

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

Tamjid Hasan Fahim, Syed Asif Johan, Saad Bin Maksud

Department of Computer Science & Engineering
Rajshahi University of Engineering and Technology
Rajshahi-6204, Bangladesh

{tamjidhfahim, connect.syedasifjohan, sbinmaksud}@gmail.com

Abstract

Automating the classification of psychological defense mechanisms is a critical yet challenging frontier in clinical natural language processing. General-purpose Large Language Models (LLMs) struggle to apply fine-grained ordinal frameworks like the Defense Mechanism Rating Scales due to the implicit nature of clinical cues and a fundamental clinical reasoning gap. These models exhibit severe extreme response bias, systematically gravitating toward the scale’s endpoints while failing to resolve nuanced, mid-level defenses. In this paper, we present our third-place system for the PSYDEFDETECT Shared Task at BioNLP 2026, designed specifically to overcome this failure mode. We propose a hybrid architecture that synergizes label-flattened generative retrieval with an LLM classifier fine-tuned via the distillation of supervised clinical reasoning traces. This dual approach, grounding decisions in rubric criteria while leveraging task-specific supervision, successfully mitigates the observed bias, achieving an accuracy of 67.37% and a macro-F1 of 39.56%. Our work provides empirical evidence that tightly integrating targeted clinical supervision with dynamic rubric-grounded retrieval significantly outperforms the raw parameter scale of un-tuned foundation models.

1 Introduction

Understanding how individuals psychologically defend against emotional distress is central to psychotherapeutic practice. The *Defense Mechanism Rating Scales* (DMRS; Di Giuseppe and Perry 2021) provide a validated framework for categorizing these unconscious processes across seven ordinal maturity levels (1–7). While essential for clinical assessment, manual scoring requires specialized training and does not scale to the large volumes of data generated in digital mental health interventions.

The PSYDEFDETECT Shared Task at BioNLP 2026 (Na et al., 2026a) formalizes the challenge of automating this process. This task presents three fundamental challenges. First, defense mechanisms exhibit subtle overlap between adjacent levels, where labels share significant semantic and pragmatic features, challenging standard classifiers that treat them as discrete categories rather than an ordered spectrum. Second, general-purpose LLMs exhibit a clinical reasoning gap. They lack the domain-specific reasoning required to apply the rigid criteria of a professional rubric, leading to a disconnect between surface-level sentiment and the underlying unconscious processes. Finally, defense mechanisms are expressed through implicit and context-dependent cues, resulting in a low signal-to-noise ratio. Accurate classification therefore depends less on explicit lexical indicators and more on latent patterns.

In this paper, we present a highly competitive hybrid system designed to address these inherent challenges, securing third place among 21 registered teams in the official shared task evaluation. We introduce a teacher-student distillation framework that encodes structured chain-of-thought clinical reasoning into a smaller LLM, eliminating the need for long-context rubric-handbook prompting at inference time. We further augment this supervised fine-tuning (SFT) pipeline with a dynamic DMRS-Q item retrieval framework to ground predictions in the rubric’s nuanced criteria. Crucially, by flattening the clinical rubric during retrieval and removing ordinal identifiers from the prompt and output space, the system mitigates the model’s tendency toward extreme response bias. Ultimately, our results demonstrate that automating nuanced psychological assessment depends less on continuously scaling foundation models and more on tightly integrating targeted clinical supervision with dynamic, rubric-grounded retrieval.

2 Related Work

The application of Natural Language Processing to mental health has expanded rapidly, moving from depression and suicide risk detection (Coppersmith et al., 2018; Gkotsis et al., 2017) to modeling clinical dialogue dynamics like empathy and therapeutic intent (Pérez-Rosas et al., 2019). However, Na et al. (2025) highlight fragmentation in LLM research on psychotherapy and limited theoretical integration, suggesting the automated analysis of defense mechanisms remains a highly significant yet underexplored frontier. Most prior psycholinguistic analyses of therapy transcripts have relied on lexicon-based features or shallow classifiers (Miner et al., 2022; Kim et al., 2025). These approaches struggle with the deep pragmatic context and subtle cues of the DMRS framework, where signals are encoded in implicit narrative structure rather than explicit lexical content.

A primary obstacle in modeling these mechanisms is fine-grained ordinal classification (Zhang et al., 2017). Defense mechanisms lie on a continuum, and adjacent levels often differ only by subtle shifts in intent, affect, or defensive strategy. In this setting, direct prompting of LLMs often reveals *extreme response bias* — a tendency to choose extreme categories on an ordinal rating scale, leading to overly extreme or highly confident judgments. A related failure mode is *extreme label bias*, where predictions disproportionately collapse toward the most salient extreme classes rather than reflecting the full graded structure of the rubric (Reif and Schwartz, 2024; Wang et al., 2024). This behavior exposes a deeper clinical reasoning gap: general-purpose LLMs do not reliably apply the nuanced, rule-bound distinctions required to separate highly overlapping ordinal states.

To bridge this gap, recent advancements in chain-of-thought prompting (Wei et al., 2022) and knowledge distillation (Ho et al., 2023; Magister et al., 2023) offer a pathway to teach models the requisite intermediate clinical logic. Concurrently, Retrieval-Augmented Generation (RAG; Lewis et al. 2020) has proven effective at grounding models in complex texts, often improving downstream reasoning and classification performance (Shi et al., 2024). By retrieving matching patterns directly from the clinical handbook, models can bypass the need to classify utterances in a vacuum. Our work synthesizes these paradigms, demonstrating that LLMs with task-specific supervision through distilled rea-

soning traces, combined with rubric-grounded retrieval, and chain-of-thought generation, provide a robust solution to the complexities of clinical NLP.

3 Methodology

We frame defense mechanism classification as a supervised sequence classification task over the PSYDEFCONV (Na et al., 2026b) dataset. Following the shared-task setup, each seeker utterance is assigned one of nine task labels (0–8): the seven DMRS maturity levels plus two auxiliary categories. We first establish traditional encoder-only baseline, followed by a progression of seven LLM-based systems designed to systematically address the specific challenges of this clinical classification task.

3.1 Approach 1: Encoder-Based BERT Variants

To assess whether a simpler encoder-only architecture could solve the task, we fine-tuned two BERT-family models: the larger *bert-large-uncased* (Devlin et al., 2018), and *mental-bert-base-uncased* (Ji et al., 2022), a smaller but domain-adapted variant pre-trained on mental health-related Reddit posts. We represent the dialogue history as a single sequence using explicit speaker-role markers ([SEEKER] and [SUPPORTER]), concatenate the context and target utterance with the [SEP] token, and apply left-side truncation to preserve the most recent content when the input exceeds 512 tokens. We split the training data into train and validation subsets using a stratified 85/15 split to preserve the label distribution, and optimize the models with class-weighted cross-entropy to address class imbalance. Both models are trained for 5 epochs with a learning rate of 2×10^{-5} .

3.2 Approach 2: Direct Prompting with Strong LLM

In this few-shot baseline, we provide the full DMRS handbook ($\sim 12k$ tokens) and explicitly include three examples for each of the defense levels. We instruct Gemini 3.1 Pro Preview (Temperature = 0.1) to predict the defense level by generating a structured, four-step chain-of-thought rationale (See Appendix A.): (i) *Context Analysis* (situating the utterance within the dialogue history); (ii) *Goal Identification* (determining the communicative intent of the target utterance); (iii) *DMRS Alignment* (mapping clues to the handbook criteria); and (iv) *Differential Analysis* (ruling out adja-

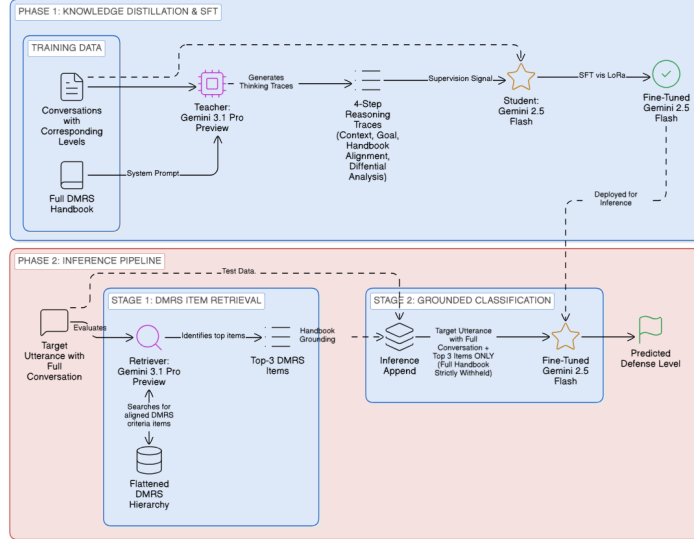


Figure 1: Methodology Diagram of the Best Performing System (Approach 7).

cent levels). Despite the model’s one-million-token context window, it produces a highly skewed prediction distribution, concentrating near the ordinal extremes, especially levels 0 (*No Defenses*) and 7 (*High-Adaptive Defenses*). We treat this extreme response bias as a key failure mode of direct long-context prompting, motivating the development of the subsequent approaches.

3.3 Approach 3: SFT of Small LLM with Knowledge Distillation

We introduce task supervision by fine-tuning *Ministral-8B-Instruct-2410* via a teacher-student distillation paradigm. A teacher model (Gemini 3.1 Pro Preview) with access to the full DMRS handbook and ground-truth label generates structured reasoning traces for each training example, following the four-step rationale from Approach 2. We then fine-tune the student model on these traces using a balanced and trimmed training set. By distilling the rubric’s knowledge directly into the generated reasoning traces, the handbook and the few-shot examples can be withheld at inference time, eliminating the massive prompt overhead while successfully preserving the required clinical logic.

3.4 Approach 4: SFT of Strong LLM

To evaluate the effect of model scale on task performance, we apply the same knowledge distillation framework from Approach 3 to Gemini 2.5 Flash, a substantially larger model. The model is fine-tuned via Google Vertex AI (LoRA rank = 16; 5 epochs; learning_rate_multiplier=1) on the full train-

ing set of the teacher-generated reasoning traces. This experiment tests whether a stronger model can better resolve the subtle distinctions within the DMRS hierarchy than the 8B-parameter student model used in Approach 3.

3.5 Approach 5: DMRS-Q Items Retrieval with Weighted Voting

To address the needle-in-a-haystack problem of locating relevant passages within a long clinical rubric, we reframe the classification task as extraction. We observe that a foundation LLM, despite poor ordinal classification performance, is a capable candidate *extractor*. We flatten the DMRS hierarchy by removing all explicit ordinal identifiers, including level numbers, hierarchy markers, and level-specific groupings, and prompt Gemini 3.1 Pro Preview to perform generative retrieval with explicit justification, identifying the top-3 DMRS-Q items most aligned with a target utterance within its conversational context. This flattening technique removes the ordinal target structure from the retrieval space, preventing the model from triggering its extreme response bias during extraction. The retrieved items are subsequently mapped back to their parent levels (0–8), and a final prediction is made via weighted voting. We evaluated two heuristic voting schemes: a strict top-1-dominant setting [0.7, 0.3, 0.1], and a more balanced alternative [0.45, 0.35, 0.20] that allows the combined weight of the second and third retrievals to outweigh the first when they agree.

System	Description	Accuracy	Precision	Recall	Macro-F1
1a	bert-large-uncased Finetuning	45.6	21.1	20.6	20.2
1b	mental-bert-base-uncased Finetuning	35.0	26.0	24.7	22.3
2	Direct Prompting (gemini-3.1-pro-preview)	38.3	37.6	14.9	20.5
3	SFT Small LLM (Ministral-8B-Instruct-2410)	42.6	25.3	28.4	25.3
4	SFT Strong LLM (gemini-2.5-flash)	52.1	33.0	23.7	26.5
5a	Retrieval + Voting [0.7, 0.3, 0.1]	64.8	35.0	37.3	34.5
5b	Retrieval + Voting [0.45, 0.35, 0.20]	65.0	39.0	39.6	37.0
6	Retrieval + Pro Re-ranking	64.7	41.1	37.2	35.7
7	System 4 + Retrieval	67.4	42.6	40.9	39.6
8a	System 7 + Pro Re-ranking (18 samples)	66.9	41.2	40.4	38.6
8b	System 7 + Pro Re-ranking (131 samples)	66.9	42.0	37.0	35.9

Table 1: Performance of all evaluated systems corresponding to the approaches described in Section 3. All values are percentages. Bold denotes the best result per column.

3.6 Approach 6: Retrieval + Stronger LLM Re-ranking

Static voting cannot effectively resolve cases where the top-3 retrieved items map to contradictory levels, as it lacks any mechanism for reasoning over conflicting evidence. We therefore propose a two-stage pipeline in which the retriever from Approach 5 produces candidate levels, and Gemini 3.1 Pro Preview re-evaluates them with explicit chain-of-thought reasoning. We passed *only* the corresponding handbook sections rather than the full rubric as system prompt. This narrows the model’s attention to at most 3 candidate levels, directly mitigating the needle-in-a-haystack problem.

3.7 Approach 7: Supervision + Retrieval (Best System)

Our best performing method synthesizes the strengths of task-specific supervision (Approach 4) and dynamic handbook grounding (Approach 5).

We prompt the fine-tuned Gemini 2.5 Flash classifier (from Approach 4) with the full conversational context, the target utterance, and the dynamically retrieved top-3 DMRS-Q items (from Approach 5), as illustrated in Figure 1. The full DMRS handbook is strictly withheld from the system instruction to eliminate inference overhead and avoid context dilution.

We designed this hybrid architecture to test whether the model could leverage its internalized SFT clinical reasoning while grounding its prediction in the narrow, dynamic context of the retrieved items.

3.8 Approach 8: Selective Re-ranking via Stronger LLM

Building on our best-performing system (Approach 7), we test whether a stronger foundation model can serve as an effective tie-breaker in cases

of severe retrieval contradiction. While the fine-tuned Flash model excels generally, we hypothesized that the superior reasoning capacity of Gemini 3.1 Pro Preview could better resolve edge cases, despite lacking task-specific fine-tuning. Predictions from Approach 7 are retained when all top-3 retrievals agree ($\sim 75\%$ of cases). When retrievals conflict, the stronger Pro model is invoked to resolve the prediction using explicit chain-of-thought reasoning over the handbook sections of the candidate levels. We evaluate this cascade under two disagreement thresholds: cases where all three retrievals differ (18 samples), and cases where at least two differ (131 samples).

4 Results

Table 1 reports Accuracy, Precision, Recall, and macro-F1 for all systems on the PSYDEFCONV test set (472 examples), with metrics computed over the positive classes (1–8) only. The detailed class-wise behavior of the best system (System 7) is shown in Appendix Figure 2.

Encoder baselines are weak lexical benchmarks.

Although encoder-based models establish a foundation, they lack the capacity to explicitly reason over subtle cues and the complex DMRS rubric. Notably, the larger bert-large-uncased model underperforms its smaller but domain-adapted counterpart, mental-bert-base-uncased, suggesting that domain adaptation is more beneficial than model scale for this task.

Supervision is essential; scale is not a substitute.

The state-of-the-art LLM (System 2) completely fails despite accessing the full DMRS handbook within its large context window, achieving performance comparable to BERT-based baselines (System 1). The foundation model exhibits severe

extreme response bias, collapsing into confident, plausibly reasoned, yet clinically ungrounded predictions at the extreme ends of the scale (levels 0 and 7). This confirms that fine-grained ordinal classification fundamentally requires task-specific SFT.

Furthermore, comparing our supervised models reveals the nuanced impact of model scale. Although the substantially larger Gemini 2.5 Flash (System 4) improves accuracy substantially over the 8B-parameter student (System 3), its macro-F1 gain is surprisingly modest. This suggests that scale alone does not resolve the minority-class and ordinal ambiguity challenges. Together, these results demonstrate that the clinical supervision signal is the true bottleneck; prompt engineering and raw parameter scale alone cannot overcome the clinical reasoning gap.

Retrieval is the single most impactful intervention. Reframing the task as generative extraction (System 5) is the most effective isolated intervention, yielding a substantial performance leap without any additional training. As hypothesized, removing the ordinal identifiers from the prompt and output spaces via rubric flattening successfully neutralizes extreme response bias while mitigating the needle-in-a-haystack problem for downstream inference.

Interestingly, comparing the voting weight configurations in Systems 5a and 5b reveals that down-weighting the top-1 item improves overall performance. This suggests that while the foundation LLM is an excellent candidate extractor, its raw initial ranking is not a sufficiently reliable proxy for direct classification.

Supervision and retrieval are strongly complementary. Our best system (System 7) combines dynamically retrieved DMRS-Q items with the fine-tuned Gemini 2.5 Flash classifier, producing a substantial +13.1 macro-F1 improvement over using SFT alone (System 4). Furthermore, we find that the fine-tuned model demonstrates strong resilience to retrieval noise¹, successfully correcting predictions even in cases where all three retrieved candidates map to incorrect levels. This suggests that retrieval and SFT capture robust, mutually reinforcing signals: retrieval surfaces localized DMRS evidence to ground the prediction and narrow the decision space, while SFT provides the clinical

reasoning needed to interpret, weigh, and sometimes override that evidence under noisy retrieval conditions.

General-purpose cascaded re-ranking amplifies retrieval noise. Systems 6 and System 8 expose the strict limitations of relying on un-tuned foundation models to evaluate retrieved candidates. Replacing static voting with LLM re-ranking (System 6) actively degrades performance relative to heuristic voting (System 5). Because both systems evaluate the exact same retrieved items, this demonstrates that general-purpose reasoning lacks the clinical logic necessary to accurately weigh candidate labels—proving the bottleneck is the reasoning process itself, not just upstream retrieval accuracy. Furthermore, invoking this stronger foundation model as a selective tie-breaker for contradictory retrievals (System 8) underperforms our SFT-driven hybrid approach (System 7). In both cascade scenarios, the explicit clinical logic internalized by the fine-tuned model proves far more reliable at resolving ambiguities and recovering from retrieval noise. Ultimately, these results reiterate that task-specific supervision is strictly superior to raw parameter scale, even at the re-ranking stage.

5 Conclusion

Our results show that raw parameter scale alone cannot overcome the clinical reasoning gap inherent in foundation models for fine-grained ordinal defense mechanism classification. Instead, task-specific supervised fine-tuning and retrieval-based grounding provide complementary signals that substantially improve performance. A key caveat is that our best system (System 7) has train-inference asymmetry: retrieval is injected only at inference time, while the fine-tuned classifier is trained without retrieved DMRS-Q items. This means the student learns from reasoning traces alone and then adapts to appended handbook evidence at test time, which may limit its ability to fully exploit the retrieved context. A natural next step is retrieval-augmented fine-tuning, where retrieved items are included during training so the model can learn to integrate them end-to-end. More broadly, future work in clinical NLP should focus on improving upstream retrieval precision, refining knowledge distillation for supervision, and ultimately deepening the synergy between targeted clinical supervision and rubric-grounded retrieval.

¹We provide a qualitative case study in Appendix B.

Limitations

Our study has several limitations. First, our best system exhibits train-inference asymmetry, since retrieved DMRS-Q items are provided only at inference time and not during fine-tuning. We discuss this as a major direction for retrieval-augmented training. Second, our distillation as well as retrieval pipeline depends on proprietary models, which introduces reproducibility concerns and non-trivial API costs, limiting accessibility for resource-constrained research groups. Third, due to compute constraints, the Ministral-8B model in Approach 3 was trained for only 2 epochs on a trimmed, balanced subset, and Gemini 2.5 Flash SFT was not extended beyond 5 epochs despite still improving, suggesting our results are a lower bound on what supervised fine-tuning can achieve. Finally, our retrieval step relies on a single generative model without re-ranking diversity, making the pipeline vulnerable to systematic retrieval errors that downstream models cannot always recover from.

References

- Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. [Natural language processing of social media as screening for suicide risk](#). *Biomedical Informatics Insights*, 10:1178222618792860.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.
- Mariagrazia Di Giuseppe and J. Christopher Perry. 2021. [The hierarchy of defense mechanisms: Assessing defensive functioning with the defense mechanisms rating scales q-sort](#). *Frontiers in Psychology*, Volume 12 - 2021.
- George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J. P. Hubbard, Richard J. B. Dobson, and Rina Dutta. 2017. [Characterisation of mental health conditions in social media using informed deep learning](#). *Scientific Reports*, 7(1).
- Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. [Large language models are reasoning teachers](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882. Association for Computational Linguistics.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. [Mental-BERT: Publicly available pretrained language models for mental healthcare](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.
- Samuel Kim, Oghenemaro Imieye, and Yunting Yin. 2025. [Interpretable depression detection from social media text using llm-derived embeddings](#). *Preprint*, arXiv:2506.06616.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. [Teaching small language models to reason](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.
- Adam S. Miner, Scott L. Fleming, Albert Haque, Jason A. Fries, Tim Althoff, Denise E. Wilfley, W. Stewart Agras, Arnold Milstein, Jeff Hancock, Steven M. Asch, Shannon Wiltsey Stirman, Bruce A. Arnow, and Nigam H. Shah. 2022. [A computational approach to measure the linguistic characteristics of psychotherapy timing, responsiveness, and consistency](#). *npj Mental Health Research*, 1(1).
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the psydefdetect shared task at bionlp 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.
- Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.
- Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. [What makes a good counselor? learning to distinguish between high-quality](#)

- and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.
- Yuval Reif and Roy Schwartz. 2024. [Beyond performance: Quantifying and mitigating label bias in LLMs](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6784–6798, Mexico City, Mexico. Association for Computational Linguistics.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. [Replug: Retrieval-augmented black-box language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 8371–8384. Association for Computational Linguistics.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. [Ordinal common-sense inference](#). *Transactions of the Association for Computational Linguistics*, 5:379–395.

Appendix

A Prompt Listings

```
class ClinicalReasoning(BaseModel):
    context_trigger: str = Field(
        description="Description of the
        stressor in the dialogue."
    )
    psychological_goal: str = Field(
        description="What the speaker is
        trying to achieve/avoid."
    )
    handbook_alignment: str = Field(
        description="Specific evidence
        from the handbook that justifies the
        label."
    )
    differential_diagnosis: str = Field(
        description="Why this isn't a
        higher or lower-level defense."
    )

class DefensePrediction(BaseModel):
    clinical_reasoning:
    ClinicalReasoning
    defense_level: int = Field(
        description="The numeric defense
        level (0-8).")
    label: str = Field(description="The
    specific name of the defense
    mechanism used.")

SYSTEM_INSTRUCTION = f"""
You are an expert clinical psychologist
and data annotator. Your task is to
analyze dialogues and classify the
psychological defense mechanism used
in the 'current_text_to_classify'
based on the Defense Mechanisms
Rating Scales (DMRS) hierarchy.

You must assign exactly one label from
the list below:
{LABEL_DESCRIPTIONS}

Your classification must be grounded in
the DMRS hierarchy given below. This
following comprehensive HANDBOOK
serves as your core classifying
guideline:
{HANDBOOK_TEXT}

Here are some examples of how to reason
through the task:
{FEW_SHOT_EXAMPLES}

CORE INSTRUCTIONS:
1. Primacy of Context: Always read the
preceding dialogue to understand
what triggered the '
current_text_to_classify'.
2. Function-Oriented: Ask yourself, "
What psychological goal is the
speaker trying to achieve?"
3. Handbook Grounded: Match the behavior
to the specific criteria in the
DMRS Handbook. Reason through why
specific criteria are met.
```

```
4. You must maintain hierarchical
integrity explicitly reason
through why the classification does
not drift into higher (more adaptive
) or lower (more pathological)
levels by verifying that all
exclusionary criteria for the
selected level are met.
5. Distinguish Emotion from Defense:
Saying "I am sad" is Level 0. A
defense requires distortion,
avoidance, or transformation.
6. Always pick the single most accurate
label (0 8 ) from the LABEL
REFERENCE above.
7. Output strict JSON matching the
requested schema.
"""
```

Listing 1: Base System Prompt primarily utilized in Approach 2. Subsequent approaches adapt this template by entirely withholding the handbook and few-shot examples during inference for supervised models, substituting the handbook with dynamically retrieved top-3 DMRS-Q items, or isolating specific candidate sections of the handbook for re-ranking.

```
class ClinicalReasoning(BaseModel):
    context_trigger: str = Field(
        description="Description of the
        stressor in the dialogue."
    )
    psychological_goal: str = Field(
        description="What the speaker is
        trying to achieve/avoid."
    )
    handbook_alignment: str = Field(
        description="Specific evidence
        from the handbook that justifies the
        defense level (0-8).")
    differential_diagnosis: str = Field(
        description="Why this isn't a
        higher or lower-level defense."
    )

SYSTEM_INSTRUCTION = f"""
You are an expert clinical psychologist
and data annotator. Your task is to
analyze dialogues and generate the
exact clinical reasoning (thought
trace) that perfectly justifies the
PROVIDED psychological defense
mechanism level for the '
current_text_to_classify', based on
the Defense Mechanisms Rating Scales
(DMRS) hierarchy.

The correct true label level will be
given to you in the prompt. You must
explain *why* it is the correct
level.
{LABEL_DESCRIPTIONS}

Your classification reasoning must be
grounded in the DMRS hierarchy given
below. This following comprehensive
```

```

HANDBOOK serves as your core
classifying guideline:
{HANDBOOK_TEXT}

CORE INSTRUCTIONS:
1. Primacy of Context: Always read the
preceding dialogue to understand
what triggered the '
current_text_to_classify'.
2. Function-Oriented: Ask yourself, "
What psychological goal is the
speaker trying to achieve or avoid?"
3. Handbook Grounded: Match the behavior
to the specific criteria in the
DMRS Handbook. Reason through why
specific criteria are met pointing
towards the provided correct label.
4. Hierarchical Integrity: You must
maintain hierarchical integrity
explicitly reason through why the
classification does not drift into
higher or lower levels by verifying
that all exclusionary criteria for
the selected level are met.
5. Emulate the Correct Path (CRITICAL):
Provide the thought process as if
you independently arrived at the
provided correct label.
6. Output strict JSON matching the
requested schema, ensuring your
output 'defense_level' matches the
provided correct level.
"""

```

Listing 2: System Prompt for the Reasoning Trace Generation by the Teacher Model.

```

SYSTEM_INSTRUCTION = f"""
You are an expert clinical psychologist
and data annotator.

Below is the Psychological Defense
Mechanism Coding Handbook:
{FLATTEN_HANDBOOK_TEXT}

YOUR TASK:
You will be provided with a dialogue and
a 'current_text_to_classify'.
You must SCAN THROUGH THE HANDBOOK and
pick out the MOST MATCHING 3 DMRS-Q
items for the given text.

CRITICAL INSTRUCTIONS:
1. FOCUS EXPLICITLY ON THE DMRS-Q ITEMS.
Compare the dialogue's scenario
against the specific behaviors
described in the DMRS-Q items.
2. Identify the corresponding defense
mechanism (such as 'Splitting', '
Acting Out', 'Humor', 'Repression').
3. Base your selections strictly on the
Definitions, Given Scenarios,
Examples, and Distinctions provided
in the handbook.
4. Provide exactly 3 matches. If one
defense is blatantly obvious,
provide the top 3 DMRS-Q items that
best capture the nuances of the
interaction.
"""

```

Listing 3: System Prompt for DMRS-Q Items Retrieval.

```

class DMRSItemMatch(BaseModel):
    dmrs_q_item: str = Field(
        description="The exact DMRS-Q
        Item number and text from the
        handbook (e.g., 'ITEM 45: At times
        when expressing...')."
    )
    match_justification: str = Field(
        description="Detailed
        explanation of why this specific
        DMRS-Q item matches the conversation
        based on the handbook's definition,
        scenario, examples, and
        distinctions."
    )
    sublevel_name: str = Field(
        description="The specific
        defense mechanism name from the
        handbook (e.g., 'Passive Aggression',
        'Rationalization', 'Devaluation'
        etc.). Do NOT use the 0-8 level
        numbers."
    )

class ExtractionResult(BaseModel):
    top_3_matches: list[DMRSItemMatch] =
    Field(
        description="Exactly 3 most
        matching DMRS-Q items and their
        details."
    )

```

B Qualitative Case Study: SFT as a Retrieval Error-Correction Mechanism

To examine why supervised fine-tuning combined with retrieval (System 7) improves performance beyond direct prompting with a large foundation model (System 2), retrieval-based weighted voting (System 5), and retrieval-based re-ranking with an unsupervised LLM (System 6), we analyze a representative test case in which all three non-supervised prediction approaches described above fail, while the supervised model correctly predicts the gold label.

This analysis is especially important because the training dataset is highly imbalanced. Table 2 shows the label distribution in the dataset. Within the seven DMRS maturity-level classes, Level 5 is the smallest class, with only 48 examples. Therefore, correct prediction of a Level 5 case by the supervised model is meaningful: **the model is not simply following the majority-class tendency, but is able to identify a minority mid-level defense that all the other systems miss.**

Label	Count
0	296
1	108
2	61
3	99
4	84
5	48
6	172
7	968
8	28

Table 2: Label distribution of the training set. The distribution is highly imbalanced, with Level 7 being the dominant class and Level 5 being the smallest class among the seven DMRS maturity-level labels.

System	Predicted level	Correct?
Gold	5	–
System 2	0	No
System 5	7	No
System 6	7	No
System 7	5	Yes

Table 3: Prediction outcomes for test_00377, a Level 5 minority-class case where Systems 2, 5, and 6 fail, while System 7 recovers the correct label.

In test sample test_00377, where the true gold label is Level 5, the un-tuned foundation model (System 2, Gemini 3.1 Pro Preview) misinterprets the clinical nuance as a neutral conversational response and predicts Level 0, a textbook illustration of the severe extreme response bias that plagues direct prompting strategies on ordinal scales. For the retrieval-only framework (System 5), the top-3 retrieved DMRS-Q items map to levels 7, 7, and 5 respectively, causing the static heuristic voting scheme to collapse toward the incorrect majority, outputting Level 7. Even when advanced general-purpose reasoning is applied via LLM re-ranking (System 6), the un-tuned foundation model fails to resolve the conflicting evidence and yields Level 7 with plausible but ungrounded reasoning. In sharp contrast, our best hybrid architecture (System 7) successfully demonstrates its resilience to upstream retrieval noise. By harmonizing its internalized supervised reasoning with dynamic handbook grounding, it successfully corrects the misleading retrieval weights to accurately predict the gold Level 5 label.

Specifically, the supervised model recognizes that the target utterance immediately succeeds an unresolved, emotionally charged dialogue regarding abandonment, interpersonal loss, and financial insecurity. Rather than misinterpreting this context as a neutral conversational shift, the model utilizes its internalized clinical logic to identify a Level 5 defensive strategy (*Displacement*), wherein the

speaker defuses acute emotional distress by pivoting to a safer, tangential topic. Crucially, the model’s differential diagnosis proves that it understands why the top-retrieved Level 7 items are incorrect: it explicitly recognizes the speaker is merely sharing a casual anecdote, not actively applying collaborative advice (*Affiliation*) or utilizing an established coping mechanism (*Sublimation*).

Importantly, while the correct DMRS-Q item is surfaced during the initial retrieval phase, its position as the third-ranked candidate causes it to be overridden by conflicting evidence within both the heuristic voting (System 5) and un-tuned re-ranking (System 6) pipelines. This diagnostic outcome strongly substantiates our central finding: **although retrieval-augmented frameworks excel at capturing relevant rubric clauses, task-specific supervised reasoning remains indispensable for reliably interpreting and weighing that evidence when candidate items are noisy or highly competitive.**

The complete dialogue transcript for the test instance test_00377 along with the corresponding reasoning trace outputs for Systems 2, 5, 6, and 7 are presented below. To maximize readability, superficial formatting artifacts, such as escaped new-line characters and markdown emphasis markers, have been removed, while the verbatim wording of the model generations has been strictly preserved.

B.1 Full Conversation and Target Utterance

```

=== test_00377 ===
Supporter: Hello how are you?
Seeker: hello im looking for someone to talk to
im fine how are you
Supporter: I'm doing ok I'm glad you are good. Is
it snowing by you? Merry Christmas!
Seeker: thats great and no its not snowing its
very cold thow merry christmas to you also
Supporter: How can I help you today?
Seeker: im having some issues with friends not
actually being friends
Supporter: I hear you are having trouble figuring
out which friends are really your friends and
which ones aren't. Is that about right?
Seeker: yes
Supporter: I understand that... Did something
happen that makes you feel like that?
Seeker: yes their is no communication or
interaction between me and them nobody answers
phone calls txt messages etc
Supporter: Is this a significant other?
Seeker: no friends
Supporter: Are they ignoring you?
Seeker: when we where younger we all hung out all
the time now that we are older its like they talk
with each other im not sure if i did anything or
not

```

Supporter: Are you sure they are okay? Or not away with family for the holidays?
 Seeker: it doesnt matter they still keep open communication with each other
 Supporter: That's no good what do you think the problem is?
 Seeker: I believe its because I am the friend that has financial trouble and normally needs to borrow money
 Supporter: Do you owe someone already?
 Seeker: no anything i owe i always pay back perhaps they are tired of being asked
 Supporter: That can get in the way of a friendship sometimes, but I try not to get ahead of myself and jump the gun. They may have lost a phone or something may have come up... That very well could be... Are you working?
 Seeker: no im not currently I also think sometimes to myself that they don't answer because they think I'm asking for money
 Supporter: I see, Have you tried for unemployment?
 Seeker: yes i dont qualify
 Supporter: I understand that. What if that is the case? You may need to talk to them and let them know how you feel about that. How long has this been going on?
 Seeker: I'm nervous about asking them why and it happens on and off like when they feel lie it i guess i will say maybe 4 months
 Supporter: I understand that, but if you want to get to a point that you feel better about it then you will have to speak to them.
 Seeker: yes i know i hope everything be immediately resolved and not lead into other problems and end of friendships
 Supporter: Do you see them everyday? I understand that, are you able to go over there.
 Seeker: no i dont live in the same city anymore when i go out of town i can visit but with covid19 rules
 Supporter: Well I can see that you have put the effort in to try and keep the relationship so it sounds like the ball is in their court if they want to do the same.
 Seeker: yes
 Supporter: Well that is about all that can be done and wait for them to come around...Do you have any hobbies?
 Seeker: yes i like to read and play music
 Supporter: I like to go on runs every morning. It helps to clear my mind and makes it easier to brush things off when they bother me. What instrument do you play?
 Seeker: my sister goes on runs in the morning time and says it relax her for the day i think i migh i played the piano when i was in highschool
 Target Utterance: my sister goes on runs in the morning time and says it relax her for the day i think i migh i played the piano when i was in highschool

B.2 System 2 Direct Prompting Reasoning Trace

```
{
  "clinical_reasoning": {
```

```
    "context_trigger": "The Supporter shifted the conversation from the Seeker's primary stressor (distant friends) to potential coping mechanisms, specifically mentioning running and asking what instrument the Seeker plays.",
    "psychological_goal": "The Seeker aims to maintain the conversational flow by answering the Supporter's direct question about the instrument and acknowledging their comment about running.",
    "handbook_alignment": "This utterance is a factual and functional response to the Supporter's prompt. According to the handbook, utterances that serve a purely conversational or social function and do not engage with emotional conflict or psychological content are classified as Level 0. Furthermore, reporting factual details of an event or engaging in social small talk on neutral topics are explicitly listed as situations that are typically not defenses.",
    "differential_diagnosis": "This is not a highly adaptive defense like Sublimation because the Seeker is merely recalling a past hobby (playing piano in high school) and considering running, rather than actively channeling their current emotional distress into a creative or athletic pursuit. It does not qualify as any lower-level defense because there is no evidence that a difficult emotion or internal conflict is being avoided, distorted, or transformed."
  },
  "defense_level": 0,
  "label": "No Defense / Neutral Utterance"
}
```

B.3 System 5 Retrieval Trace

```
{
  "dmrs_q_item": "ITEM 44: When the subject describes seeking help from others, there is a sense of having learned something from the interchange.",
  "match_justification": "The Seeker is actively engaging with the Supporter's shared experience about running to clear the mind. By stating 'i think i migh' (meaning they might start running), the Seeker demonstrates that they are taking away a potential new coping strategy and learning from the supportive interchange.",
  "sublevel_name": "Affiliation",
  "defense_level": 7
},
{
  "dmrs_q_item": "ITEM 97: Following experiences of emotional distress or conflict, the subject engages in sports or other physical activities which are an invigorating outlet for any lingering frustrations.",
  "match_justification": "In the context of feeling distressed and anxious about their friendships, the Seeker considers taking up running after hearing that it helps relax the mind and brush things off. This shows an inclination toward using physical activity as a healthy, sublimatory outlet for their current emotional frustrations.",
```

```

    "sublevel_name": "Sublimation",
    "defense_level": 7
  },
  {
    "dmrs_q_item": "ITEM 69: When confronting emotionally charged topics, the subject tends not to address concerns directly and fully but wanders off to tangentially related topics that are emotionally easier for the subject to discuss or prefers to pay attention to someone else dealing with a similar situation. This can include preferring to read or watch a film portraying people dealing with similar problems.",
    "match_justification": "The conversation has shifted from the highly distressing topic of the Seeker's failing friendships and financial insecurities to the safe, neutral topic of hobbies (running, playing the piano). The Seeker readily participates in this shift, focusing on these easier, tangential subjects rather than dwelling on the immediate emotional conflict.",
    "sublevel_name": "Displacement",
    "defense_level": 5
  }
}

```

- Sublimation (Level 7): Sublimation involves actively channeling distressing feelings into a socially acceptable, creative, or physical outlet. While running and playing the piano are classic examples of sublimatory activities, the Seeker is only considering running in the future ("i think i migh") and played the piano in the past ("when i was in highschool"). Because they are not currently engaging in the activity to process their immediate distress, Sublimation is not the best fit.

- Affiliation (Level 7): Affiliation involves turning to others for support and collaboratively finding solutions or enhancing coping skills. By listening to the Supporter's personal experience with running, validating it, and stating "i think i migh" [try it], the Seeker demonstrates that they are actively learning from the supportive interchange. They are taking away a tangible, adaptive coping strategy from the conversation, which perfectly aligns with the criteria for Affiliation (specifically ITEM 44: "there is a sense of having learned something from the interchange").

4. Verdict:

The Seeker is utilizing the supportive relationship to adopt a new, healthy coping strategy, making Affiliation the most accurate defense mechanism.

Final Selected Defense Level

7

B.4 System 6 Foundation LLM Re-ranking Trace

1. Context:

The Seeker has been discussing their distress over failing friendships and financial insecurities. The Supporter, recognizing that there is little more to be done about the friends at this moment, shifts the topic to hobbies as a way to cope. The Supporter mentions that they go on runs to clear their mind and asks what instrument the Seeker plays.

2. Function:

The Seeker responds to both parts of the Supporter's message. They validate the Supporter's coping strategy by mentioning their sister's similar experience with running, express an openness to trying it themselves ("i think i migh"), and answer the direct question about the instrument they play. The psychological function here is engaging with the support offered and considering a new coping mechanism.

3. Evaluation:

- Displacement (Level 5): Displacement involves the subject unconsciously shifting the focus from a distressing topic to a safer, tangential one to avoid emotional discomfort. However, in this dialogue, the Supporter initiated the topic shift to hobbies. The Seeker is simply following the conversational lead and answering direct questions. Therefore, the Seeker is not actively using displacement.

B.5 System 7 Supervised Reasoning Trace

Context Trigger: The Supporter has just concluded the discussion about the Seeker's distressing friendship issues, suggesting that the Seeker has done all they can and the ball is in the friends' court. The Supporter then asks about the Seeker's hobbies.

Psychological Goal: To avoid the lingering emotional pain, anxiety, and helplessness associated with the unresolved friendship conflict by readily shifting the conversational focus to a neutral, safe, and positive topic.

Handbook Alignment: The utterance aligns with Displacement (Level 5). The DMRS handbook defines Displacement as keeping unacceptable motives out of awareness, where feelings surface indirectly. Specifically, ITEM 69 states: 'When confronting emotionally charged topics, the subject tends not to address concerns directly and fully but wanders off to tangentially related topics that are emotionally easier for the subject to discuss...' The Seeker was just discussing a highly emotionally charged topic (feeling ignored and abandoned by friends due to financial issues). The Supporter then shifts the topic to hobbies. The Seeker readily engages in this new, emotionally easier topic, talking about their sister's running and their past piano playing, thereby displacing the focus from the distressing conflict to a safe, tangential subject.

Differential Diagnosis: This is not a Level 0 (Neutral Utterance) because, in the immediate context of a highly distressing and unresolved emotional conflict, the abrupt and enthusiastic shift to a completely unrelated topic serves a defensive function to escape the pain. It is not Level 7 (Affiliation or Sublimation) because the Seeker is not actively using the Supporter's advice to collaboratively solve the friendship problem (Affiliation), nor are they describing an active, established coping mechanism to channel their frustration (Sublimation); they are merely agreeing with the Supporter's hobby and sharing a casual, unrelated anecdote. It is not Level 6 (Isolation of Affect) because the Seeker is not discussing the distressing event in a detached manner; they have completely changed the subject to avoid the event altogether.

Defense Level: 5

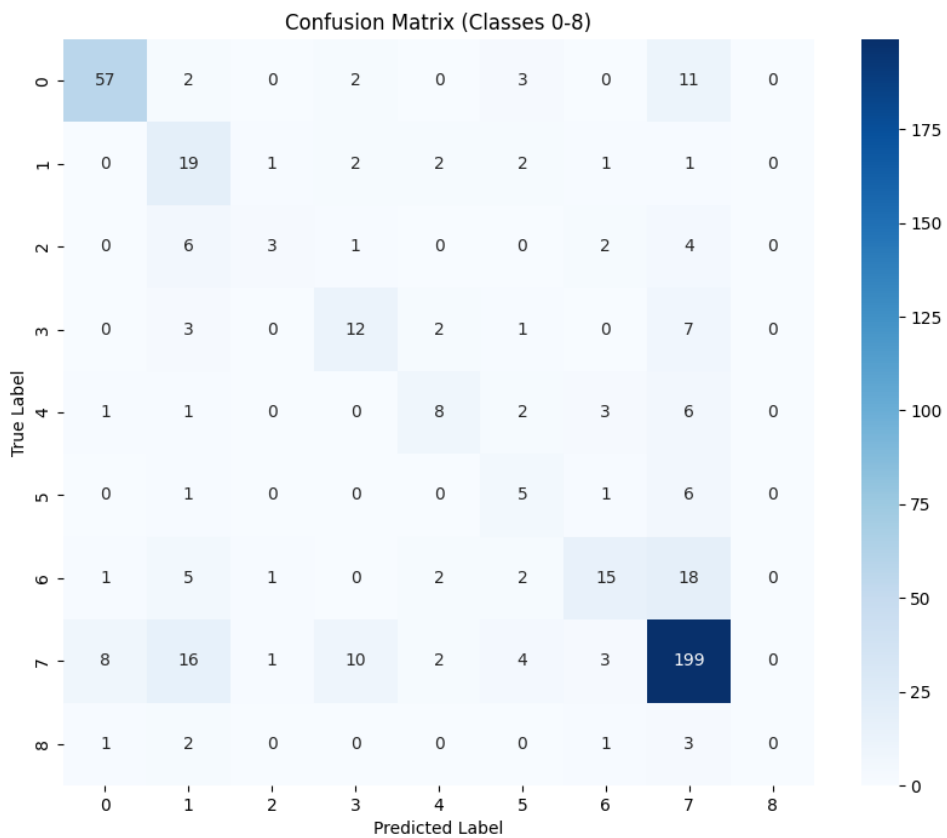


Figure 2: Confusion Matrix of the best performing system (Approach 7). The system demonstrates high accuracy on the most frequent classes, though it exhibits a slight prediction bias towards Level 7 when evaluating ambiguous mid-level defenses.