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

Liudmila Babakova
Moscow State Pedagogical University
babakovalyuda@gmail.com

Christopher Luongo-Vázquez
Universidad de La Rioja
chluongo@unirioja.es

Ilia Stepin
Universidad Autónoma de Madrid
ilia.stepin@uam.es

Abstract

Psychological defense detection is one of essential present-day challenges in clinical practice. The state-of-the-art natural language processing (NLP) tools aim to automate this task. However, their potential and efficiency remain largely unexplored. This manuscript attempts to address this problem from various perspectives: it first explores the efficiency of direct large language model (LLM)-prompting. Then, it applies NLP techniques for LLM fine-tuning applied to the psychological defense classification task. Finally, it attempts to generate states of mind based on the speaker's psychological state. The results show that the complexity of the task requires further improvement of the software solutions used.

1 Introduction

One of the present-day psychotherapy challenges consists in automating psychological defense detection. Large language models (LLMs) have been shown to have great potential to revolutionize psychotherapy. Nevertheless, their effectiveness, when applied to treating mental health issues, remains underexplored, as they have only been shown to offer static solutions to single mental health-related problems (Na et al., 2025).

This manuscript presents (1) a description of the methodological aspects, (2) experimental design as well as (3) the empirical results obtained while participating in the shared task PsyDefDetect (Na et al., 2026a). Essentially, it explores the potential of some of the most widely used LLMs applied to a corpus of multi-turn dialogues that are labeled to highlight a possible psychological defense mechanism (Na et al., 2026b).

The remainder of the manuscript is structured as follows. Section 2 describes the methodology applied as well as the experimental settings. Section 3 reports the results obtained. Finally, section 4 concludes the manuscript.

2 Methodology

Four series of experiments have been carried out as part of this work. First, prompting-related experiments took place for an LLM to directly classify the given (non-)defensive utterances (see Section 2.1 for details). Subsequently, encoder fine-tuning experiments were conducted to estimate the effect of hyperparameter fine-tuning of the given language models (see Section 2.2 for details). Afterwards, a novel technique for state-of-mind generation was proposed and tested in the context of psychological defense classification (see Section 2.3 for details). Finally, LLM fine-tuning makes part of the concluding series of experiments. The corresponding methodology is described in Section 2.4.

All the experiments were carried out using the PsyDefConv dataset (Na et al., 2026b) containing nine classes (0 – no defense, 1 – action defenses, 2 – major image-distorting, 3 – disavowal defenses, 4 – minor image-distorting, 5 – neurotic defenses, 6 – obsessional defenses, 7 – high-adaptive defenses, 8 – needs more information).

2.1 LLM prompting

The prompting experiments were designed to include three internal stages: (1) direct prompting, (2) few-shot prompting, and (3) two-step psychological defense classification. Four baseline models have been used in all the corresponding experiments. Namely, these are DeepSeek-v3.2 (DeepSeek-AI et al., 2025), Qwen3 (Qwen3-235B-A22B) (Team, 2025), Qwen2.5-72B-Instruct (Team, 2024), and GLM-5 (GLM-5-Team et al., 2026). First, all the aforementioned models were prompted to predict the type of psychological defense mechanism (if any). Subsequently, they were enhanced with several input examples to enable in-context learning. Finally, two-step predictions were made. First, all the utterances underwent a binary classification task (defensive vs.

Proceedings of the BioNLP 2026 (Shared Tasks), pages 104–108
July 3-4, 2026 ©2026 Association for Computational Linguistics

non-defensive). Second, all the utterances claimed defensive were subsequently assigned labels of the corresponding defensive dataset classes. To do so, both binary and multiclass classification requests were enhanced with few-shot prompts generated in stage (2).

2.2 Encoder fine-tuning

In the second phase, we evaluate encoder-based models on a simplified binary formulation of the task aimed at improving recognition of non-defensive utterances. We compare two labeling strategies. In the *Unmerged* setting, models are trained on the original multiclass labels and later evaluated with respect to the binary distinction. In the *Merged* setting, label 0 (*No Defense*) is preserved, while all non-zero labels are collapsed into a single positive class representing the presence of any defense mechanism.

Several transformer encoders were used in the corresponding experiments. Namely, those are BERT (Devlin et al., 2018), ModernBERT (Warner et al., 2024), RoBERTa (Liu et al., 2019), and DeBERTa-v3 (He et al., 2021). For each model and labeling setup, we explore the effect of setting multiple learning rates (2×10^{-4}, 2×10^{-5}, and 2×10^{-6}) in order to assess optimization sensitivity and identify robust configurations. The best-performing binary classifier from this phase is later incorporated as the filtering component in our hybrid retrieval-based system.

2.3 State-of-mind generation

As part of the original pipeline, we additionally propose a hybrid framework that combines LLM-based semantic abstraction, instance-based classification, and binary filtering. First, each dialogue context together with the target utterance is transformed into a structured *state-of-mind* representation using GPT-5.2. Rather than operating directly on raw dialogues, the model generates a compact description of the speaker's psychological state, intentions, and coping behavior, providing a normalized representation in which defensive patterns become more comparable across instances.

We then perform classification in this representation space using a k-nearest neighbors (KNN) strategy. The generated state-of-mind descriptions are embedded with Gemma (Schechter Vera et al., 2025)[1]. As for each test example, we retrieve the nearest training representations ($k \in 3, 5, 7$). The final label is assigned by majority voting over the retrieved neighbors. This is argued to allow the system to leverage semantic similarity between psychologically related situations instead of relying only on surface lexical cues.

Finally, to better address the strong class imbalance, we incorporate the best-performing binary classifier from Section 2.2 (RoBERTa *Merged*, learning rate=2×10^{-5}), which distinguishes *No Defense* (Level 0) from *Any Defense* (Levels 1–8). In the final hybrid setup, utterances predicted as *No Defense* are directly assigned label 0, while all the remaining instances are classified using the KNN module. In this way, the encoder serves as a precision-oriented gate for trivial non-defensive cases, while the retrieval component handles finer-grained defense level prediction.

2.4 LLM fine-tuning

In the concluding phase of the experiments, a single baseline LLM (Llama-3.2-3B) (AI@Meta, 2024) was fine-tuned to assess the corresponding effect on psychological defense mechanism classification. The quantized low-rank adaptation technique (QLoRA) (Dettmers et al., 2023) was used to fine-tune the baseline LLM. Two learning rates (1×10^{-4} and 1×10^{-5}) were employed in the corresponding experiments.

3 Experimental results

This section reports the experimental results of all the four groups of the experiments whose methodology is described in section 2. Table 1 reports precision, recall, and F1-scores for all the baseline models tested within all the prompting experiment stages. It is worth noting that DeepSeek performance increases, as few-shot prompting and two-step classification are incorporated at later stages of the experiment whereas this effect is not observed in the case of all the other baseline models used.

Table 2 overviews the encoder fine-tuning experimental results. Remarkably, all the LLMs tested show better performance in the *Merged* setting.

Table 3 presents the state-of-mind generation experimental results as a function of the number of k-nearest neighbors. It can be seen that the model's performance slightly increases along with the increase in the number of neighbors.

Table 4 reports the effects of the QLoRA-based LLM fine-tuning. These do not appear to signifi-

[1] huggingface.co/google/embeddinggemma-300m

Model	Precision	Recall	F1-score
Stage 1: Direct prompting (baseline models)			
DeepSeek-v3.2	0.182	0.213	0.155
Qwen3	0.238	**0.226**	**0.180**
Qwen2.5-72B	0.192	0.204	0.152
GLM-5	**0.348**	0.216	0.178
Stage 2: Few-shot prompting			
DeepSeek-v3.2	**0.369**	**0.259**	0.194
Qwen3	0.328	0.248	**0.203**
Qwen2.5-72B	0.344	0.221	0.175
GLM-5	0.302	0.209	0.172
Stage 3: Defense classification			
DeepSeek-v3.2	**0.276**	0.133	0.083
Qwen3	0.204	**0.137**	0.071
Qwen2.5-72B	0.073	0.112	0.031
GLM-5	0.247	0.148	**0.088**

Table 1: First-phase experimental results (prompting).

Labeling	Learning rate	F1-score
Model: *BERT*		
Unmerged	2e-4	0.076
	2e-5	0.221
	2e-6	0.138
Merged	2e-4	0.454
	2e-5	0.836
	2e-6	0.794
Model: *ModernBERT*		
Unmerged	2e-4	0.159
	2e-5	0.234
	2e-6	0.131
Merged	2e-4	0.857
	2e-5	0.853
	2e-6	0.774
Model: *RoBERTa*		
Unmerged	2e-4	0.076
	2e-5	0.215
	2e-6	0.122
Merged	2e-4	0.454
	2e-5	0.913
	2e-6	0.796
Model: *DeBERTa-v3*		
Unmerged	2e-4	0.030
	2e-5	0.030
	2e-6	0.076
Merged	2e-4	0.144
	2e-5	0.144
	2e-6	0.454

Table 2: Second-phase experimental results (encoder fine-tuning).

# neighbors	Precision	Recall	F1-score
3	0.174	0.173	0.170
5	0.175	0.176	**0.171**
7	**0.190**	**0.177**	**0.171**

Table 3: Third-phase experimental results (LLM-based state-of-mind representation).

Learning rate	Precision	Recall	F1-score
1e-4	**0.139**	**0.201**	**0.164**
1e-5	0.130	0.135	0.116

Table 4: Fourth-phase experimental results (LLM fine-tuning).

cantly differ depending on the changes in the learning rate employed.

Based on the insights from the experiments carried out, the final, 7-neighbor state-of-mind generation model was passed on to the shared task organizers so that the test results could be calculated. These are presented in Table 5.

4 Conclusion

The predominantly negative results obtained emphasize the complexity of the psychological defense mechanism classification task. They highlight, for instance, the need for extensive use of data preprocessing techniques. It is particularly necessary when using the provided dataset due to a high degree of class imbalance among the data instances. In addition, better elaborated fine-tuning mechanisms are needed to improve the performance of the models tested.

As part of future work, a more thorough comparative analysis of various language models seems required. The experimental results do not allow us to conclude how effective the presented conceptualization of LLM-based state-of-mind representations is. That said, they pave the way toward further refining of the methodological basis of the experiments presented in this manuscript.

Limitations

Regarding the first-phase experiments (prompting), these only include a limited set of LLMs. A com-

Accuracy	Precision	Recall	F1-score
0.6144	0.2366	0.1660	0.1612

Table 5: Test results.

parison against a wider set of models as well as the use of alternative techniques reshaping the original prompts might help improve the performance.

As for the second-phase experiments (encoder fine-tuning), only a limited number of models and hyperparameters has been used for fine-tuning. To be precise, only the learning rate for four LLMs has been modified while running the corresponding experiments. In addition, the experimental results reported do not allow us to make conclusions on their statistical significance.

When it comes to the third-phase experiments (state-of-mind generation), these employ only one classic machine learning classifier. It is therefore necessary to explore the potential of state-of-the-art classifiers to ensure adequate classification of, for example, defensive vs. non-defensive cases. In addition, the results reported indicate that improved data preprocessing might help improve the model performance, particularly when dealing with class imbalance.

Finally, the fourth-phase experiments (LLM fine-tuning) include only one preselected LLM. In both experiments carried out, only the learning rate was changed as part of fine-tuning. This may be argued not to be sufficient to improve the discriminatory capacity of the corresponding classifier.

References

AI@Meta. 2024. Llama 3 model card.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLORA: efficient finetuning of quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

GLM-5-Team, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, and 167 others. 2026. GLM-5: from Vibe Coding to Agentic Engineering. *Preprint*, arXiv:2602.15763.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *Preprint*, arXiv:2111.09543.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

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

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

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

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 69 others. 2025. EmbeddingGemma: Powerful and Lightweight Text Representations. *Preprint*, arXiv:2509.20354.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *Preprint*, arXiv:2412.13663.

A LLM prompts

This appendix reports the LLM prompts used across the first- and third-phase experiments.

In the first-phase experiments, the following prompt was used: "You are a strict classifier. Choose exactly one label: 0 or 1. Use the following guide: 0 - No Defenses: Functional utterances that maintain conversational flow without engaging conflict. 1 - Action Defenses: Passive Aggression, Help-Rejecting Complaining, Acting Out. Distress is released by acting on the environment instead of reflecting. 2 - Major Image-Distorting: Splitting (self/other), Projective Identification. Reduces anxiety via all-good/all-bad distortions of self or other. 3 - Disavowal Defenses: Denial, Rationalization, Projection, Autistic Fantasy. Rejects threatening reality by denying, excusing, blaming, or fantasizing. 4 - Minor Image-Distorting: Devaluation/Idealization (self or other), Omnipotence. Softer distortions temporarily inflate or deflate self-esteem. 5 - Neurotic Defenses: Repression, Dissociation, Reaction Formation, Displacement. Keeps unacceptable motives out of awareness; feelings surface indirectly. 6 - Obsessional Defenses: Isolation of Affect, Intellectualization, Undoing. Uses excessive logic or symbolic acts to separate feelings from events. 7 - High-Adaptive Defenses: Affiliation, Altruism, Anticipation, Humor, Self-Assertion, Self-Observation, Sublimation, Suppression. Mature coping that integrates emotion and thought to channel affect constructively. 8 - Needs More Information: Label used when an utterance is too ambiguous or lacks context. Return only JSON in one line: {"label": <0 or 1>, "reason": "short reason"}."

In the third-phase experiments, the following prompt was used: "You are a psychologist, skilled in Cognitive Behavioral Therapy (CBT), Eye Movement Desensitization and Reprocessing (EMDR) protocol, Acceptance and Commitment Therapy, and Mindfulness practices, providing empathetic guidance in psychotherapy. Your task is to briefly describe the current state of mind of the seeker based on the ongoing conversation between seeker and supporter. Focus on features most relevant to identifying psychological defense mechanisms — specifically: how the seeker relates to distressing emotions (avoids, intellectualizes, acts out, suppresses, projects, denies, or integrates them); whether their perception of self or others appears distorted (idealized, devalued, split into all-good/all-bad); how they respond to help (accept, resist, reject); the degree of self-awareness and reflective capacity; and any signs of fantasy, displacement, rationalization, or mature coping such as humor or self-assertion. Describe in 2-4 sentences."