.8 ± 15.8

Table 2: Data Statistics of PSYDEFCONV.

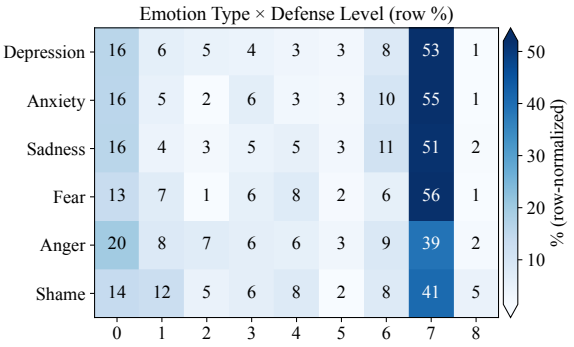

Figure 1: Distribution of defense levels (x-axis) across different emotions (y-axis). Values indicate the proportion of each defense level within a given emotion.

and immature defenses are less common. Notably, 17.4% of utterances are either non-defensive or contextually ambiguous, reflecting emotionally neutral or under-specified dialogue turns.

Distribution of Defense Levels Across Emotions. Figure 1 shows that mature defenses (Level 7) dominate across all emotions, indicating a general preference for adaptive coping. However, variations emerge across emotion types. *Shame* exhibits the highest proportion of low-level defenses (Levels 1–4), suggesting more fragmented and immature regulation. *Anger* stands out for its elevated use of Level 0 (no defense), implying more unfiltered or confrontational expressions. In contrast, *depression*, *anxiety*, and *sadness* exhibit highly similar distributions across all nine defense levels, reflecting a consistent and balanced defensive response.

Emotion-Specific Trajectories. Figure 2 illustrates how defense strategies evolve over the course of dialogues for seeker utterances expressing *shame* and *sadness*. While both emotions initially rise from non-defensive to more regulated responses, shame triggers a sharper early increase in neurotic defenses, followed by a steeper decline. In contrast, sadness exhibits a flatter trajectory centered around mature defenses. These patterns suggest that shame is associated with more reactive and volatile defense regulation compared to the steadier coping observed in sadness.

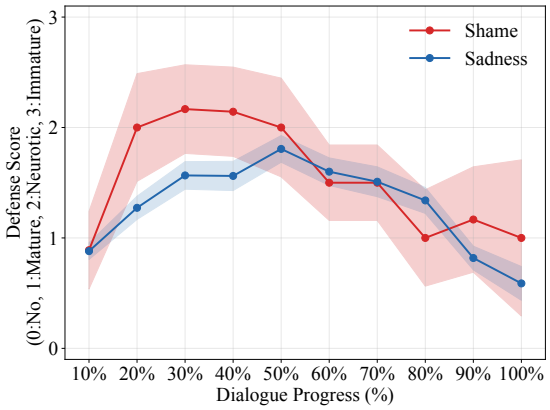

Figure 2: Average defense scores over dialogue progress for seeker utterances expressing *shame* and *sadness*. Shame shows stronger early defensiveness and greater fluctuation than sadness.

4 DMRS CO-PILOT

To support annotation, we develop DMRS CO-PILOT, an automated system that operationalizes the DMRS framework to analyze target text. We iteratively refine the system across annotator training iterations, incorporating feedback from annotators and calibration outcomes. As shown in Figure 3, the system implements a four-stage pipeline that contextualizes the target text using the background information, filters candidate defense mechanisms, validates the candidates, and synthesizes evidence into ranked recommendations. These recommendations are delivered as pre-annotation suggestions, improving efficiency and consistency.

4.1 System Architecture

We formalize the pipeline as a function Φ that maps the *target text* and its *background information* to a ranked pair of analytical conclusions. Let the background information be B, and the target text be x_t. The input to the pipeline is (B, x_t); the output is a ranked pair (C_1, C_2), where each conclusion specifies a concrete defense level from the predefined label set and provides a brief rationale that summarizes supporting evidence from the background information and the target text. The process comprises four sequential stages. We define

$$\Phi := f_4 \circ f_3 \circ f_2 \circ f_1. \quad (1)$$

Applying Φ to (B, x_t) yields

$$(C_1, C_2) = \Phi(B, x_t). \quad (2)$$

Stage 1: Stressor Identification (f_1). The first stage contextualizes the analysis by conditioning

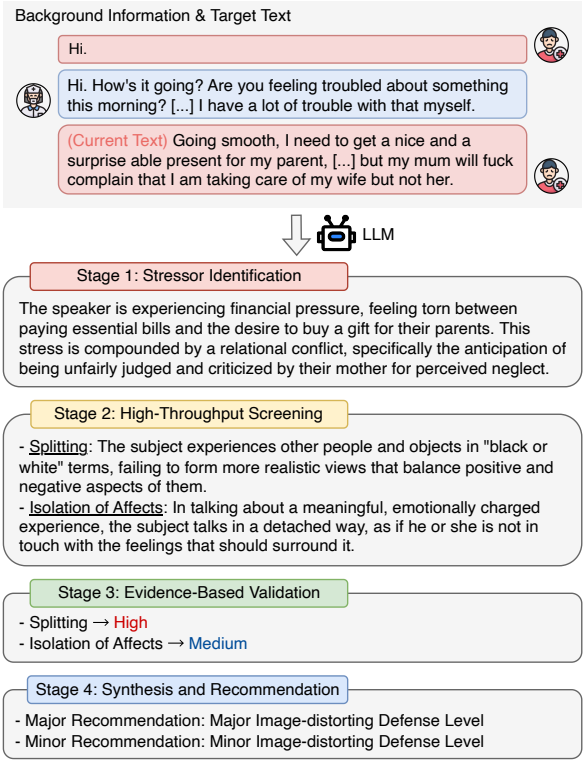

Figure 3: The four-stage analysis pipeline of DMRS Co-Pilot. The system follows a cascaded design, moving from broad contextualization and screening to deep validation and synthesis, allowing for the strategic use of different models for tasks with varying computational and reasoning requirements.

on the background information and the target text, and then identifies a candidate psychological stressor. The function f_1 takes (B, x_t) as input and outputs a natural-language hypothesis s:

$$s = f_1(B, x_t). \tag{3}$$

Stage 2: High-Throughput Screening (f_2). Let $\mathcal{I} = \{I_1, \ldots, I_{150}\}$ denote the full set of DMRS descriptive items. The screening stage reduces this set to a compact candidate subset by assessing each item's relevance to (B, x_t) conditioned on the hypothesized stressor s. The function f_2 takes (B, x_t, s) as input and returns, for each item, a binary relevance judgment and a brief rationale explaining the decision, from which the relevant subset $\mathcal{I}_{\text{rel}}$ is derived:

$$\mathcal{I}_{\text{rel}} = f_2(B, x_t, s), \qquad \mathcal{I}_{\text{rel}} \subseteq \mathcal{I}. \tag{4}$$

Stage 3: Evidence-Based Validation (f_3). For each relevant item $I_j \in \mathcal{I}_{\text{rel}}$, the function f_3 performs in-depth validation against (B, x_t), conditioned on the hypothesized stressor s. Each validated item yields a structured tuple $V_j = (c_j, e_j)$, where $c_j \in \{\text{High}, \text{Medium}, \text{Low}\}$ is a categorical confidence score and e_j is textual evidence summarizing the support in (B, x_t). The overall output is the set of validated pairs:

$$\mathcal{V} = f_3(\mathcal{I}_{\text{rel}}, B, x_t, s). \tag{5}$$

Stage 4: Synthesis and Recommendation (f_4). The final stage synthesizes $\mathcal{V}$ together with the contextual information (B, x_t) and the hypothesized stressor s to produce two ranked analytical conclusions. The function f_4 takes $(\mathcal{V}, B, x_t, s)$ as input and outputs a primary and a secondary conclusion; each conclusion consists of a selected item, a defense-level label from the predefined label set, a concise rationale, and salient relational cues grounded in (B, x_t):

$$(\mathcal{C}_1, \mathcal{C}_2) = f_4(\mathcal{V}, B, x_t, s). \tag{6}$$

4.2 Implementation Details

The pipeline's conceptual functions are implemented with large language models (LLMs), with model choice determined by each stage's needs. Functions requiring deep contextual reasoning and synthesis (f_1, f_3, f_4) use Gemini 2.5 Pro, while the high-throughput screening function (f_2) uses the more efficient Gemini 2.5 Flash for scalability. To support reproducibility and reliability, all prompts run with a low temperature of 0.2, and the outputs of f_2, f_3, and f_4 conform to a strict JSON schema for machine readability. The pipeline is engineered for robustness with parallel execution for f_2 and f_3, automated API call retries, and structured fallbacks for error handling.

4.3 System Evaluation

Operational Efficiency. To evaluate the operational efficiency of DMRS Co-Pilot, we conducted a full within-subject crossover experiment with a psychology doctoral student (zero prior DMRS experience). The experiment utilized a corpus of 126 unique samples, divided into two equal sets ($N = 63$). Each sample was annotated twice, once manually and once with the assistance of DMRS Co-Pilot, following a counterbalanced order to eliminate potential learning effects. As summarized in Table 3, the system demonstrated a significant reduction in annotation time. For Set A, the tool provided a 40.6% speed-up, reducing the average time per task from 33.10s to 19.66s. In Set B, the annotation time was consistent at 22.94s for both manual and assisted conditions. Across

Operational Efficiency (Full within-subject crossover, N=126)			
Condition	Set A (1–63)	Set B (64–126)	Mean (N=126)
Manual Only (s)	33.10	22.94	28.02
DMRS Co-Pilot (s)	19.66	22.94	21.30
Speed-up	**+40.6%**	**+0.0%**	**+24.0%**

Table 3: Efficiency evaluation of DMRS Co-Pilot. Each of the 126 unique samples was annotated twice.

the entire 126-sample corpus, the use of DMRS Co-Pilot resulted in a 24.0% overall efficiency gain, bringing the mean annotation time down from 28.02s to 21.30s. These results indicate that the system effectively streamlines the complex DMRS annotation process for expert users.

Human Agency and Bias Mitigation. To evaluate human agency and ensure that the AI-assisted workflow did not induce automation bias, we analyzed the interaction logs to determine the rejection rate of the system's recommendations. This rate represents the proportion of instances where the expert annotator rejected both final recommendations of DMRS Co-Pilot to independently identify a more clinically appropriate defense level. The results reveal a 24.0% rejection rate, providing evidence that the expert remained critically engaged with the clinical logic throughout the task. Unlike the direct-labeling frameworks investigated in recent studies where annotators frequently conform to model outputs due to an anchoring effect (Schroeder et al., 2025), our architecture exposes the reasoning chain from stressor identification to item validation. This transparency allows the expert to review intermediate evidence and manually override the final suggestions when they do not align with clinical standards. The consistent rejection of nearly one-quarter of the recommendations confirms that the workflow prevents model-driven conformity and ensures that the final annotations are a product of critically-vetted expert judgment.

Clinical Validity. To assess the clinical rigor of the system-assisted outputs, a professor of psychiatry and a doctoral student independently reviewed 45 analyses produced by DMRS Co-Pilot, rating evidence support, clinical plausibility, and insightfulness on a seven-point scale. The audit yielded mean scores of 4.62 for evidence supportiveness, 4.44 for plausibility, and 4.40 for insight, with medians of 5 across all dimensions. Notably, the system demonstrated even stronger performance in

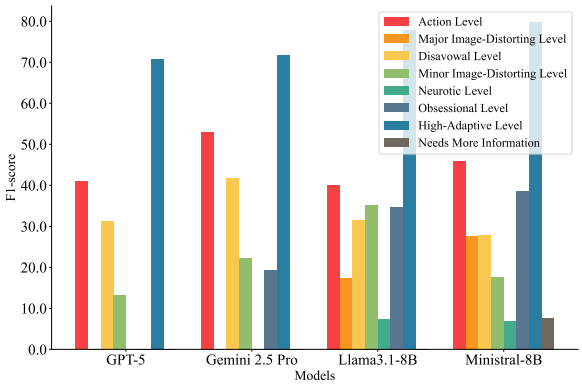

Figure 4: F1-score of the models with respect to different levels of psychological defense mechanisms.

cases involving intermediate-level defenses (Levels 4 and 5), where the mean score increased to 5.68. These findings demonstrate that DMRS Co-Pilot provides professionally-grounded clinical reasoning with significant medical reference value, transcending simple categorical labeling to offer deep, evidence-based diagnostic insights.

5 Experiments

5.1 Experimental Setup

We conduct experiments in both zero-shot and fine-tuning settings using state-of-the-art LLMs, with the prompt illustrated in Appendix E. For zero-shot prompting, we experiment with GPT-5[1], GPT-5 mini, Gemini 2.5 Pro (Comanici et al., 2025), Kimi-K2 (`0905`, Bai et al., 2025b), DeepSeek-V3.2 (Liu et al., 2025), and Qwen3-Next (Yang et al., 2025), where the default/recommended hyperparameters are adopted. We also fine-tune models including Llama 3.1-8B (Grattafiori et al., 2024), Ministral-8B[2], GLM-4-9B (Zeng et al., 2024), Qwen3-4B, Qwen3-8B, and InternLM3-8B (Cai et al., 2024b). During the fine-tuning process, we set the number of epochs to 10, the batch size to 1, the gradient accumulation step to 8, and the learning rate to 1e-4. To ensure a consistent evaluation, the dataset is partitioned using an 80/20 stratified utterance-level split. Performance is evaluated using accuracy, along with macro precision, recall, and F1-scores, computed both over positive classes (1-8) and all classes, in line with established practices for multi-class classification (Wang et al., 2022, 2024b; Chen et al., 2025).

[1] https://openai.com/index/gpt-5-system-card/
[2] https://mistral.ai/news/ministraux

Model	ACC	P	R	F1
Zero-shot Prompted LLMs				
GPT-5	52.75	27.59	16.56	19.53
GPT-5 mini	54.03	26.30	16.99	18.41
Gemini 2.5 Pro	**56.36**	27.49	26.12	25.99
Kimi-K2	41.10	21.28	21.38	17.58
DeepSeek-V3.2 (w/o)	39.83	24.77	19.97	16.15
DeepSeek-V3.2 (w/)	55.72	**29.66**	**27.53**	**26.17**
Qwen3-Next (w/o)	40.89	26.91	19.09	14.88
Qwen3-Next (w/)	44.28	27.23	21.58	20.68
Fine-tuned LLMs				
Llama3.1-8B	62.92	33.24	30.08	30.51
Ministral-8B	**64.83**	**33.97**	**30.45**	**31.48**
GLM-4-9B	62.92	30.10	29.53	28.61
Qwen3-4B	60.59	30.49	28.53	28.46
Qwen3-8B	61.44	30.10	28.91	28.39
InternLM3-8B	63.98	33.47	29.93	30.53

Table 4: Overall performance of the evaluated models on the PSYDEFCONV dataset over *positive classes*. "w/o" and "w/" denote the thinking process is disabled/enabled. The best and the second-best performance in each setup are highlighted in **bold** and underlined, respectively.

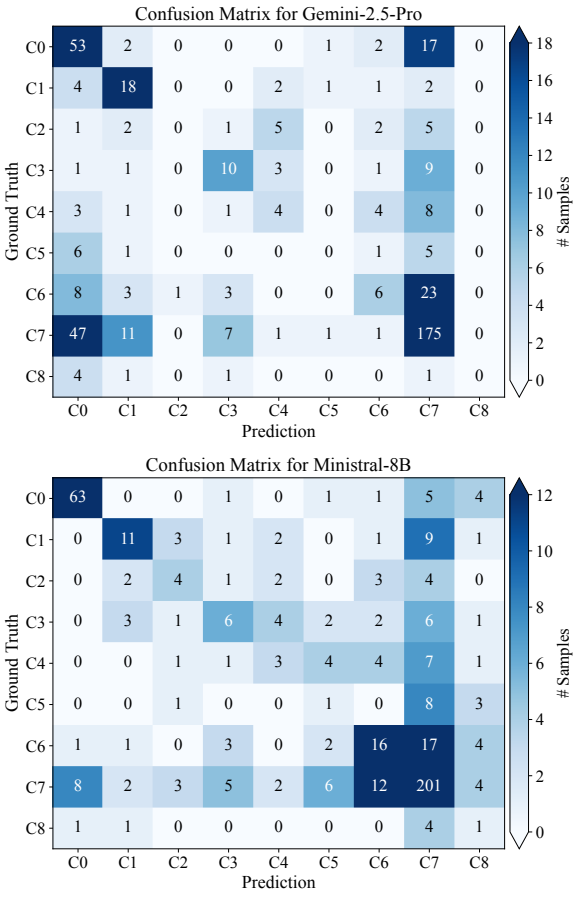

Figure 5: Confusion matrices of Gemini 2.5 Pro and Ministral-8B on the PSYDEFCONV dataset.

5.2 Experimental Results

Tables 4 and 6 summarize the overall performance of the models on the PSYDEFCONV dataset. Among the zero-shot prompted models, Gemini 2.5 Pro and DeepSeek-V3.2 demonstrate superior performance, while Ministral-8B and InternLM3-8B emerge as the top-performing fine-tuned models. Despite these results, all models exhibit unsatisfactory effectiveness, as even the best fine-tuned macro F1-score reaches only approximately 30. This limitation is primarily due to the dataset's inherent imbalance, which poses significant challenges in accurately detecting fine-grained psychological defense mechanisms in conversations. Notably, the thinking capabilities of DeepSeek-V3.2 and Qwen3-Next both yield performance improvements, with DeepSeek-V3.2 achieving a substantial increase of approximately 40% in accuracy and 62% in macro F1-score.

Figures 4 and 5 illustrate the F1-scores across different psychological defense levels and the confusion matrices of the model predictions. Both zero-shot prompted and fine-tuned models exhibit a strong bias toward predicting the *High-Adaptive level*, along with the action, disavowal, and minor image-distorting levels. Notably, many samples from other levels, particularly the action level and obsessional level, are frequently misclassified as High-Adaptive level. Although fine-tuning improves the model's ability to predict underrepresented levels (e.g., neurotic level), the dataset's inherent imbalance exacerbates the skew toward the dominant High-Adaptive level. Furthermore, "needs more information" remains the most challenging class to predict, with Llama3.1-8B performing especially poorly, as it fails to accurately classify any sample into this category.

6 Conclusion and Future Work

We study defensive functioning in supportive conversations by aligning clinical theory with what is observable in dialogue. We annotate DMRS levels rather than mechanisms, release PSYDEFCONV with substantial agreement, and provide DMRS CO-PILOT to support efficient and consistent labeling. Benchmarks indicate that the task remains challenging, with clear headroom and a tendency to overpredict mature defenses. Our dataset and tools offer a reproducible basis for future work and encourage models that use richer context and theory-aware supervision.

Moving forward, we aim to develop methods to mitigate biases toward majority defense levels and investigate how defense mechanisms evolve across multi-turn therapeutic interactions. Ultimately, we hope to leverage these insights to build generative models capable of delivering more adaptive and theory-aware emotional support.

Limitations

The primary limitations of our work include the language being limited to English and the skewed distribution of instances at the High-Adaptive level, reflecting the natural prevalence of psychological defense mechanisms in real-world contexts. While these limitations do not diminish our contributions, we encourage future practitioners to (1) extend the annotation to additional languages, such as Chinese, Japanese, and Spanish; and (2) generate synthetic dataset samples to mitigate the imbalance. Both directions can be effectively facilitated through the proposed DMRS CO-PILOT framework.

Ethical Considerations

Licenses. We use conversations from ESConv (Liu et al., 2021) under its stated terms. Our release contains only derived annotations, document identifiers, and code. We do not redistribute the original dialogues beyond what the ESConv license allows. Downstream users must obtain ESConv separately and agree to its conditions. Our annotations are released for research use and require compliance with the ESConv license.

Annotator Details. The two primary annotators are co-authors based in the United States and Australia. The within-subject timing study was conducted by a co-author based in Australia who is a psychology doctoral student. Independent expert evaluation of the system's clinical validity was performed by a professor of psychiatry and a doctoral student based in the United States; these experts were compensated at a rate of 20 USD per hour. All are adult researchers fluent in English. These activities involved only members of the research team and independent experts working with de-identified, publicly available text; participation was voluntary and could be discontinued at any time.

Ethics Review. This work involves secondary analysis of publicly available dialogue data and new annotations created by trained raters. No personal identifying information is collected or inferred. Because the content concerns mental health, we provided clear guidance on breaks and access to support resources. The protocol was reviewed by the ethics committee of Xi'an Jiaotong-Liverpool University and determined to be exempt from ethics review, as it involves only secondary analysis of publicly available datasets and does not include primary data collection.

Societal Impact. The resource and models are intended for research on language and defensive functioning. They are not a diagnostic tool and should not be used to make clinical or legal decisions about individuals. Misuse risks include stigma, unwarranted profiling, or surveillance outside supportive contexts. The dataset is in English and reflects the domains covered by ESConv, so results may not generalize across cultures or clinical settings. We encourage users to report limitations, check for bias, and avoid normative claims about people or groups.

Acknowledgments

We would like to thank Shumao Yu for the insightful discussions and Zac Imel for his expert evaluation and constructive suggestions. We also gratefully acknowledge the support of the Label Studio Academic Program for providing access to the annotation platform used in this study. This work was partially supported by ARC LP 240200698.

References

Xin Bai, Guanyi Chen, Tingting He, Chenlian Zhou, and Yu Liu. 2025a. Emotional supporters often use multiple strategies in a single turn. *Preprint*, arXiv:2505.15316.

Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, and 149 others. 2025b. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024a. EmpCRL: Controllable empathetic response generation via in-context commonsense reasoning and reinforcement learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5734–5746, Torino, Italia. ELRA and ICCL.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024b. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Tong Chen, Zimu Wang, Yiyi Miao, Haoran Luo, Yuanfei Sun, Wei Wang, Zhengyong Jiang, Procheta Sen, and Jionglong Su. 2025. MedFact: A large-scale Chinese dataset for evidence-based medical fact-checking of LLM responses. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32340–32353, Suzhou, China. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

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*, Volume 12 - 2021.

Sigmund Freud. 1936. Inhibitions, symptoms and anxiety. *The Psychoanalytic Quarterly*, 5(1):1–28.

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.

Yining Hua, Hongbin Na, Zehan Li, Fenglin Liu, Xiao Fang, David Clifton, and John Torous. 2025a. A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, 8(1):230.

Yining Hua, Steve Siddals, Zilin Ma, Isaac Galatzer-Levy, Winna Xia, Christine Hau, Hongbin Na, Matthew Flathers, Jake Linardon, Cyrus Ayubcha, and 1 others. 2025b. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World Psychiatry*, 24(3):383–394.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261, Bangkok, Thailand. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, and 180 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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.

Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025. Detecting conversational mental manipulation with intent-aware prompting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9176–9183, Abu Dhabi, UAE. 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.

J Christopher Perry and Melissa Henry. 2004. Studying defense mechanisms in psychotherapy using the defense mechanism rating scales. *Advances in psychology*, 136:165–192.

J Christopher Perry, Marianne E Kardos, and Christopher J Pagano. 1993. The study of defenses in psychotherapy using the defense mechanism rating scales (dmrs). *The concept of defense mechanisms in contemporary psychology: Theoretical, research, and clinical perspectives*, pages 122–132.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2025. Just put a human in the loop? investigating LLM-assisted

annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25771–25795, Vienna, Austria. Association for Computational Linguistics.

Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large language models and empathy: Systematic review. *J Med Internet Res*, 26:e52597.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

George E Vaillant. 2012. *Adaptation to life*. Harvard University Press.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024a. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 470–475, Bangkok, Thailand. Association for Computational Linguistics.

Zimu Wang, Hongbin Na, Rena Gao, Jiayuan Ma, Yining Hua, Ling Chen, and Wei Wang. 2025. From posts to timelines: Modeling mental health dynamics from social media timelines with hybrid LLMs. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 249–255, Albuquerque, New Mexico. Association for Computational Linguistics.

Zimu Wang, Lei Xia, Wei Wang, and Xinya Du. 2024b. Document-level causal relation extraction with knowledge-guided binary question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955, Miami, Florida, USA. Association for Computational Linguistics.

Donald W Winnicott. 2018. Ego distortion in terms of true and false self. In *The person who is me*, pages 7–22. Routledge.

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.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025. SweetieChat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4646–4669, Abu Dhabi, UAE. Association for Computational Linguistics.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, and 38 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

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.

Xiangyu Zhao, Yaling Shen, Yiwen Jiang, Zimu Wang, Jiahe Liu, Maxmartwell H Cheng, Guilherme C Oliveira, Robert Desimone, Dominic Dwyer, and Zongyuan Ge. 2025. It hears, it sees too: Multimodal llm for depression detection by integrating visual understanding into audio language models. *Preprint*, arXiv:2511.19877.

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. *Preprint*, arXiv:2308.11584.

A Annotation Handbook: Operational Protocol

This appendix summarizes the coding protocol used to annotate defense functioning in supportive dialogues. It is designed to ensure consistency, traceability, and reproducibility. Definitions of individual defenses are omitted here.

A.1 Scope and Units

- **Target of annotation** The help seeker's utterances only.
- **Unit** One utterance at a time. An utterance may contain multiple sentences if they form a single turn.
- **Context window** Use only the dialogue prior to and including the target utterance. Do not use future turns.
- **Label set** DMRS levels 1 to 7. Two auxiliary labels are included: 0 for no defense and 8 for needs more information.

A.2 Core Principles

- **Primacy of context** Judge function with respect to the preceding dialogue.
- **Function over form** Ask what the utterance achieves for the speaker in relation to stress or conflict.
- **Emotion is not a defense** Pure feeling statements are not defenses unless there is clear avoidance, distortion, or transformation.
- **Acknowledge mature coping** Mark adaptive responses when they are supported by local evidence.

A.3 Workflow

- Read the prior context and the target utterance. Form a brief hypothesis of the salient stressor.
- Decide whether the target utterance is a neutral or phatic act. If so, assign 0 and proceed to the next item.
- If the utterance cannot be judged with the available context, assign 8.
- Otherwise select one primary level from 1 to 7 that best explains the utterance's function.
- Optionally record one secondary candidate level if evidence is close.
- Extract minimal evidence spans from the text. Write a one to two sentence rationale that links evidence to the level choice.
- Record any uncertainties or edge cases in the notes field.

A.4 Evidence and Rationale

- Quote only what is necessary. Prefer short spans over paraphrase.
- Ground each claim in the target utterance or its immediate context.
- Avoid inferences about stable traits unless the dialogue explicitly supports them.
- When evidence is weak, state why and consider label 8.

A.5 Disambiguation Rules

- **Multiple signals in one utterance** Choose the dominant function. If two are truly inseparable, prefer the lower maturity level or mark 8 when evidence is insufficient.
- **Mature coping vs phatic closings** Simple thanks, greetings, and farewells are label 0. Rich reflections on received help may support a mature level if explicitly stated.
- **Negative talk about others** Require signs of self esteem protection or blame shifting. Factual criticism alone is not a defense.
- **Intellectual style** Distinguish descriptive or analytic language from efforts to keep feelings at a distance. The latter supports an obsessional level when grounded in text.
- **Across time requirements** Signals that demand longitudinal evidence are rarely observable in short dialogues. Do not infer them without explicit cues in the current context.

A.6 Typical Label 0 Situations

- Greetings and farewells.
- Simple thanks or brief acknowledgments such as yes, no, okay.
- Logistical or factual questions and answers.

- Neutral small talk and bodily states.
- Clarifications of meaning without defensive intent.

A.7 Quality Control

- Two trained annotators label independently and blind to each other.
- Use a shared handbook and examples for calibration. Hold regular reviews to refine decision rules.
- Compute agreement statistics on double labeled items. Resolve disagreements by consensus to create gold labels.
- Maintain an audit trail with conversation id, utterance id, primary level, secondary level when used, stressor hypothesis, evidence spans, rationale, and notes.

A.8 Use with DMRS CO-PILOT

- Pre annotations include a stressor hypothesis, candidate items, evidence summaries, and two ranked level suggestions.
- Annotators must verify or revise suggestions based on the text. Record whether the suggestion was helpful.
- When suggestions overreach beyond local evidence, prefer a conservative label or 8.

B Further Data Statistics

Table 5 presents a comprehensive statistical overview of the PSYDEFCONV dataset. Among the 200 seekers, the most prevalent presenting problems are ongoing depression (29.0%) and job crises (24.5%), with anxiety (28.5%) and depression (27.5%) being the dominant expressed emotions. Regarding the defense mechanism annotations ($N = 2,336$), Level 7 (High-Adaptive) is the most frequent level, accounting for 51.8% of the data. When aggregated into broader categories, Mature Defenses represent the majority (51.8%), followed by Immature Defenses (18.9%) and No Defense (17.4%), reflecting a diverse spectrum of psychological coping strategies within the dataset.

Categories	Num	Proportion
Seeker's Problems		
Ongoing Depression	58	29.0%
Job Crisis	49	24.5%
Breakup with Partner	42	21.0%
Problems with Friends	26	13.0%
Academic Pressure	25	12.5%
Overall	200	100.0%
Seeker's Emotions		
Anxiety	57	28.5%
Depression	55	27.5%
Sadness	50	25.0%
Fear	18	9.0%
Anger	15	7.5%
Shame	5	2.5%
Overall	200	100.0%
Seeker's Defense Levels		
0 No Defense	371	15.9%
1 Action Level	136	5.8%
2 Major Image-Distorting Level	77	3.3%
3 Disavowal Level	124	5.3%
4 Minor Image-Distorting Level	105	4.5%
5 Neurotic Level	61	2.6%
6 Obsessional Level	216	9.2%
7 High-Adaptive Level	1,211	51.8%
8 Needs More Information	35	1.5%
Overall	2,336	100.0%
Seeker's Defense Categories		
No Defense (0, 8)	406	17.4%
Mature Defenses (7)	1,211	51.8%
Neurotic Defenses (5, 6)	277	11.9%
Immature Defenses (1, 2, 3, 4)	442	18.9%
Overall	2,336	100.0%

Table 5: Distribution of the PSYDEFCONV dataset across seekers' presenting problems, expressed emotions, annotated defense levels, and aggregated defense categories.

C Further Data Analysis

Heatmaps of Defense Distributions. Figure 6 shows that mature defenses (Level 7) dominate across all emotions, indicating a general preference for adaptive coping. However, variations emerge across emotion types. Shame exhibits the highest proportion of low-level defenses (Levels 1–4), suggesting more fragmented and immature regulation. Anger stands out for its elevated use of Level 0 (no defense), implying more unfiltered or confrontational expressions. In contrast, *depression*, *anxiety*, and *sadness* exhibit highly similar distributions across all nine defense levels, reflecting a consistent and balanced defensive response. Turning to problem contexts, Figure 7 reports row normalized

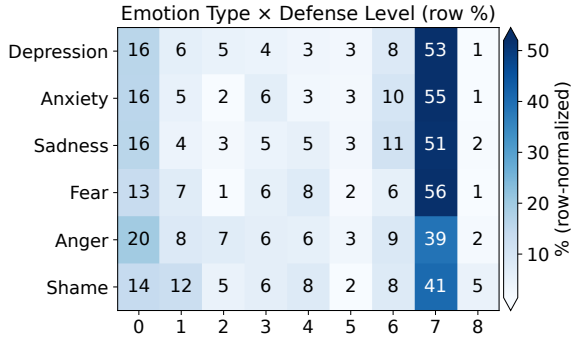

Figure 6: Defense levels by emotion. Level 7 dominates. Shame shows more low levels. Anger shows more Level 0. Depression anxiety and sadness are similar.

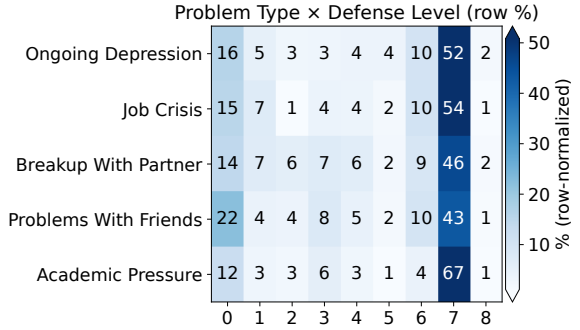

Figure 7: Defense levels by problem context. Level 7 dominates. Academic pressure is highest at Level 7. Problems with friends shows more Level 0. Breakup with partner is higher at Levels 2 to 4.

percentages that mirror the dominance of Level 7 across contexts, with academic pressure showing the largest share at Level 7 and the smallest shares at low levels, problems with friends showing the highest share at Level 0, breakup with partner concentrating more mass in Levels 2 to 4 than other contexts, while ongoing depression and job crisis align closely with the overall pattern.

Progress of Defense Strategies Across Emotions and Problem Contexts. Figure 8 and Figure 9 trace average defense scores across dialogue progress from 10 percent to 100 percent. On the vertical axis larger values indicate more immature and neurotic defenses while lower values indicate mature or no defense. Across emotions and problem contexts the curves generally rise in the early to middle stages then return toward more regulated responding by the end. Shame shows the sharpest early ascent followed by a clear decline. Sadness depression and anxiety follow similar and smooth paths near mature defenses. Anger moves toward the lowest scores near the closing turns which suggests more direct expression with minimal defensive filtering. By context job crisis shows the most pronounced arch with a high middle peak. Ongoing depression and problems with friends peak in the middle with moderate amplitude. Breakup with partner climbs early then holds a plateau before easing. Academic pressure varies the least and ends with a clear return toward mature defenses. These patterns highlight the middle of the dialogue as a critical window for defense regulation and complement the findings in the main text.

Model	P	R	F1
Zero-shot Prompted LLMs			
GPT-5	28.01	23.75	22.40
GPT-5 mini	27.22	23.40	21.62
Gemini 2.5 Pro	<u>29.07</u>	<u>31.07</u>	<u>28.93</u>
Kimi-K2	21.84	23.89	19.29
DeepSeek-V3.2 (w/o)	25.53	24.41	18.96
DeepSeek-V3.2 (w/)	**31.38**	**32.76**	**29.51**
Qwen3-Next (w/o)	27.97	25.12	18.63
Qwen3-Next (w/)	27.74	27.48	23.34
Fine-tuned LLMs			
Llama3.1-8B	39.29	<u>36.22</u>	36.73
Ministral-8B	**39.78**	**36.40**	**37.45**
GLM-4-9B	36.74	35.43	35.00
Qwen3-4B	36.32	34.70	34.57
Qwen3-8B	36.46	34.89	34.67
InternLM3-8B	<u>39.77</u>	36.09	<u>36.88</u>

Table 6: Overall performance of the evaluated models on the PSYDEFCONV dataset across *all classes*. "w/o" and "w/" denote the thinking process is disabled/enabled. The best and the second-best performance in each setup are highlighted in **bold** and <u>underlined</u>, respectively.

D HumanSignal Annotation Interface

Figure 10 presents the Humansignal annotation screen on page 1. The left column shows the conversation context with seeker and supporter turns. The right panel states the task and highlights the sentence to annotate. Annotators choose a defense level from 0 to 8 where 0 is no defense and 7 is highly adaptive while 8 denotes need more information. The panel also includes structured fields for hypothesized stressor primary conclusion secondary conclusion and validated items. This figure documents the data collection workflow used to generate defense labels and complements the quantitative figures by showing the exact labeling environment.

E Prompt for Experiments

To evaluate the zero-shot performance of LLMs on the PSYDEFCONV dataset, we designed a struc-

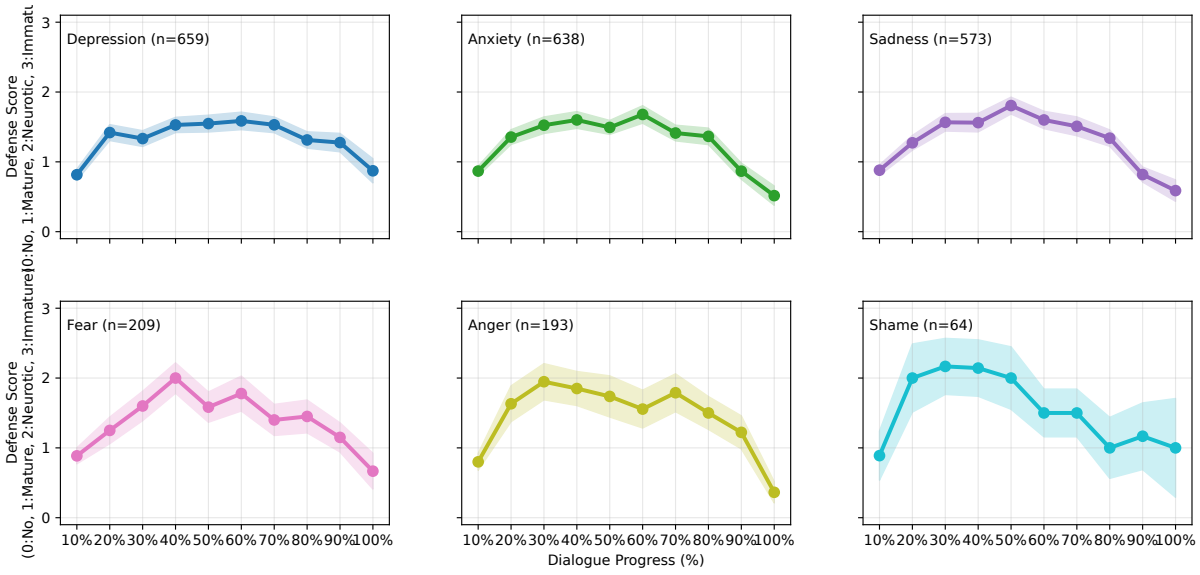

Figure 8: Defense score trajectories across dialogue progress by emotion. Each curve shows the mean score with a shaded band that indicates variability. Higher scores indicate more immature and neurotic defenses and lower scores indicate mature or no defense. Shame rises early then declines. Anger ends near the lowest scores. Sadness depression and anxiety follow closely aligned paths.

tured prompt that provides the model with the necessary DMRS level definitions and the dialogue context.

F DMRS Co-Pilot System Prompts

To ensure interpretability and replicability of the DMRS CO-PILOT system, we include in this appendix the full text of the four prompting templates used across the analysis pipeline. Each stage in the pipeline corresponds to one dedicated prompt that operationalizes the respective analytical function f_1–f_4 described in Section 4.1.

Stage 1 (*Stressor Identification*) elicits a concise hypothesis about the psychological stressor expressed in the target utterance. Stage 2 (*High-Throughput Screening*) filters the complete DMRS item set to a small subset of candidates relevant to the identified stressor. Stage 3 (*Evidence-Based Validation*) performs fine-grained item-level validation, assessing evidential alignment and confidence. Finally, Stage 4 (*Synthesis and Recommendation*) integrates validated evidence into two ranked analytical conclusions that form the system's final output.

The following figures reproduce these prompts verbatim as used in the deployed system.

G Experimental Results on All Classes

The experimental results, computed across the full label space, are presented in Table 6.

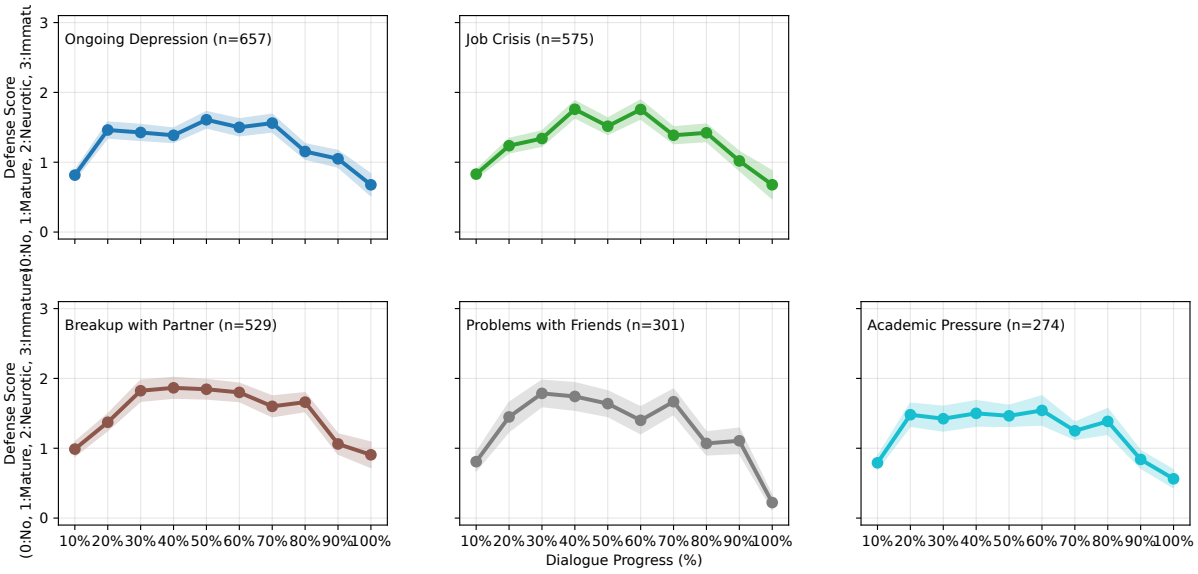

Figure 9: Defense score trajectories across dialogue progress by problem context. Most contexts peak in the middle then return to lower scores at the end. Job crisis shows the clearest arch. Breakup with partner rises early then holds a plateau before easing. Ongoing depression and problems with friends show moderate peaks. Academic pressure varies little and ends lower than the middle.

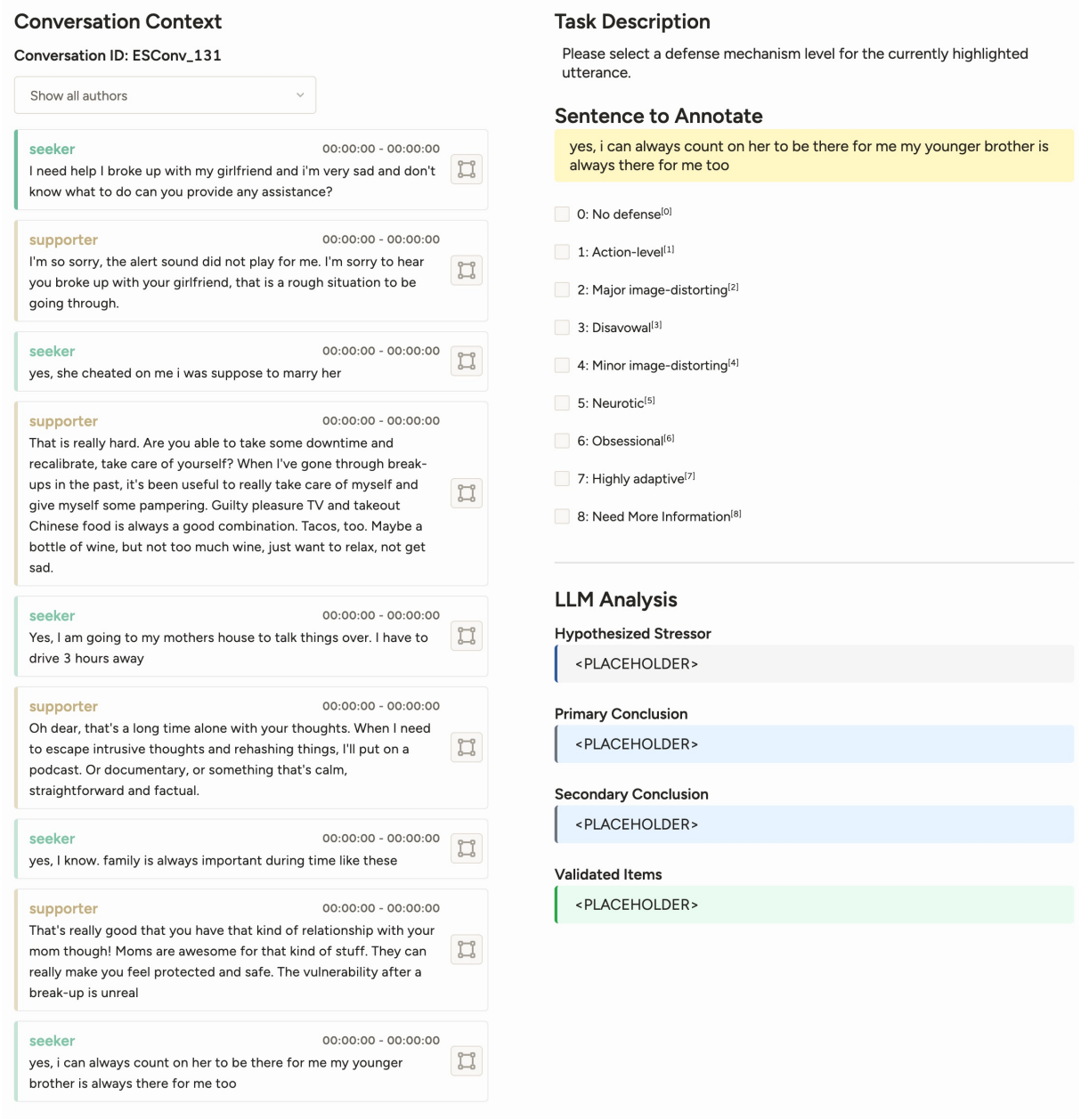

Figure 10: HumanSignal interface for defense level annotation. Left column shows the full dialogue. Right panel prompts the rater to select a level from 0 to 8 and provides fields for analysis and validation.

Prompt for Experiments

```
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).
You may reason internally, but the final answer must follow the required output format.

- Dialogue context:
{conversation}

- Target utterance:
{current_text}

Available labels (Defense Mechanism Rating Scale tiers):
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 (Evidence triggers suspicion of a defense but is insufficient to confirm any
tier)

Return exactly one line:
label: <0-8>
No additional content is allowed.
```

Figure 11: Prompt used for experiments, which instructs LLMs to determine the specific level of psychological defense mechanism based on the dialogue context and the target utterance.

DMRS Co-Pilot Stage 1: Stressor Identification Prompt (f_1)

```
You are a psychological analysis expert. Your task is to first objectively describe the primary function
 and emotional content of the target utterance, then determine if it reveals any psychological stressor.

Target Utterance (primary focus): "{target_utterance}"
Dialogue Context (reference only): "{dialogue_context}"

Step 1: Describe the primary function of this utterance (e.g., greeting, factual statement, emotional
expression, request, etc.)

Step 2: Determine IF the utterance expresses any discernible pressure, conflict, or psychological
tension. Many utterances, especially simple social conventions (like greetings) or neutral statements,
do not contain psychological stressors.

Step 3: If and only if you identified a genuine stressor in Step 2, describe what that stressor is in
1-2 sentences. If no stressor was identified, explicitly state: "No specific psychological stressor
identified in this utterance."

Important: Simple greetings, basic social exchanges, and neutral factual statements should typically
result in "No specific psychological stressor identified" unless there is clear evidence of underlying
distress in the specific wording or context.

Provide your final answer: Either a concise description of the identified stressor, or "No specific
psychological stressor identified in this utterance."
```

DMRS Co-Pilot Stage 2: High-Throughput Screening Prompt (f_2)

```
You are a meticulous DMRS screening analyst. Your primary responsibility is to conduct a rigorous
precise analysis to ensure only high-quality candidates are passed to the expert review stage.

**Analysis Context:**
- Target Utterance: "{target_utterance}"
- Dialogue Context: "{dialogue_context}"
- Identified Stressor: "{stressor_text}"

**DMRS Item to Analyze:**
- Item ID: {item['item_id']}
- Defense Name: {item['defense_name']}
- Level: {item['level_name']}
- Description: {item['item_description']}

**Relevance Assessment:**
- **Relevant**: The target utterance shows some connection to the DMRS item description
- **Not Relevant**: The target utterance shows no connection to the DMRS item description

Provide a JSON object with exactly these fields:
- item_id: {item['item_id']}
- relevance: "Relevant" or "Not Relevant"
- brief_analysis: string (MUST follow this micro-format within the string: Start with "Matching Points:
" to describe specific aspects that align with the DMRS item description - analyze word choices,
emotional tone, psychological patterns, and functional purposes step by step. Then start with "Non-
matching Points: " to describe specific aspects that do not align or are missing - be thorough and
analytical. Think carefully point by point about the psychological mechanisms involved.)

**Critical Instructions:**
1. Analyze step-by-step, examining every aspect of the target utterance against the DMRS item
description
2. Use the micro-format strictly: Matching Points: [matching aspects] Non-matching Points: [non-matching
 aspects]
3. Be thorough and precise in your analysis
```

Figure 12: Prompt used in Stage 1 (*Stressor Identification*). This stage identifies whether the target utterance expresses any psychological pressure, conflict, or tension, yielding a concise hypothesis of the potential stressor that contextualizes subsequent analysis. And prompt used in Stage 2 (*High-Throughput Screening*). This stage performs large-scale relevance filtering of all DMRS items against the identified stressor, producing a compact subset of candidate items with accompanying brief rationales.

DMRS Co-Pilot Stage 3: Evidence-Based Validation Prompt (f_3)

```
You are an expert psychological analyst conducting a rigorous validation of defense mechanisms. Your
task is to perform a strict, evidence-based verification of whether the target utterance truly
demonstrates this specific DMRS item.

**Context:**
- Target Utterance: "{target_utterance}"
- Dialogue Context: "{dialogue_context}"
- Hypothesized Stressor (assumed correct): "{stressor_text}"

**Defense Mechanism: {defense_name}**
- Level: {level_name}

**DMRS Item Being Validated:**
- Item ID: {item_id}
- Description: {item_description}

**Analysis Requirements:**

1. **Points of Conformance Analysis**:
   - Examine how the target utterance aligns with key elements in this DMRS item
   - Be specific about which aspects of the utterance match the item description
   - Consider word choice, emotional tone, functional purpose, and psychological mechanisms

2. **Flaws & Points of Inference Analysis**:
   - Identify evidence that requires inference rather than direct observation
   - Note any contradictions or inconsistencies
   - Identify missing core elements that would strengthen the case
   - Consider alternative explanations for the behavior
   - Assess degree of speculation required

3. **Confidence Assessment**:
   - **High**: Clear, direct evidence with minimal inference required
   - **Medium**: Reasonable evidence but requires some inference or has minor gaps
   - **Low**: Weak evidence, requires significant inference, or has major gaps
   - **Not Applicable**: The utterance shows no discernible connection to the DMRS item, even with
   significant inference

4. **Justification Synthesis**:
   - Briefly synthesize the most critical arguments from your conformance and flaws analyses.
   - Conclude with a final, decisive reason for your chosen confidence level.

**Required JSON Output Format:**
Provide a JSON object with exactly these fields:
- item_id: {item_id}
- defense_name: "{defense_name}"
- level_name: "{level_name}"
- item_description: "{item_description}"
- confidence: "High" or "Medium" or "Low" or "Not Applicable"
- points_of_conformance: string (detailed analysis of matching aspects)
- flaws_and_inference: string (critical analysis of gaps and required inferences)
- justification: string (brief 1-2 sentence justification for the confidence rating)

Remember: Be rigorous and critical. Do not overstate the evidence.
```

Figure 13: Prompt used in Stage 3 (*Evidence-Based Validation*). This stage validates each candidate DMRS item through detailed comparison with the target utterance, assessing evidential conformance, inferential gaps, and confidence levels for each item.

DMRS Co-Pilot Stage 4: Synthesis and Recommendation Prompt (f_4)

```
You are an expert psychological analyst performing a final item-centric analysis of defense mechanisms
based on validated evidence.

**Context:**
- Target Utterance: "{target_utterance}"
- Dialogue Context: "{dialogue_context}"
- Hypothesized Stressor: "{hypothesized_stressor}"

**Evidence Summary:**
- Total applicable items: {statistics['applicable_items']}
- High confidence: {statistics['high_confidence']}
- Medium confidence: {statistics['medium_confidence']}
- Low confidence: {statistics['low_confidence']}

**All Validated Items:**
{items_json}

**Your Analytical Task:**
You must perform a unified analysis that starts from identifying the most critical evidence (key items)
and evaluates each independently. Your analysis should produce exactly TWO conclusions.

**Understanding "No Defense" Situations:**
Before analysis, recognize that many utterances contain NO defense mechanisms. These include:
- Simple greetings or thanks ("Hi", "Thank you")
- Basic responses ("Yes", "No", "OK")
- Factual event reporting or objective descriptions
- Information-seeking questions ("When is our next appointment?")
- Social small talk (weather, traffic)
- Direct physical sensations ("I feel cold")
- Genuine clarification requests ("Sorry, can you explain that again?")
- Objective personal information ("I'm 30 years old")

**Step 1: Identify and Evaluate the Top 1 Item**
- Find the most prominent item with the strongest evidence
- Independently evaluate this item's quality:
  - If evidence is solid (e.g., High or clear Medium confidence with strong context): Recommend this
  specific defense
  - If this is the only evidence but it's ambiguous: Conclude "Need more information" (because even the
  strongest signal is unclear)
  - If no items exist OR the utterance falls into the "No Defense" categories above: Conclude "No
  defense"

**Step 2: Identify and Evaluate the Top 2 Item**
- Find the second most prominent item
- Independently evaluate its quality:
  - If evidence is clear: Recommend this specific defense
  - If evidence is ambiguous or weak: Conclude "Need more information" (noting there's a secondary
  signal about [defense name] that needs clarification)

**Critical Guidelines:**
1. Each conclusion must be independent - don't force recommendations if evidence is weak
2. "Need more information" is a valid primary or secondary conclusion when evidence is ambiguous
3. Be specific about WHY you're concluding what you're concluding
4. Don't mechanically follow confidence levels - evaluate the actual evidence quality
5. **IMPORTANT**: Your rationale MUST include:
   - Analysis of how the DMRS item description relates to the target utterance
   - Direct quotes from the target utterance as evidence (use quotation marks)
   - Connection to the dialogue context and hypothesized stressor
   - Reference to specific elements from the item description that match/don't match the utterance

**Required JSON Output Format:**
Provide your analysis as a JSON object with exactly this structure:
{JSON Output Format}

Remember: Your goal is to provide the most accurate and honest assessment of the evidence, not to force
defense recommendations where evidence is weak.
```

Figure 14: Prompt used in Stage 4 (*Synthesis and Recommendation*). This final stage integrates all validated evidence to produce two independent, ranked analytical conclusions—each comprising a defense-level label, rationale, and contextual justification grounded in the dialogue.