Eraserhead at PsyDefDetect: Prompt Design and Class Rebalancing for Psychological Defense Mechanism Detection

Muhammad Abu Horaira, Mehreen Rahman, Nahian Chowdhury
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004029, u2004033, u2004026}@student.cuet.ac.bd

Abstract

We describe the Eraserhead system submitted to the PsyDefDetect shared task at BioNLP 2026, which frames psychological defense level detection as a nine-class utterance classification problem over supportive dialogue. Our system is based on Qwen3-14B and combines clinically informed prompt design, per-label oversampling, and careful inference settings for stable prediction. A central challenge of the task is strong class imbalance, with High-Adaptive responses appearing far more often than several minority classes. This makes it easy for models to favor the majority class and achieve reasonable accuracy while performing poorly on rarer categories. To address this, we iteratively adjusted oversampling targets based on error analysis and predicted label distributions across submission rounds. Our final system achieved an official macro F1 of 0.3418 on Leaderboard 1 and 0.3947 on Leaderboard 2, ranking 7th among the 21 registered teams on both leaderboards. We further analyze the main failure modes of the system, especially the difficulty of distinguishing Minor Image Distorting defenses from High-Adaptive responses and the persistent tendency to over-predict the majority class. These findings highlight the broader difficulty of modeling psychological function from text alone.

1 Introduction

Psychological defense mechanisms are unconscious processes through which individuals manage anxiety and emotional distress in interpersonal interaction (Vaillant, 1992). The Defense Mechanism Rating Scales (Perry, 1990) provide a clinically grounded hierarchy of defense levels. Recent work has also shown growing interest in the use of large language models in psychotherapy related settings (Na et al., 2025).

The PsyDefDetect shared task (Na et al., 2026a) formulates this problem as a nine class utterance level classification task over supportive conversations. The task is challenging because label distinctions depend on psychological function rather than surface wording, while the class distribution is also highly imbalanced.

Our system, Eraserhead (CodaBench: `nahian_abu`), combines clinically informed prompt design, per-label oversampling, and careful inference design. The main contributions of this work are as follows:

- We develop a theoretically grounded prompt design based on DMRS, incorporating explicit label definitions and targeted disambiguation rules for fine-grained psychological classification.
- We introduce a per-label oversampling strategy with empirically tuned target distributions to reduce majority class bias and improve representation of minority defense categories.
- We provide a focused analysis of prediction patterns and failure modes, highlighting the difficulty of distinguishing subtle defense mechanisms from text alone.

2 Related Work

Recent advances in NLP have substantially expanded the scope of computational mental health and psychological text analysis. Early work in this area often relied on domain-adapted transformer models trained on mental health-related social media data. A representative example is MentalBERT (Ji et al., 2022), which was pre-trained on social media data and showed improved performance across several mental health detection benchmarks. Despite these results, its reliance on social media corpora introduces domain bias and may limit generalization to other settings, especially supportive or therapeutic conversations. Such models are also

Proceedings of the BioNLP 2026 (Shared Tasks), pages 54–58

typically optimized for symptom or condition detection rather than for identifying deeper psychological processes.

Subsequent studies explored improved architectures and broader applications. DEENT (Narvaez Burbano et al., 2025) was proposed for depression detection on social media and showed gains over traditional machine learning baselines. Likewise, Ajayi et al. (2025) examined the detection of mental health conditions and cyberbullying from social media using machine learning methods. Another line of work has emphasized domain-adaptive pre-training. Chinese MentalBERT (Zhai et al., 2024), for example, leveraged large-scale social media corpora to improve Chinese mental health text analysis and outperformed several baseline models. Related work on expressive narrative stories (Tang et al., 2024) further suggests that language models can capture subtle linguistic patterns in emotionally rich text, while still showing limitations in deeper semantic reasoning.

Overall, prior work demonstrates that transformer-based models are effective for mental health text classification, especially in social media settings. However, most studies focus on surface-level categorization rather than fine-grained psychological mechanisms. In contrast, our work addresses defense mechanism classification grounded in the Defense Mechanism Rating Scales, where labels may appear similar on the surface but differ in underlying psychological function. This makes conversational context, clinically informed prompt design, and class imbalance handling especially important.

3 Task and Data

PsyDefDetect is a nine-class utterance-level classification task for detecting levels of psychological defense mechanisms in supportive conversations (Na et al., 2026b). Given a short dialogue context and a target utterance from the help-seeker, the model must predict one DMRS label.

The task is challenging for two reasons. First, the distinctions between labels are psychologically subtle and often depend more on underlying function than on surface wording. Second, the label distribution is highly imbalanced, with High-Adaptive responses appearing much more often than several minority classes. This imbalance makes prediction of the rarer categories especially difficult and motivates the rebalancing approach used in our system.

4 System Overview

Our classification system is based on Qwen3-14B[1], which we fine-tuned with 4-bit QLoRA on Kaggle T4 GPUs.

4.1 Model and Fine-Tuning

We fine-tuned Qwen3-14B using Low-Rank Adaptation through the Unsloth library[2], which enables memory-efficient 4-bit quantized training. LoRA adapters were applied to all attention and feed-forward projection matrices (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`) with LoRA rank 16 and scaling factor 16. Training used AdamW 8-bit with a learning rate of 2×10^{-4}, cosine decay, an effective batch size of 4 (per-device batch size 1 with gradient accumulation of 4), 3 epochs, and a 2,048-token context window.

4.2 Class-Rebalanced Oversampling

The training set is highly imbalanced, with Label 7 (High-Adaptive) dominating several other classes. When trained on the original distribution, the model tends to favor the majority class, producing high accuracy but poor macro F1. This pattern is also visible on the leaderboard, where many systems achieve relatively strong accuracy while still struggling on macro F1.

To address this, we used per-label oversampling with empirically tuned target counts. Minority classes were oversampled to improve representation, while the majority class (Label 7) was capped below its natural frequency to reduce its dominance without removing it entirely. This helped preserve calibration while limiting majority-class bias. These target counts were adjusted through iterative empirical tuning across multiple submission rounds. After each run, we compared the predicted label distribution on the test set with the training prior. Over-predicted labels had their targets reduced, while under-predicted labels had their targets increased. We did not use a separate held-out validation split for these adjustments; instead, oversampling targets were revised across official submission rounds by comparing predicted label distributions with the training prior. Table 1 summarizes this process across two rounds.

[1] https://huggingface.co/unsloth/Qwen3-14B-unsloth-bnb-4bit

[2] https://github.com/unslothai/unsloth

L	Name	v1	v3 Error	v4
0	No Defenses	200	+5.5%	200
1	Action Defenses	180	−0.5%	180
2	Major Image-Distorting	140	−1.0%	140
3	Disavowal Defenses	130	+6.6%	130
4	Minor Image-Distorting	160	+3.3%	**80**
5	Neurotic Defenses	120	−0.4%	120
6	Obsessional Defenses	220	+0.8%	220
7	High-Adaptive Defenses	600	−8.3%	**900**
8	Needs More Information	70	−0.5%	70
	Total	**1820**		**2040**

Table 1: Oversampling targets across versions. "v3 Error" denotes predicted frequency minus training frequency after the v3 submission, where positive values indicate over-prediction and negative values indicate under-prediction. Boldface highlights the targets changed for v4: Label 4 was reduced after over-prediction, while Label 7 was increased after under-prediction.

One important finding is that Label 7 was difficult to calibrate reliably. In v3, it remained under-predicted relative to its training prior, which motivated increasing its oversampling target in v4. This likely reflects an asymmetry in the DMRS framework: other labels require *positive evidence* for a specific defense mechanism, whereas Label 7 is often assigned when such evidence is absent. As a result, an LLM biased toward cooperative and interpretive readings may still default too easily to a more adaptive classification.

4.3 Prompt Design

Prompt design was the most influential part of the system. The classification prompt evolved through four versions shaped by error analysis from prior submissions: v1 used basic label descriptions, v2 added full label definitions with sub-mechanism names, v3 introduced targeted Critical Distinctions rules, and v4 refined those rules while adding the explicit rule-out step described below.

The final prompt (v4) is structured as a clinical rubric with three parts. Key Principles instructs the model to focus on *psychological function* rather than surface wording and clarifies that direct emotional expression does not by itself constitute a defense. Full Label Definitions list all nine DMRS levels together with their sub-mechanisms, which reduced uncertainty more effectively than high-level descriptions alone. Finally, Critical Distinctions provides eight targeted rules for the label pairs most often confused in earlier submissions:

- *Label 0 vs. 7:* A specific Label 7 sub-mechanism must be identifiable; otherwise, assign 0.
- *Label 1 vs. 7:* Label 1 rejects all offered solutions while complaining; Label 7 is genuinely open to help.
- *Label 3 vs. 4:* Label 3 involves evasion of a *fact or reality*, whereas Label 4 distorts the *image of self or others*.
- *Projection vs. Displacement (3 vs. 5):* Projection means the speaker is unaware of their own feeling and attributes it to others; Displacement means the feeling is recognized but redirected toward the wrong target.
- *Label 4 vs. 2:* Label 4 preserves some degree of reality testing, whereas Label 2 is more absolute and leaves little room for ambivalence.
- *Label 6 vs. 7 (Isolation of Affect):* Label 6 implies that emotion is genuinely absent; Label 7 still shows engagement, even when understated.
- *Label 6 (Undoing) vs. 7:* Undoing tends to appear hesitant and self-contradictory rather than analytical.
- *Plain self-disclosure:* Purely descriptive reporting of a past feeling with no distortion or mature coping mechanism defaults to Label 0; genuinely unclassifiable utterances should be assigned Label 8.

The prompt instructs the model to identify the psychological function, rule out the closest alternative, and output a single digit from 0 to 8. This rule-out step was added to discourage the model from selecting the first plausible label.

4.4 Inference

At inference time, the prompt is wrapped in the model's native chat template with `add_generation_prompt=True` and `enable_thinking=False`. We disabled Qwen3's thinking mode because, in early experiments, chain-of-thought generation often shifted predictions toward overly charitable High-Adaptive interpretations. Non-thinking mode reduced this tendency and produced more stable, more easily parseable single-digit outputs.

System	LB1-Acc	LB1-F1	LB2-F1
Nürnberg NLP (1/21)	0.7013	0.4200	0.4732
UTS (2/21)	0.6737	0.4055	0.4450
PerceptionLab (3/21)	0.6737	0.3956	0.4402
zzucs (6/21)	0.6441	0.3585	0.4135
Eraserhead (7/21)	**0.6462**	**0.3418**	**0.3947**
zzunlp (8/21)	0.6758	0.3300	0.3909

Table 2: Official leaderboard results. LB1 = Leaderboard 1 for positive classes (Labels 1–8), and LB2 = Leaderboard 2 for all classes. Rankings are reported with respect to the 21 registered teams in the official evaluation.

We decoded with temperature 0.1 and a maximum of 10 new tokens. The predicted label was extracted using a two-stage regular expression: first matching `label: [0-8]`, then falling back to the first standalone digit in the valid range.

5 Results

5.1 Main Results

Table 2 presents our official leaderboard results together with selected comparison systems.

Our system achieved an official LB1 macro F1 of 0.3418 and an official LB2 macro F1 of 0.3947, placing it 7th among the 21 registered teams in the official evaluation. Its LB1 accuracy of 0.6462 is also reasonably close to that of several higher-ranked systems, such as zzunlp at 0.6758. This suggests that the model is fairly competitive in overall agreement with the gold labels, but still falls behind the strongest systems in handling minority classes, particularly in terms of precision and recall.

6 Analysis

6.1 The Label 4 Failure: Why Minor Image-Distorting Is the Hardest Class

The clearest weakness in our system is Label 4 (Minor Image-Distorting), which was predicted less often than expected. This class requires the model to detect an *unrealistically* inflated or deflated view of a person while preserving some reality testing, a distinction that often depends on broader emotional context unavailable in a short conversation window.

These cases are difficult because their surface language can still appear reasonable or adaptive. Our model often resolved this ambiguity in favor of Label 7, the safer prediction under the imbalanced training distribution. Reducing the Label 4 oversampling target in v4 then pushed the model toward under-prediction, suggesting that the boundary between Labels 4 and 7 is not well captured by frequency calibration alone.

6.2 Label 7 as a Fallback Class

Even after increasing the oversampling target for Label 7, this class remained difficult to control consistently. This suggests that the problem is not purely one of class frequency, but also of decision boundary definition. In the DMRS framework, other labels require *positive evidence* for a specific defense mechanism, whereas Label 7 is selected when such evidence is weak or absent. As a result, Label 7 becomes a natural fallback under uncertainty. Because LLMs tend toward cooperative and charitable interpretations, they may be especially likely to resolve ambiguous cases in favor of a more adaptive reading.

6.3 Confusion Among Labels 3, 6, and 7

We observe persistent confusion among Labels 3 (Disavowal), 6 (Obsessional), and 7 (High-Adaptive). Although these labels often look similar on the surface, they differ in psychological function: Label 3 avoids an uncomfortable truth, Label 6 is analytical but affectively detached, and Label 7 reflects genuine emotional engagement.

This distinction is difficult to recover from text alone, since affect is often conveyed through prosody and other nonverbal cues that are absent from transcripts. As a result, models fall back on surface heuristics, such as mapping analytical language to Label 6 and positive framing to Label 7. This reflects a broader difficulty of text-only DMRS assessment rather than a limitation of our system alone.

7 Conclusion

We presented Eraserhead, a Qwen3-14B based system for the PsyDefDetect shared task that combines clinically informed prompt design, class rebalanced oversampling, and careful inference settings. Our results show that this approach can achieve competitive performance, while our analysis highlights persistent difficulty around Label 4 and the tendency to over-predict Label 7. More broadly, the task underscores the challenge of inferring psychological function from text alone, especially under strong class imbalance.

Limitations

This work has several limitations. The task is based on text-only dialogue, which excludes prosodic and nonverbal cues that may be important for distinguishing psychological defense mechanisms. This especially affects subtle distinctions among Labels 3, 6, and 7, as well as the boundary between Label 4 and Label 7.

Our class-rebalancing strategy was tuned across official submission rounds by comparing predicted label distributions with the training prior, without using a separate held-out validation split. Although effective in practice, this limits how strongly we can interpret the results in terms of generalization.

Finally, our results are based on a single model family and a specific prompt design, so they may not transfer directly to other architectures or prompting setups. More broadly, reliable text-only classification remains challenging because labels can appear similar on the surface while differing in psychological function.

Ethical Considerations

This work examines automatic detection of psychological defense mechanisms in supportive conversations. Since the task relates to psychological functioning, model predictions should not be treated as clinical judgments. The system is intended only for shared-task research, not for diagnosis, treatment, or mental health assessment.

Misclassification and bias are also important concerns. The model may be affected by class imbalance and may over-predict majority or fallback labels such as High-Adaptive in ambiguous cases. In real-world use, such errors could lead to misleading interpretations of a speaker's psychological state. Language models may also reflect biases from pretraining data, especially when interpreting emotionally expressive or culturally diverse language.

To reduce these risks, we present the system strictly as a research prototype, report results transparently, and discuss its limitations. Any future use in mental health settings would require human oversight, stronger validation, and proper ethical review.

References

Edward Ajayi, Martha Kachweka, Mawuli Deku, and Emily Aiken. 2025. A machine learning approach for detection of mental health conditions and cyberbullying from social media. *arXiv preprint arXiv:2511.20001*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

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.

Robinson Narvaez Burbano, Oscar Mauricio Caicedo Rendon, and Carlos A. Astudillo. 2025. An encoder-only transformer model for depression detection from social network data: The deent approach. *Applied Sciences*, 15(6):3358.

J. Christopher Perry. 1990. *Defense Mechanism Rating Scales (DMRS), 5th Edition*. McGill University / DMRS Research Group.

Jinwen Tang, Qiming Guo, Yunxin Zhao, and Yi Shang. 2024. Decoding linguistic nuances in mental health text classification using expressive narrative stories. In *2024 IEEE 6th International Conference on Cognitive Machine Intelligence (CogMI)*, pages 207–216.

George E. Vaillant. 1992. *Ego Mechanisms of Defense: A Guide for Clinicians and Researchers*. American Psychiatric Press, Washington, DC.

Wei Zhai, Hongzhi Qi, Qing Zhao, Jianqiang Li, Ziqi Wang, Han Wang, Bing Xiang Yang, and Guanghui Fu. 2024. Chinese mentalbert: Domain-adaptive pretraining on social media for chinese mental health text analysis. *arXiv preprint arXiv:2402.09151*.