

# CS\_Metro at PsyDefDetect: Detecting Psychological Defense Mechanisms in Mental Health Dialogues with Summarization-Enhanced Transformer Ensembles

Oarisa Rebayet\*, Radiul Walee\*, Symom Hossain Shohan,  
Kawsar Ahmed and Mohammed Moshiul Hoque

Department of Computer Science and Engineering  
Chittagong University of Engineering and Technology

u2104129@student.cuet.ac.bd, moshiul\_240@cuet.ac.bd

## Abstract

Detecting psychological defense mechanisms in supportive conversations is essential for assisting mental health practitioners. Natural language processing techniques are increasingly integral to such systems, enabling automated classification of defense levels to better understand help-seeker behavior and resistance patterns. In PsyDefDetect at BioNLP 2026, we address the task of nine-class defense level classification on the PSYDEFCONV corpus. We propose a three-stage pipeline combining LLM-based dialogue summarization, domain-specific transformer fine-tuning, and rule-based ensemble prediction. Additionally, we evaluate three mental health domain-specific transformers (Mental-BERT, Mental-RoBERTa, Mental-XLNet) alongside fine-tuned LLMs (Qwen3-4B, Qwen3-1.7B, Mistral-7B) under different input conditions. Experimental results on the released test-set gold labels show that our ensemble approach achieves the best performance, reaching 34.69% macro F1 and surpassing the baseline by 4.69 percentage points. On the official PsyDefDetect Leaderboard 1 (labels 1–8), the submitted system achieved a Macro-F1 score of 23.46%, ranking 15th out of 21 teams, while on Leaderboard 2 (labels 0–8), it achieved 30.04%, securing 14th place. These findings demonstrate that domain-specific transformers substantially outperform generic LLM fine-tuning on this specialized clinical task.

## 1 Introduction

Psychological defense mechanisms are automatic cognitive and behavioral strategies that individuals employ to manage emotional distress and interpersonal conflict (Vaillant, 1992). While defenses serve adaptive functions, their rigid or excessive use is strongly associated with poor mental health outcomes, reduced help-seeking behavior, and resistance to therapeutic intervention (?). In clinical

and supportive conversations, the ability to automatically identify defense levels in a speaker’s utterances could assist mental health practitioners in monitoring client progress, tailoring interventions, and reducing annotation burden in large-scale clinical studies. Despite the growing application of natural language processing (NLP) to mental health (Chancellor and Choudhury, 2020), the automatic detection of psychological defense mechanisms remains largely underexplored. The task is inherently challenging, as psychological defense mechanisms are highly dependent on contextual information, and remain difficult to reliably assess even for trained clinicians.

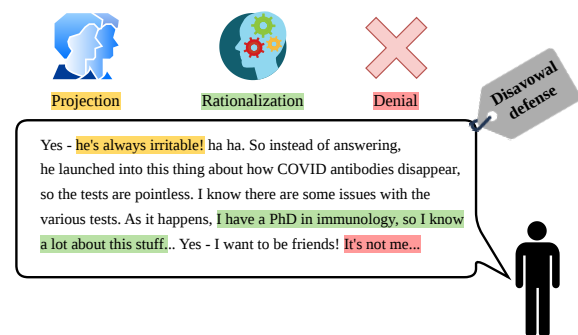


Figure 1: An example of defense level classification. Highlighted spans show the linguistic signals that indicate a Disavowal defense (Label 3) in the target utterance.

The recently introduced PSYDEFCONV corpus (Na et al., 2026b), comprising 200 therapeutic dialogues across nine defense levels, establishes a benchmark for this problem. However, even strong language models achieve only around 30% macro F1 in zero-shot and fine-tuning settings, and they consistently tend to overpredict mature defenses due to class imbalance. As illustrated in Figure 1, the task requires identifying such defense-indicative signals within helpseeker utterances to assign one of nine DMRS-defined defense levels.

To address these challenges, we participated in Psy-DefDetect@BioNLP 2026 (Na et al., 2026a), a shared task on detecting psychological defense levels in supportive conversations. Our contribution are as follows:

- We propose a dialogue summarization strategy using Qwen3-4B-Instruct (Team, 2025) to reduce conversational noise and provide global context for utterance-level defense classification.
- We systematically evaluate mental health domain-specific transformer models and fine-tuned LLMs, providing comparative analysis across all nine defense levels. Our results show that transformers with summarization guidance achieve more competitive performance than raw LLM fine-tuning on this specialized clinical task.
- We introduce a rule-based ensemble that leverages per-class model strengths identified through validation analysis, achieving more balanced predictions across defense levels.

## 2 Literature Review

Modeling mental health conversations poses unique challenges for NLP systems due to their context dependent nature, and has recently been explored extensively in the context of large language models for psychotherapy (Na et al., 2025). The ESConv corpus (Liu et al., 2021) laid the foundation for this area, providing 1,053 dialogues annotated with eight support strategies. Extended ESConv and ESConv-SRA (Madani and Srihari, 2025) generate strategy conditioned continuations and analyze how large language models (LLMs) maintain coherence across turns. Multi-Strategy ESConv (Bai et al., 2026) demonstrates that LLMs outperform supervised models in producing multi-strategy replies. Additionally, Zheng et al. (2023a) present AugESC, which finetunes GPT-J 6B on ESConv to complete dialogue threads, while Zheng et al. (2023b) create ExTES via ChatGPT in-context generation. Zhang et al. (2024) introduce ESD-CoT, extending ESConv with explicit reasoning chains.

Defense mechanisms are unconscious strategies individuals employ to manage emotional distress (Freud, 1936). The Defense Mechanism Rating Scales (DMRS) (Perry et al., 1993) is a validated, empirically grounded instrument that arranges defenses into a seven level hierarchy from Level 1

(Action defenses) to Level 7 (High Adaptive defenses) (Perry and Henry, 2004; Di Giuseppe and Perry, 2021). The PSYDEFCONV corpus (Na et al., 2026b), the first conversational dataset annotated with DMRS defense levels comprises 2,336 labeled utterances across nine defense classes. However, strong language models achieve only around 31% macro F1 in fine tuning settings, with a marked tendency to overpredict the dominant High Adaptive class (51.8% of the dataset) due to severe class imbalance.

Previous work on defense mechanism detection has either focused on supporting strategies in emotional support conversations or applied general mental health classification techniques without explicitly modeling defensive functioning. We address this gap through a three-stage pipeline: LLM-based dialogue summarization to reduce noise, domain specific transformer fine tuning with class imbalance handling, and a rule-based ensemble to counteract overprediction of the dominant class.

## 3 Dataset and Task Description

PsyDefDetect@BioNLP 2026 (Na et al., 2026b) is a shared task focused on analyzing psychological defense mechanisms in supportive conversations. The task requires identifying the defense level of a target utterance within its multi-turn dialogue context, spanning 9 categories described in Appendix 4. Table 1 summarizes the corpus statistics.

Corpus Statistic	Train	Test
Dialogues	200	189
Utterances	1,864	472
Average turns per dialogue	22.3	16.9
Average words per utterance	18.9	18.3
Seeker-to-Supporter ratio	1.04:1	1.06:1

Table 1: Dataset statistics for train and test splits.

## 4 System Overview

This study addresses psychological defense mechanism classification from therapeutic dialogues as a nine-class prediction problem. Our pipeline comprises three sequential stages: LLM-based dialogue summarization, domain-specific transformer fine-tuning, and a rule-based ensemble that combines model predictions. Figure 2 represents the architecture of the system. The implementation and source code are publicly available on GitHub.<sup>1</sup>

<sup>1</sup>[https://github.com/0arisaR/CS\\_Metro-at-PsyDefDetect-BioNLP-2026](https://github.com/0arisaR/CS_Metro-at-PsyDefDetect-BioNLP-2026)

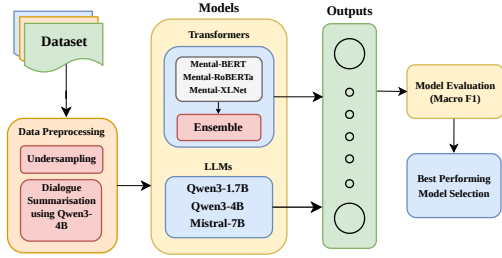


Figure 2: Overview of the proposed methodology pipeline, including dialogue summarization, transformer-based classification, LLM fine-tuning, and evaluation.

#### 4.1 Data Preprocessing

The dataset presents two primary challenges: irrelevant or noisy information that can misguide the model and a notable class imbalance. To mitigate these issues, we apply dialogue summarization to extract the most relevant content from each conversation and use class imbalance handling techniques to ensure more balanced learning across all classes.

**Dialogue Summarization:** Each dialogue was summarized using **Qwen3-4B-Instruct** to capture the main intent of the conversation. The dialogues were first formatted by concatenating speaker turns as "Speaker: text" and then passed to the model with instructions to produce a concise summary within 100 words. The structured prompt used for this summarization step is provided in Appendix 7.

**Class Imbalance Handling:** The training data exhibits a strong class imbalance, with Label 7 (High-Adaptive) originally comprising 968 samples, making it disproportionately dominant. To mitigate this effect, we apply targeted undersampling, restricting Label 7 to a maximum of 550 samples. However, since undersampling alone does not fully resolve the imbalance across the remaining classes, we further compute inverse-frequency class weights on the undersampled data.

$$w_c = \frac{N}{9 \times \text{count}(c)} \quad (1)$$

where  $N$  is the total number of training samples and  $\text{count}(c)$  is the number of samples in class  $c$ .

#### 4.2 Models

We evaluate two categories of models for multi-label classification:

**Domain-Specific Transformer Models** We employ transformer models which are pretrained on mental-health related corpora such as Mental-BERT-base and Mental-RoBERTa-base (Ji et al., 2022), and Mental-XLNet-base (Ji et al., 2023).

**Fine-tuned Large Language Models** We further fine-tune instruction-tuned LLMs, including Qwen3-1.7B, Qwen3-4B (Team, 2025), and Mistral-7B<sup>2</sup> (Jiang et al., 2023), to evaluate their effectiveness on the task.

#### 4.3 Training

The models are trained end-to-end using the cross-entropy loss function. All experiments follow a unified hyperparameter configuration, as summarized in Table 2.

Hyperparameter	Value
Epochs	20
Learning rate	$2 \times 10^{-5}$
Batch size	16
Weight decay	0.01

Table 2: Training hyperparameters.

#### 4.4 Rule-Based Ensemble

We construct a rule-based ensemble based on per-class performance observed on the validation set. The models exhibit complementary strengths across labels, with predictions selected from the model that is relatively more reliable for a given class. In cases where no clear preference emerges, the ensemble defaults to the most stable model. Final predictions are stored along with individual model outputs for evaluation.

### 5 Results and Discussion

Table 3 summarizes the performance of all models on the test set across labels (0–8). The results correspond to evaluation on the released test-set gold labels. In the official PsyDefDetect Leaderboard 1 (labels 1–8), our submission achieved a Macro-F1 score of 23.46%, ranking 15th out of 21 teams, while in Leaderboard 2 (labels 0–8), it achieved a Macro-F1 score of 30.04%, securing 14th place.

<sup>2</sup><https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>

<i>Transformers</i>					
Model	Approach	Acc	Rec	Prec	F1
MentalBERT	Raw	56.36	28.80	28.25	28.36
	Summ	52.97	33.32	31.53	31.21
	$\Delta$	-3.39	+4.52	+3.28	+2.85
MentalRoBERTa	Raw	57.84	33.21	31.79	32.20
	Summ	54.45	31.56	30.66	29.57
	$\Delta$	-3.39	-1.65	-1.13	-2.63
MentalXLNet	Raw	46.40	23.17	21.69	22.17
	Summ	51.06	33.38	27.94	29.65
	$\Delta$	+4.66	+10.21	+6.25	+7.48
Ensemble	Summ	<b>51.48</b>	<b>40.00</b>	<b>36.00</b>	<b>34.69</b>
<i>LLMs (Few-shot, 9-example)</i>					
Llama-3.2-3B-Instruct	Raw	30.72	22.15	18.64	14.65
	Summ	31.14	22.99	19.51	18.25
	$\Delta$	+0.42	+0.84	+0.87	+3.60
Qwen3-4B-Instruct	Raw	20.34	21.19	21.83	16.15
	Summ	29.03	21.72	24.95	19.48
	$\Delta$	+8.69	+0.53	+3.12	+3.33
Phi-4-mini-instruct	Raw	33.47	18.14	20.29	16.73
	Summ	34.96	23.60	22.27	19.46
	$\Delta$	+1.49	+5.46	+1.98	+2.73
<i>LLMs (Fine-tuned)</i>					
Qwen3-4B	Raw	63.77	31.31	36.73	30.04
Qwen3-1.7B	Raw	54.45	30.19	25.27	25.58
Mistral-7B	Raw	47.03	29.39	26.56	26.59

Table 3: Performance comparison of transformer-based models and LLMs (both fine-tuned and few-shot), reported in %. 'Summ' denotes summarisation of raw conversations; 'Raw' uses full dialogue history; few-shot uses 9 examples per label.

**Summarization reduces input noise but affects transformer models differently across architectures.** The results suggest that dialogue summarization has a non-uniform effect on transformer-based models. MentalBERT and MentalRoBERTa show a small decrease in accuracy (-3.39 for both), while recall and F1 increase, indicating that summarization helps recover relevant information but may slightly affect precision. In contrast, MentalXLNet improves across all metrics, with a notable gain in F1 (+7.48) and recall (+10.21), suggesting better use of compressed dialogue context.

**Model architecture influences sensitivity to compressed dialogue representations.** The results show that transformer variants respond differently to summarization. MentalRoBERTa performs best on raw dialogues but degrades more with summarized inputs, whereas MentalXLNet benefits more consistently from compression. MentalBERT remains comparatively stable but achieves lower over-

all performance. Overall, stronger contextual representations appear more effective at handling compressed inputs.

**Large language models leverage longer context windows but remain competitive rather than outperforming transformer ensembles.** Fine-tuned LLMs showed strong performance, with Qwen-4B (QLoRA) performing best, followed by Qwen-1.7B and Mistral-7B. Unlike encoder-based transformers, these models were trained on raw dialogues, as their larger context windows reduce the need for summarization and make them less sensitive to irrelevant conversational context. However, despite this advantage, LLMs did not surpass the transformer ensemble, suggesting that carefully optimized transformer models with summarization remain highly effective for this specialized classification task.

**Few-shot learning remains below fine-tuned baselines.** Although summarization improves performance, the best few-shot F1 score (19.48) remains well below the transformer ensemble (34.69) and fine-tuned Qwen3-4B (30.04). This gap suggests that limited in-context examples are insufficient for capturing the nuanced patterns in mental health dialogues, emphasizing the importance of task-specific fine-tuning.

## 5.1 Per-Class Behaviour Analysis

Figure 3 illustrates the class-wise performance distribution across all models. In particular, No Defense and High-Adaptive exhibit higher accuracy across all models, indicating that these classes dominate the decision space. In contrast, defense categories such as Disavowal and Neurotic show low performance.

The observed distribution suggests a strong bias toward high-frequency, while minority and psychologically nuanced categories remain unstable across architectures. This indicates that the classification boundary is primarily shaped by data imbalance rather than model capacity.

## 5.2 Error Analysis

Figure 4 presents the confusion matrix of the ensemble system, highlighting systematic error patterns across all classes. The misclassification structure is highly non-uniform and reveals strong semantic entanglement between several defense mechanism categories.

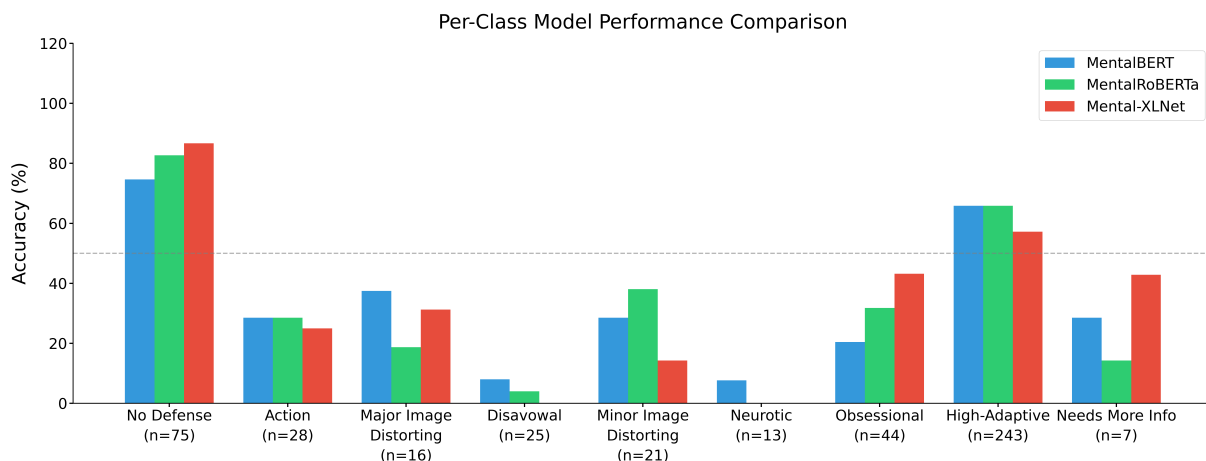


Figure 3: Per-class accuracy comparison across models.

True Class \ Predicted Class	NoDef	Action	MajDist	Disavow	MinDist	Neurotic	Obsess	HighAdapt	NeedInfo
NoDef	55	0	0	0	0	2	1	16	1
Action	0	11	4	0	3	1	5	4	0
MajDist	0	0	5	0	4	0	3	4	0
Disavow	1	5	3	2	8	0	1	5	0
MinDist	0	1	4	0	3	2	3	8	0
Neurotic	1	1	0	0	1	1	0	9	0
Obsess	1	4	1	1	2	2	15	18	0
HighAdapt	10	9	8	1	11	3	21	177	3
NeedInfo	0	1	0	0	0	0	1	4	1

Figure 4: Confusion matrix of ensemble predictions over Labels 0–8.

A clear error pattern is the consistent bias toward High-Adaptive, which behaves as a dominant sink class. Samples from Disavowal, Minor Image Distorting, Neurotic, and Obsessional classes are often misclassified as High-Adaptive, suggesting that uncertain cases are frequently mapped to a high-frequency category.

There is also notable confusion between Neurotic and High-Adaptive, indicating that the model does not clearly separate these two and may treat them as closely related states.

Minor Image Distorting is particularly unstable, with predictions spread across several nearby classes, showing weak and inconsistent learning for this category.

## 6 Conclusion

This study investigates psychological defense mechanism classification within conversations us-

ing transformer-based models and fine-tuned large language models. MentalBERT, MentalRoBERTa, and MentalXLNet were evaluated under raw and summarized settings alongside an ensemble approach. The results indicate that summarization consistently benefits MentalXLNet, whereas MentalBERT and MentalRoBERTa exhibit mixed behavior. Fine-tuned LLMs, including Qwen-4B, Qwen-1.7B, and Mistral-7B, were evaluated on raw dialogues leveraging their larger context windows. These models do not consistently outperform the transformer ensemble. Future research could explore synthetic augmentation for minority classes, dialogue compression strategies, and cross-lingual transfer to enhance generalization.

## Limitations

While this study contributes to psychological defense mechanism classification in supportive conversations, several limitations exist. Summarization may discard subtle emotional cues critical for understanding psychological distress. Resource constraints prevented exploration of larger LLMs, alternative compression techniques, multi-turn context, and cross lingual transfer. Models were evaluated on a single dataset (PSYDEFCONV), which may not generalize across diverse populations. Finally, parameter-efficient fine-tuning may not fully utilize the representational capacity of larger models compared to full fine-tuning.

## Ethical Considerations

The system presented in this study are intended to support, not substitute, clinical professionals. Predictions are subject to potential misclassifica-

tion and may inherit biases present in the training data. Real-world deployment necessitates human oversight and compliance with applicable data protection regulations.

## Acknowledgements

We express our sincere gratitude to the organizers of the PsyDefDetect@BioNLP 2026 shared task for conducting the shared task and for providing the dataset and evaluation framework. We also acknowledge the support of the Directorate of Research and Extension and the NLP Lab at Chittagong University of Engineering and Technology (CUET), Bangladesh, for their continuous encouragement and support.

## References

- Xin Bai, Guanyi Chen, Tingting He, Rui Fan, Chenlian Zhou, and Yu Liu. 2026. [Emotional supporters often use multiple strategies in a single turn](#). *Neurocomputing*, 693:133897.
- 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.
- 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*, 12:718440. PMID: 34366995.
- Sigmund Freud. 1936. [Inhibitions, symptoms and anxiety](#). *The Psychoanalytic Quarterly*, 5:1–28. Original work published 1936.
- 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 LREC*.
- Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, Erik Cambria, and Jörg Tiedemann. 2023. [Domain-specific continued pretraining of language models for capturing long context in mental health](#). *arXiv preprint arXiv:2304.10447*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.
- Navid Madani and Rohini Srihari. 2025. [Steering conversational large language models for long emotional support conversations](#). In *Proceedings of the Third Workshop on Social Influence in Conversations (SICoN 2025)*, pages 109–123, Vienna, Austria. Association for Computational Linguistics.
- Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. [A survey of large language models in psychotherapy: Current landscape and future directions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.
- 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, Marianne E. Kardos, and Christopher J. Pagano. 1993. [The study of defenses in psychotherapy using the defense mechanism rating scales \(dmrs\)](#). In *The Concept of Defense Mechanisms in Contemporary Psychology: Theoretical, Research, and Clinical Perspectives*, pages 122–132.
- John Perry and Melissa Henry. 2004. [Studying defense mechanisms in psychotherapy using the defense mechanism rating scales](#). *Advances in Psychology*, 136.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- George E. Vaillant. 1992. *Ego Mechanisms of Defense: A Guide for Clinicians and Researchers*. American Psychiatric Press. Original work published June 1, 1992.
- Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. [ESCoT: Towards interpretable emotional support dialogue systems](#). In *Proceedings*

of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. *AugESC: Dialogue augmentation with large language models for emotional support conversation*. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023b. *Building emotional support chatbots in the era of llms*. *ArXiv*, abs/2308.11584.

## A Appendix

### A.1 Defense Mechanism Labels and Dataset Distribution

Table 4 presents the definition of nine defense mechanism labels along with their class distribution across train and test splits.

Label	Level	Definition	Train	Test
0	No Defense (NoDef)	No defensive content.	296	75
1	Action (Action)	Distress acted out externally.	108	28
2	Major Image Distorting (MajDist)	All-good or all-bad distortions.	61	16
3	Disavowal (Disavow)	Denial, projection, fantasy.	99	25
4	Minor Image Distorting (MinDist)	Idealization or devaluation.	84	21
5	Neurotic (Neurotic)	Motives kept out of awareness.	48	13
6	Obsessional (Obsess)	Logic separates feeling from event.	172	44
7	High-Adaptive (HighAdapt)	Mature emotional coping.	968	243
8	More Info Needed (NeedInfo)	Too ambiguous to annotate.	28	7
<b>Total</b>			<b>1864</b>	<b>472</b>

Table 4: DMRS label taxonomy with train/test distribution (Perry et al., 1993).

### A.2 QLoRA Fine-Tuning Configuration

We employ parameter-efficient fine-tuning using QLoRA to adapt a 4-bit quantized Qwen3-4B-Instruct model for the classification task.

Category	Configuration
Quantization	4-bit NF4 with bfloat16 compute
LoRA Rank ( $r$ )	16
LoRA Scaling ( $\alpha$ )	32
Target Modules	Attention and MLP projections
Epochs	3
Learning Rate	$1 \times 10^{-4}$
Effective Batch Size	8
Max Sequence Length	512 tokens

Table 5: QLoRA hyperparameter configuration for fine-tuning Qwen3-4B-Instruct.

### A.3 Few-Shot Prompt Structure

All evaluated models used a standardized 9-shot prompting framework for defense mechanism clas-

sification. Each prompt included a fixed system instruction, nine in-context examples and a final target query. The complete prompt structure is shown below.

#### System Prompt

You are a clinical psychologist specializing in psychodynamic defense mechanisms. Classify the psychological defense mechanism in the given therapy exchange.

#### Defense levels (choose exactly one):

- 0 – No Defense: No clear defense mechanism observed.
- 1 – Action: Acting out, passive aggression, help-rejecting complaining.
- 2 – Major Image-Distorting: Splitting, projective identification, autistic fantasy.
- 3 – Disavowal: Denial, projection, rationalization.
- 4 – Minor Image-Distorting: Devaluation, idealization, omnipotence.
- 5 – Neurotic: Repression, reaction formation, displacement.
- 6 – Obsessional: Intellectualization, isolation of affect, undoing.
- 7 – High-Adaptive: Humor, sublimation, anticipation, altruism, affiliation.
- 8 – Needs More Info: Not enough context to classify.

Your response must end with exactly: **Label: (digit)**

Example: Label: 3

No explanation after the label.

### A.4 Few-Shot Examples

To construct the few-shot prompting framework, one representative demonstration example was selected for each defense mechanism category (labels 0–8). Each example comprises an LLM-generated session summary, a target utterance, and its corresponding gold-standard label. The demonstrations were formatted using a unified chat-based structure to maintain consistency across all evaluated models. The complete set of nine in-context examples is presented in Table 6.

Table 6: Nine-shot in-context examples. *Session summary* is an LLM-generated condensation of the full conversation history; *Target utterance* is the text the model must label.

L	Class	Session summary	Target utterance
0	No Defense	The dialogue begins with a simple greeting from the Seeker. Overall conclusion: no meaningful points, insights, or conclusions can be drawn due to lack of content beyond a greeting.	<i>"hi"</i>
1	Action	The seeker feels isolated and lethargic due to the cold season and lack of family or friends, living in a hotel during the pandemic. They struggle with motivation to work or write and have no social connections. The supporter suggests planning Christmas activities to break the cycle of loneliness, but the seeker acknowledges they have no support system currently. Key insight: isolation and lack of social connection significantly impact emotional well-being, highlighting the need for intentional connection despite current circumstances.	<i>"I live in a hotel currently. No family or friends around, especially with the virus."</i>
2	Major Image-Distort	The seeker expresses feelings of depression and self-doubt, struggling to feel like a good person and believing they are a failure. The supporter validates their pain, offers empathy, and affirms the seeker's potential to become anyone they desire. The key insight is that while the supporter is not a professional, they provide emotional support and encourage self-worth, suggesting that personal transformation is possible despite current feelings of failure.	<i>"I am not sure of that. I feel like a failure."</i>
3	Disavowal	The seeker experiences overwhelming anxiety and depression without a clear trigger, fueled by overthinking. The supporter emphasizes the importance of professional help and shares that engaging in enjoyable hobbies helps distract from racing thoughts. Both acknowledge that seasonal changes, like winter, can worsen symptoms. Key insight: consistent self-care, hobbies, and professional support are vital for managing anxiety and depression on bad days.	<i>"That makes a lot of sense. Winter makes it tougher too I think."</i>
4	Minor Image-Distort	The seeker feels isolated and lethargic due to the cold season, lack of family, and pandemic-related restrictions. Living in a hotel with no nearby support, they struggle with motivation and loneliness, despite being a writer. While they try dating sites, they find them impersonal. The supporter suggests planning a positive future event, like a Christmas gathering next year, to break the cycle of isolation and build hope.	<i>"I try the dating sites, but they seem too far away or impersonal. Just the current season has me down."</i>
5	Neurotic	The seeker expresses deep frustration and emotional distress over recurring negative events, particularly their mother's situation in Mexico with a deported partner and financial mismanagement. The supporter validates their feelings and asks about past efforts. Key insight: the seeker struggles with a sense of helplessness and cycles of worsening circumstances, highlighting emotional exhaustion and a lack of perceived progress in personal or familial situations.	<i>"Hm. I tried moving on. And it just appears she gets worse."</i>
6	Obsessional	The seeker expresses stress, anxiety, and financial struggles due to the pandemic and job market challenges. They feel overwhelmed and emotionally drained, seeking any work opportunity. The supporter empathizes, acknowledges shared difficulties, and asks about the seeker's preferred type of work. Key insight: widespread economic hardship is impacting mental and financial well-being, and emotional support and understanding are crucial. The conversation highlights the need for empathy and tailored support in times of crisis.	<i>"anything really atm and sorry to hear that—it's tough everywhere, a lot of people having a hard time def."</i>
7	High-Adaptive	The seeker struggles with chronic back pain, anxiety, and depression for eight years, compounded by grief over their mother's death and pandemic-related delays in medical care. The supporter emphasizes the importance of talking through emotions, journaling, and seeking support. The dialogue highlights resilience, the value of emotional support, and hope amid ongoing challenges.	<i>"Thank you, I will look there for others going through the same things, it doesn't sound silly at all and you are an amazing person too. Thank you for listening and your thoughtful help."</i>
8	Needs More Info	The seeker feels shamed by friends for not donating to charity during Christmas, due to financial struggles from COVID-19. The supporter emphasizes that donations should come from genuine capacity, not guilt. The seeker should contribute based on personal values and circumstances, not others' expectations—pride in self-worth is essential.	<i>"Yeah..."</i>

### A.5 Prompt Design for Dialogue Summarization

Table 7 presents the structured prompt used for LLM-based dialogue summarization. It defines a template in which a raw dialogue is provided as input and the model generates a concise summary capturing the key information. The prompt emphasizes clarity and semantic consistency, ensuring that the generated summary reflects the core content of the dialogue without introducing unnecessary details.

Instruction Prompt
Please summarize the following dialogue in a clear and concise manner, capturing the main points, key insights, and overall conclusions. Your summary should be within 100 words, focusing on the core ideas without unnecessary details. <b>Dialogue:</b> {dialogue} <b>Summary:</b>

Table 7: Structured prompt used for LLM-based dialogue summarization.

### A.6 Example of Dialogue Summarization

To illustrate the effectiveness of the proposed prompt, an example of input dialogue and the corresponding generated summary is presented in Table 8.

Example: Input–Output Pair
<b>Dialogue:</b> <ul style="list-style-type: none"><li>• Supporter: hi</li><li>• Seeker: Hey You there?</li><li>• Supporter: Hi I stepped out, how are you?</li><li>• Seeker: I'm okay I guess. How are you?</li><li>• Supporter: Good, I am good. Is there anything I can be of an assistance today?</li><li>• Seeker: I don't know. I've been dealing with a lot this last month.</li></ul> <b>Generated Summary:</b> <p>The dialogue begins with a friendly greeting and mutual check-in. The supporter expresses well-being and offers assistance, while the seeker reports experiencing significant challenges over the past month. This indicates emotional strain and a need for support. Overall, the interaction reflects a supportive conversational exchange.</p>

Table 8: Example of input dialogue and corresponding generated summary using the proposed prompt.

## A.7 Prediction Examples

Table 9 illustrates sample predictions for the dialogue classification task. In this table, sample text inputs are presented along with their corresponding summaries, actual labels, and predicted labels from the ensemble model. This provides a qualitative view of how the model performs across different input examples by comparing the true labels with the predicted outputs.

Summary	Current Text	Actual Defense	Predicted Defense
The seeker reports chronic severe back pain causing significant anxiety and depression. The supporter offers assistance, indicating a need for emotional and possibly medical support. Key insight: persistent physical pain is deeply impacting mental health, highlighting the interconnection between physical suffering and emotional well-being. Conclusion: Addressing both the back pain and its psychological effects is essential for holistic recovery.	Hi I'm experiencing a lot of anxiety and depression because of constant severe back pain.	High-Adaptive	High-Adaptive
The seeker feels depressed, attributing it to pandemic-related isolation and lack of connection with loved ones. Despite fears of vulnerability, the supporter encourages open communication, noting that sharing emotions can relieve sadness and strengthen relationships. The key insight is that vulnerability is healthy and worthy—true love accepts one as they are. The seeker resolves to be brave and express their feelings to their boyfriend.	That is true. I should be brave enough to find out if he will be there when the chips are down. I am going to do it.	High-Adaptive	Obsessional
The seeker expresses feelings of depression and lack of motivation, attributing it to being stuck in one place. The supporter asks what caused the depression, indicating a need to explore underlying causes. The core insight is that the seeker feels emotionally stagnant and unmotivated, possibly due to a lack of change or progress in their current situation. The key takeaway is that identifying the root cause of stagnation is essential for addressing the depression.	I don't know. Maybe stuck at one place	Neurotic	Neurotic

Table 9: Sample predictions with actual and predicted defense levels.