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

Duc-Luong Tran[1], Phuong-Anh Chu[1], Hoang-Dat Do[1], Tu-Phuong Mai[1], Duy-Cat Can[1,2,3], Hoang-Quynh Le[1]*

[1]Faculty of Information Technology, VNU University of Engineering and Technology, Vietnam
[2]Platform of Bioinformatics, Lausanne University Hospital, Switzerland
[3]Faculty of Biology and Medicine, University of Lausanne, Switzerland
{22021148, 23020324, 24020060, 21020552, lhquynh}@vnu.edu.vn
duy-cat.can@chuv.ch

Abstract

The PsyDefDetect shared task focuses on classifying nine psychological defense mechanisms in multi-turn dialogues, a problem complicated by severe label imbalance and the implicit, context-dependent nature of defenses. In this work, we investigate several approaches for dialogue-level defense detection, including supervised baselines and large language model (LLM)-based pipelines. Our primary system is a retrieval-augmented LLM framework with hierarchical prediction and lightweight heuristics for decision calibration. Experiments on the PSYDEFCONV dataset show that LLM-based methods improve overall performance compared to supervised baselines, but still struggle with fine-grained distinctions, especially for minority labels. These findings highlight the challenges of modeling implicit psychological constructs in dialogue.

1 Introduction

Psychological defense mechanisms are automatic, unconscious processes that regulate responses to emotional conflict (Perry, 2014). They influence both clinical outcomes (Mrozowicz-Wrońska, 2023) and user behaviors in emotional support conversations (ESC) (Na et al., 2026b; Di Giuseppe et al., 2024). While NLP has successfully modeled explicit constructs like empathy and sentiment (Shetty et al., 2024; Liu et al., 2021), psychological defenses remain challenging because they are highly implicit, pragmatic, and context-dependent (Na et al., 2025). Unlike emotions with distinct lexical triggers, defensive behaviors evolve dynamically across multi-turn dialogues, rendering conventional single-utterance classification insufficient.

To bridge the gap between clinical theory and NLP, the PsyDefDetect shared task (Na et al., 2026a) introduces a framework for classifying defensive functioning in multi-turn dialogues using the DMRS taxonomy (Di Giuseppe and Perry, 2021). Beyond the latent nature of the task, automated detection on the PSYDEFCONV corpus faces two major structural challenges: (i) severe label imbalance driven by the natural prevalence of mature defenses, and (ii) long, noisy conversational histories that make contextual modeling difficult, as dense and extensive contexts often trigger information degradation in language models (Liu et al., 2024).

These challenges expose limitations of conventional supervised models, whose reliance on limited and imbalanced training data prevents them from capturing the subtle, context-dependent pragmatics of implicit defenses. More importantly, they highlight a deeper mismatch between the nature of psychological defenses and the assumptions of flat classification models. To address this, we leverage the external knowledge encoded in Large Language Models (LLMs) through a retrieval-augmented framework for dialogue-level inference.

Our primary approach is a retrieval-augmented LLM pipeline that operates at the dialogue level. It decomposes predictions into a coarse-to-fine hierarchy and integrates lightweight heuristics to improve robustness. In addition, we incorporate hybrid supervised–LLM filtering to handle frequent labels and a summary-based distillation module to compress noisy conversational contexts.

2 Related Work

Prior NLP work has mainly focused on explicit affective signals such as sentiment and empathy, which are often modeled with sequence classification (Rashkin et al., 2019; Sharma et al., 2020; Liu et al., 2021). More recent studies have moved toward more latent cognitive constructs, such as cognitive distortions (Maddela et al., 2023; Chen

* Corresponding author.

Proceedings of the BioNLP 2026 (Shared Tasks), pages 164–170

et al., 2023). Psychological defense mechanisms are even more implicit, pragmatic, and context-dependent (Na et al., 2025; Perry, 2014), especially in multi-turn dialogue where meaning unfolds across turns. This makes single-utterance classification insufficient and increases the need for dialogue-level modeling. In such settings, supervised models trained on limited and imbalanced data often struggle to capture the functional role of an utterance in context.

Large language models (LLMs) have shown strong potential in mental health applications (Yang et al., 2023; Xu et al., 2024) and related clinical language tasks (Galatzer-Levy et al., 2023). Their broad linguistic and world knowledge makes them attractive for low-data, context-dependent problems. However, prior work in psychotherapy-related NLP often formulates prediction as static label assignment over isolated instances (Tu et al., 2024; Bao et al., 2024), which is not well aligned with the dynamic and hierarchical nature of psychological processes (Na et al., 2025; Di Giuseppe and Perry, 2021). These limitations motivate more structured, dialogue-aware LLM pipelines for defense detection.

Retrieval-Augmented Generation (RAG) is widely used to ground LLM predictions by providing relevant examples or contextual evidence (Gao et al., 2024). In dialogue understanding, retrieval can improve prediction stability, but semantic similarity alone may be insufficient for latent psychological constructs, where similar utterances can serve different pragmatic functions depending on context (Bender and Koller, 2020). Likewise, heuristic cues may offer high-precision signals for explicit patterns, yet often generalize poorly to subtle or evolving behaviors (Chancellor and De Choudhury, 2020), motivating retrieval designs that incorporate richer discourse-level context and lightweight calibration.

3 Methodology

We tackle PsyDefDetect under two key challenges: severe label imbalance and the implicit, context-dependent nature of psychological defenses. In addition, the task involves multi-turn dialogues, where defensive behaviors are often expressed through subtle pragmatic cues and depend heavily on preceding context, making single-utterance classification insufficient.

To address these challenges, we investigate four methodological approaches. *Approach 1* establishes supervised dialogue-level baselines. *Approach 2*, our core contribution, develops a retrieval-augmented LLM pipeline to utilize the internal clinical knowledge of LLMs for identifying latent psychological patterns. Building upon this, *Approach 3* extends the framework into a hybrid supervised–LLM pipeline. Finally, *Approach 4* introduces an auxiliary summary-based distillation strategy.

3.1 Approach 1: Supervised Encoders (Baseline)

We establish a supervised baseline using a RoBERTa–LSTM dialogue encoder, where each utterance from both seeker and helper is first encoded via RoBERTa, and an LSTM aggregates these sequential representations into a dialogue-level embedding for classification. This hierarchical encoding is particularly suited to PsyDefDetect, as defense mechanisms can dynamically shift across turns and thus require interpreting each utterance in relation to its preceding context rather than in isolation.

On top of this encoder, we explore three prediction strategies:

- *Single-level classifier:* Directly predicts the final label from the full label set without any hierarchical decomposition.
- *2-level hierarchical classifier:* First predicts whether the instance belongs to label 7 or not, then classifies the remaining labels.
- *3-level hierarchical classifier:* Sequentially predicts label 0, then label 7, and finally classifies the remaining labels.

Motivated by extreme label imbalance, specifically the dominance of easily detectable labels 7 (51.93%) and 0 (15.88%), these hierarchical variants employ a sequence of strictly independent models. During inference, these models are applied sequentially: if an early stage predicts a frequent label, the process halts and outputs it; otherwise, the instance is passed to the next classifier. This early-exit strategy effectively isolates majority classes early on, allowing subsequent stages to focus entirely on more challenging and underrepresented minority classes.

3.2 Approach 2: Retrieval-Augmented LLM Inference

This is our main approach. The task is inherently difficult: labels are often implicit and expressed through complex discourse functions, while the dataset is small and highly imbalanced. Consequently, standard supervised models struggle to generalize, especially on minority classes. To address these challenges, we propose two retrieval-augmented inference strategies using the Gemini 3.1 Flash Lite model: a Direct LLM-RAG that predicts the final label in a single step, and a Hierarchical LLM-RAG that decomposes the task into coarse-grained grouping followed by fine-grained classification. Between these, we identify the hierarchical approach as our primary and most effective strategy, the complete architecture of which is illustrated in Figure 1.

Prompt construction and retrieval. We enhance the target dialogue with task-specific knowledge and contextual evidence through the following three-step pipeline:

- *Dialogue reconstruction:* We group samples by `dialogue_id` to recover their original conversation context. In PsyDefDetect, both training and test instances are truncated histories of larger dialogues, where labels may vary across turns. Reconstructing these partial dialogues enables the model to capture evolving context rather than treating samples as independent.

- *Generating explanations:* We use the LLM (Gemini 3.1 Flash Lite) to generate short explanations for labeled turns, turning each fragment into a (label + rationale) retrieval unit.

- *Retrieval:* For each target turn, we retrieve $k = 3$ dynamic dialogue fragments. Rather than using a neural embedding model, we utilize a domain-grounded symbolic matching approach. Local contextual similarity and lexical overlap are measured using Jaccard Similarity over preprocessed content tokens. The overall similarity score is a weighted linear combination of overlaps from the target utterance, preceding supporter/seeker turns, broader context, and a defined set of 17 rule-extracted discourse cues. To ensure context diversity and avoid redundancy, we employ a Maximal Marginal Relevance (MMR) selection procedure (Carbonell and Goldstein, 1998) with a diversity factor $\lambda = 0.20$. These dynamic examples are then merged with static class-representative anchor examples, capped at a maximum of 6 examples per prompt.

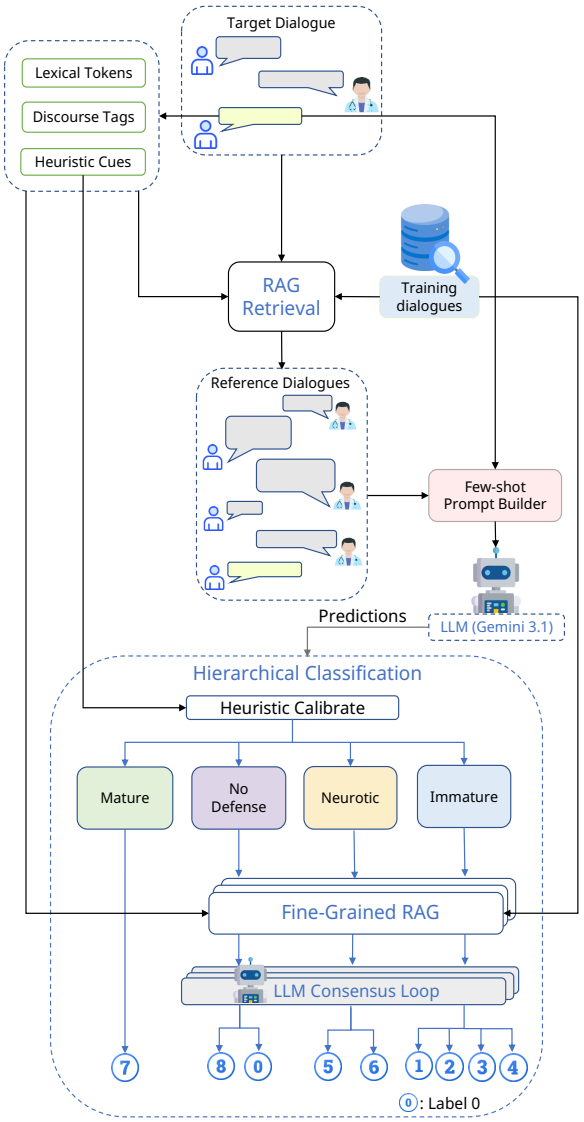

Figure 1: Architecture of the Hierarchical Retrieval-Augmented LLM Inference pipeline.

The final prompt contains reconstructed context, retrieved examples, and their explanations.

Inference strategies. On top of this pipeline, we explore two strategies using the same Gemini 3.1 Flash Lite model:

- *Direct LLM-RAG*: Directly predict the final label from the constructed prompt.

- *Hierarchical LLM-RAG:* Motivated by the severe label imbalance and the principle that decomposing complex tasks better elicits the

knowledge of LLMs (Wei et al., 2022), we decompose prediction into two sequential steps (illustrated in Figure 1): (i) predict a coarse group, and (ii) predict the final label within that group. We define four coarse-grained categories based on the DMRS taxonomy—*No Defense* (labels 0, 8), *Mature* (label 7), *Neurotic* (labels 5–6), and *Immature* (labels 1–4)—reflecting the skewed distribution in PSYDEFCONV (Na et al., 2026b). This reduces competition across semantically distant labels and focuses fine-grained prediction on a smaller, more coherent subset of classes.

Heuristic calibration. While LLMs excel at implicit semantics, they can over-analyze simple utterances or hallucinate. To ground the model, hard-coded heuristics based on 17 discourse cues are used to calibrate or override its predictions:

- *Rule 1 (Phatic/Logistical → Labels 0 or 8):* If the utterance is dominated by *phatic*, *gratitude*, or *closing* cues (e.g., "thank you", "goodbye") without substantive defense signals, it is strictly routed to *No Defense* (Label 0) or *Needs More Information* (Label 8) if exceptionally short (≤ 2 tokens).
- *Rule 2 (Action Defenses → Label 1):* If the utterance strongly triggers *barrier* or *rejecting_suggestion* cues immediately following a supporter's *suggestion* turn, predictions are calibrated toward *Action Defenses* (Label 1).
- *Rule 3 (Neurotic Overrides → Labels 5 or 6):* If the LLM predicts *No Defense*, but the text contains *abstract_control* or a mix of *fact_logistic* and *affect* tokens, the coarse group is explicitly overridden to *Neurotic Defenses*.
- *Rule 4 (Consensus Fallback):* If the fine-grained LLM fails to output a valid label within the constrained taxonomy after maximum retries, the system completely overrides the LLM and defaults to the highest-weighted heuristic label.

On the development set, we observed that these overrides are most frequently triggered under two conditions: when the LLM over-analyzes short conversational turns by assigning complex defensive mechanisms (resolved by Rule 1), and when the model fails to produce a valid output within the allowed taxonomy (resolved by Rule 4).

3.3 Approach 3: Hybrid Supervised–LLM Pipeline

Building on Approach 2, we further combine the LLM-based method with supervised filtering from Approach 1. Our design is based on two observations:

- *Frequent, easy labels (0 and 7):* These labels are abundant and relatively simple, so supervised models trained on gold labels can learn them reliably. Thus, we use the model from Approach 1 to filter them first.
- *Rare, hard labels:* The remaining labels are fewer and more ambiguous, making them difficult to learn from limited data. Hence, we apply the LLM-RAG pipeline to leverage external knowledge and avoid bias toward dominant labels.

Based on these insights, we design a hybrid pipeline: a supervised model trained on gold labels first filters samples predicted as 0 or 7, and the LLM-RAG pipeline is applied only to the remaining cases.

Although this hybrid design is conceptually appealing and better aligns model capacity with label difficulty, our experiments show that the pure LLM-RAG pipeline (Approach 2) remains more robust overall.

3.4 Approach 4: Summary Distillation (Auxiliary)

As an auxiliary direction, we explore whether LLMs can improve supervised classification indirectly via representation learning.

Motivated by the long and noisy nature of dialogue context, and the known tendency of language models to lose critical information within the intermediate parts of extended sequences (Liu et al., 2024), we use the *Phi-4 14B* model to generate a short, task-oriented summary of the preceding context for each target turn.

- Focus on the *target utterance* rather than the full dialogue.
- Remain *grounded in the dialogue context*.
- Describe the *functional role* of the utterance.

The resulting summary captures the main stressor, the role of the utterance, and its relation to prior

Method	Acc.	Prec.	Rec.	F1
Supervised Encoders				
Single-level	58.69	18.27	14.08	13.27
2-level hierarchical	**61.23**	26.63	22.43	22.92
3-level hierarchical	59.11	27.53	26.51	26.49
LLM-RAG				
Direct	58.69	31.58	24.44	26.55
Hierarchical	48.31	**41.87**	27.73	**31.13**
Hybrid Pipeline				
Filter-0,7 + RAG	58.26	28.59	**30.19**	26.83
Summary Distillation				
Phi-4 summary	52.14	25.13	23.65	23.52

Table 1: Experimental results on the test set (%). F1 is macro-averaged over all positive classes (all classes except label 0).

context in a compact form. A RoBERTa classifier is then trained on these summaries.

This direction treats the LLM as a task-aware compressor, producing a concise and discriminative representation for downstream classification.

4 Experimental Results

Our official submission to the PsyDefDetect shared task corresponds to *Approach 2 (Hierarchical LLM-RAG)*, which we identify as our primary system based on development set performance. With this configuration, our system achieved an official rank of 10 out of 21 participating teams on the shared task leaderboard.

Table 1 shows that the *hierarchical LLM-RAG* is the strongest overall approach, achieving the best Macro-F1 (*31.13%*) and Precision (*41.87%*). This confirms that coarse-to-fine prediction is more effective than direct classification for PsyDefDetect, where labels are highly imbalanced and often implicitly expressed.

Among the supervised models, the single-level classifier performs worst, while hierarchical filtering consistently improves results. In particular, the 2-level model achieves the best Accuracy (61.23%), and the 3-level model further improves Macro-F1 to 26.49%, showing that separating frequent labels before final prediction is beneficial.

A similar pattern appears for LLM-based methods. The dialogue-level RAG model already surpasses all supervised baselines in Macro-F1, and its hierarchical extension yields a further clear gain. Although the hybrid pipeline achieves the best Recall (30.19%), it remains below the pure hierarchical LLM-RAG in Macro-F1. This suggests that early supervised filtering may improve coverage but also propagates errors. Finally, the summary distillation approach improves over the weakest supervised baseline but remains less effective than direct LLM-RAG inference.

As observed in Table 1, there is a clear trade-off between overall accuracy and balanced class performance. Although the hierarchical LLM-RAG framework improves Macro-F1 from 26.55% to 31.13%, its Accuracy decreases from 58.69% to 48.31%. This improvement is primarily driven by better recognition of minority classes, especially Class 8, whose F1 score increases from 0.00% to approximately 60% after the hierarchical reformulation. Additional gains are also observed for Class 5 (+7 points) and Class 2 (+6 points). In contrast, the reduction in Accuracy is largely caused by lower performance on Class 7 (-12 points), which constitutes the majority of the test set (243 out of 397 samples, approximately 61%). Overall, these findings suggest that the hierarchical formulation shifts the model toward more balanced predictions across classes, sacrificing some performance on the dominant class in exchange for substantially improved detection of rare defense mechanisms under severe class imbalance.

5 Conclusion

In this work, we investigated four directions for psychological defense detection under severe label imbalance and the implicit, context-dependent nature of defensive behaviors. Our core method is a retrieval-augmented LLM pipeline that operates at the dialogue level, with additional hybrid and summary-distillation extensions.

Experiments show that retrieval-augmented LLM methods outperform supervised baselines, but fine-grained defense detection remains challenging, especially for less frequent labels and subtle discourse-level distinctions. Although the hybrid design is conceptually appealing, the LLM-RAG pipeline remains the most robust overall.

These findings highlight both the promise and the current limitations of LLM-based methods for modeling latent psychological constructs in dialogue. Future work should focus on stronger pragmatic grounding and richer context modeling for fine-grained clinical language understanding.

Limitations

Our framework has three notable limitations. First, the hierarchical routing strategy improves minority-class performance but degrades accuracy on the dominant class (Class 7, ~60% of samples). This occurs because errors made at the coarse-grained grouping stage misroute majority-class instances into wrong categories, which heavily damages overall accuracy due to the severe class imbalance. Second, the heuristic calibration layer — relying on 17 discourse cues — achieves high precision on surface-level patterns but lacks semantic flexibility, failing to catch implicit defense mechanisms without clear lexical markers. Finally, our reliance on a closed-source commercial LLM limits data privacy and long-term reproducibility, restricting its immediate use in strict clinical environments where open-weight, local alternatives are preferred.

References

Eliseo Bao, Anxo Pérez, and Javier Parapar. 2024. Explainable depression symptom detection in social media. *Health Information Science and Systems*, 12(1):47.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

Zhiyu Chen, Yujie Lu, and William Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304.

Mariagrazia Di Giuseppe, Katie Aafjes-van Doorn, Vera Békés, Bernard S Gorman, Karl Stukenberg, and Sherwood Waldron. 2024. Therapists' defense use impacts their patients' defensive functioning: a systematic case study. *Research in Psychotherapy: Psychopathology, Process, and Outcome*, 27(2):797.

Mariagrazia Di Giuseppe and J. Christopher Perry. 2021. The hierarchy of defense mechanisms: Assessing defensive functioning with the defense mechanisms rating scales q-sort. *Frontiers in Psychology*, 12.

Isaac R. Galatzer-Levy, Daniel McDuff, Vivek Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. 2023. The capability of large language models to measure psychiatric functioning. *arXiv preprint arXiv:2308.01834*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pages 3469–3483.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13641–13660.

Marta Mrozowicz-Wrońska. 2023. Defense mechanisms in affective disorders-the state of the art. *Psychiatria Polska*, 57(1):197–206.

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

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

Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you

only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.

J Christopher Perry. 2014. Anomalies and specific functions in the clinical identification of defense mechanisms. *Journal of clinical psychology*, 70(5):406–418.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Vishal Anand Shetty, Shauna Durbin, Meghan S Weyrich, Airín Denise Martínez, Jing Qian, and David L Chin. 2024. A scoping review of empathy recognition in text using natural language processing. *Journal of the American Medical Informatics Association*, 31(3):762–775.

Sichang Tu, Abigail Powers, Natalie Merrill, Negar Fani, Sierra Carter, Stephen Doogan, and Jinho D Choi. 2024. Automating ptsd diagnostics in clinical interviews: Leveraging large language models for trauma assessments. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 644–663.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 8(1):1–32.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077.