# transformer_1376 at PsyDefDetect: A QLoRA-Based Generative Framework for Context-Aware Psychological Defense Mechanism Detection

**Pritha Saha[1], Shuvodwip Saha[1], Anik Mahmud Shanto[2,*]**
[1] Chittagong University of Engineering & Technology, Chittagong-4349, Bangladesh
[2] Southeast University, Dhaka, Bangladesh
{priitha.saha, shuvodwipsaha}@gmail.com,
anik.mahmudshanto@seu.edu.bd
Correspondence: anik.mahmudshanto@seu.edu.bd

## Abstract

Psychological defense mechanisms play a crucial role in shaping human responses during emotionally charged conversations, yet remain underexplored in natural language processing. In this work, we address the PSYDEFCONV shared task, which involves classifying defense mechanisms in multi-turn dialogues using clinically grounded annotations based on the Defense Mechanism Rating Scales (DMRS). We propose a generative supervised fine-tuning framework that reformulates the task as conditional text generation. A pre-trained causal language model (Gemma-2-2B) is adapted using parameter-efficient fine-tuning (PEFT) with 4-bit quantization, enabling efficient training under limited computational resources. To handle class imbalance, we apply random oversampling, and we design a prompt-based input representation to incorporate conversational context effectively. Experimental results demonstrate that our generative approach is competitive with discriminative baselines while offering improved flexibility in modeling subtle and context-dependent defensive behaviors. The findings highlight the potential of generative large language models for psychologically grounded dialogue understanding tasks.

## 1 Introduction

Psychological defense mechanisms are vital adaptive or maladaptive strategies that shape how individuals regulate emotions and express distress. In emotional support dialogues, identifying these mechanisms—such as denial or intellectualization—is essential for understanding a speaker's mental state and providing appropriate support. However, defensive functioning remains largely underexplored in natural language processing (NLP) compared to surface-level tasks like sentiment analysis (Na et al., 2025).

The PsyDefDetect@BioNLP 2026 shared task (Na et al., 2026a) addresses this gap by requiring the classification of target utterances into nine categories based on the Defense Mechanism Rating Scales (DMRS). This task is particularly challenging due to the subtle, context-dependent nature of defenses, the nuanced hierarchical label space, and the presence of ambiguous cases where context is insufficient for classification.

To address these challenges, we propose a generative framework that reformulates defense mechanism identification as a conditional text generation problem rather than standard multi-class classification. We utilize parameter-efficient fine-tuning (PEFT) on a pre-trained causal language model to adapt to the task with minimal computational overhead. Our approach leverages the reasoning capabilities of large language models to capture the long-range dependencies and implicit cues necessary for identifying nuanced psychological constructs.

**Contributions.** Our main contributions are as follows:

- We introduce a parameter-efficient adaptation strategy that enables effective training under limited computational resources.
- We demonstrate the effectiveness of our approach on the PsyDefDetect shared task, highlighting its ability to capture context-dependent and nuanced defensive behaviors.

## 2 Related Work

**Psychological Analysis in NLP.** Research in computational psychology has traditionally focused on emotion recognition and sentiment analysis (Poria et al., 2017; Zhang et al., 2018). While datasets like ESConv (Liu et al., 2021) facilitate emotional support modeling, they often focus on surface-level affect. The PSYDEFCONV dataset addresses this by providing annotations based on the Defense Mechanism Rating Scales (DMRS)



*Proceedings of the BioNLP 2026 (Shared Tasks)*, pages 99–103
July 3-4, 2026 ©2026 Association for Computational Linguistics

(Perry, 1990), enabling a deeper analysis of cognitive processes in dialogue.

**Context-Aware Dialogue Understanding.** Effective dialogue modeling requires capturing dependencies across multiple turns. While early methods used recurrent architectures (Serban et al., 2016), modern transformer-based approaches leverage self-attention to encode context (Devlin et al., 2019; Wolf et al., 2019). However, identifying implicit psychological phenomena remains a challenge, as these constructs depend on subtle nuances often missed by standard encoding strategies.

**Generative Approaches for Classification.** While discriminative models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are standard for classification, recent work explores reformulating these tasks as text generation (Raffel et al., 2020; Brown et al., 2020). Conditional generation via instruction-based prompting has shown superiority in complex scenarios requiring reasoning and contextual interpretation, making it highly suitable for defense mechanism detection.

**Parameter-Efficient Fine-Tuning.** Fine-tuning large language models (LLMs) is computationally demanding. Parameter-efficient fine-tuning (PEFT) techniques, such as LoRA (Hu et al., 2021), allow for effective model adaptation by updating only a fraction of the total parameters. These methods significantly reduce resource requirements while maintaining performance levels comparable to full fine-tuning, providing a scalable solution for specialized NLP tasks.

## 3 Dataset

We conduct our experiments on the PSYDEFCONV dataset (Na et al., 2026b), released as part of the shared task on psychological defense understanding in conversations. The dataset is constructed from a subset of the ESConv corpus (Liu et al., 2021) and contains multi-turn emotional support dialogues annotated using the Defense Mechanism Rating Scales (DMRS) (Perry, 1990).

| Statistic | Value |
|---|---|
| Total Dialogues | 200 |
| Total Utterances | 4,709 |
| Annotated Utterances | 2,336 |
| Number of Classes | 9 |

Table 1: Overview of the PSYDEFCONV dataset.

The task is formulated as a 9-class classification problem, where each instance consists of a target utterance and its preceding dialogue context. The labels correspond to hierarchical levels of defensive functioning, ranging from no defense to high-adaptive defenses, along with an additional category for ambiguous cases requiring more context.

Following the shared task setup, we have used the provided train, validation, and test split. The task is particularly challenging due to its context dependency, subtle inter-class variations, and the presence of ambiguous instances.

## 4 Methodology

We propose a generative supervised fine-tuning (SFT) framework as shown in figure 1 to model psychological defense classification as a conditional text generation task. A pre-trained causal language model (Gemma-2-2B) is adapted using parameter-efficient fine-tuning (PEFT) to predict defense labels given conversational context.

### 4.1 Data Processing

To preserve class distribution, we perform a stratified 80/20 train-validation split. Due to class imbalance, we apply random oversampling to minority classes up to a fixed target size, while retaining all samples from majority classes. This strategy improves representation of rare defense categories without discarding useful data.

### 4.2 Model and Fine-Tuning

We adopt a 4-bit quantized QLoRA setup for efficient training. Low-rank adapters are injected into all linear projection layers, enabling task adaptation with minimal parameter updates. Quantization significantly reduces memory usage, allowing fine-tuning on limited hardware while maintaining performance.

### 4.3 Generative Formulation

Unlike conventional classifiers, we formulate the task as next-token prediction. Each input is constructed using a chat-style template containing the system prompt, dialogue context, and target utterance. The model is trained to generate a single token corresponding to the defense label.

To ensure proper supervision, tokens corresponding to the input prompt are masked, and loss is computed only on the generated label token. For long sequences, a left-truncation strategy is applied

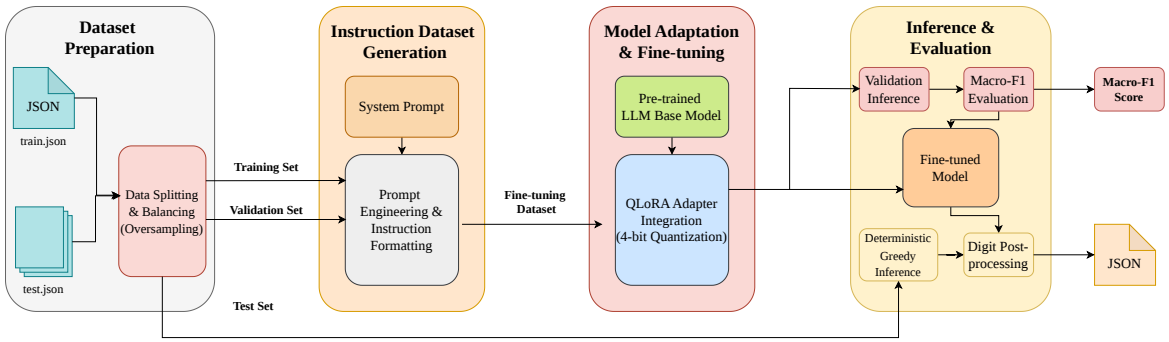


Figure 1: Our proposed framework for the PsyDefDetect task

to preserve the target utterance and label within the maximum sequence length. Since Gemma-2's chat template does not support a dedicated system role, the system instruction is prepended directly into the user turn. The full prompt template is provided in Appendix B.

### 4.4 Training Strategy

We implement a manual training loop using PyTorch with gradient accumulation to achieve an effective batch size of 16. Mixed precision training is employed to improve efficiency. A linear warmup followed by cosine decay is used for stable optimization.

### 4.5 Inference and Evaluation

During inference, we use greedy decoding to generate the predicted label deterministically. A simple post-processing step extracts the label token from the generated output. Following the shared task protocol, we report Macro-F1 score over classes 1–8, excluding the majority "No Defense" class to ensure balanced evaluation.

## 5 Experimental Analysis

### 5.1 Experimental Setup

All experiments are conducted on cloud-based GPU platforms (Kaggle and Google Colab) using NVIDIA T4 GPUs. The hardware constraints motivate our use of 4-bit quantization and parameter-efficient fine-tuning throughout.

We fine-tune Gemma-2-2B-IT via 4-bit QLoRA using the Unsloth framework. LoRA adapters are injected into all attention and feed-forward projection layers. Training uses a manual PyTorch loop with AdamW optimization, cosine decay scheduling with linear warmup, and gradient clipping. To address class imbalance, minority classes are oversampled to a fixed target count prior to training.

To isolate the contribution of the generative reformulation, we additionally train DeBERTa-v3-base as a discriminative sequence classifier under identical data conditions. The dialogue context and target utterance are encoded as a sentence pair with left-truncation applied to the context side. This model serves as an internal comparison point and is distinct from the organizer-provided official baseline.

Both systems are trained and evaluated on the same stratified 80/20 split of the official training data, with final evaluation performed on the official blind test set. Full hyperparameters for both systems are listed in Appendix A.

### 5.2 Results

Our system achieves an official Macro-F1 of 0.2475 on the blind test set (LB1), ranking 12th out of 21 registered teams. Table 2 presents a full comparison across all evaluation metrics.

Table 2: Comparison of model performance metrics(Accuracy, Precision, Recall and F1-score) on the official test set.

| Model | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|
| Official Baseline* | 0.6483 | 0.3397 | 0.3045 | 0.3148 |
| DeBERTa-v3-base | 0.5500 | 0.1202 | 0.1636 | 0.1376 |
| **Proposed Framework** | **0.5508** | **0.2669** | **0.2351** | **0.2475** |

*Organizer-provided baseline from official leaderboard.

The proposed generative system substantially outperforms the discriminative DeBERTa-v3-base encoder across all metrics, with a Macro-F1 gain of +0.11, suggesting that reformulating defense classification as conditional text generation better captures the implicit and context-dependent nature of psychological defenses. The gap relative to the official baseline reflects the inherent difficulty of

the task under limited computational resources, and points toward larger generative models as a promising direction for future work.

## 6 Conclusion

In this work, we presented a generative supervised fine-tuning framework for psychological defense mechanism classification in conversational data. By reformulating the task as conditional text generation and leveraging parameter-efficient fine-tuning with 4-bit quantization, we successfully adapted a causal language model to capture context-dependent psychological cues.

Experimental results demonstrate that the proposed method is competitive with standard discriminative baselines while offering improved flexibility in modeling implicit and nuanced defense behaviors. Future work will explore larger-scale models, improved context modeling strategies, and more robust decoding mechanisms to further enhance performance on psychologically grounded conversational tasks.

## 7 Limitations

Our approach has several limitations. First, the model relies on generative decoding, which introduces slight variability in outputs and may require strict post-processing to ensure valid label prediction. Second, the use of a relatively small causal language model (Gemma-2-2B) limits the capacity to capture highly complex conversational nuances compared to larger models. Third, class imbalance handling via oversampling may introduce redundancy and potential overfitting to minority classes. Finally, the reliance on fixed prompt templates may reduce generalization to unseen dialogue structures.

## References


Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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.

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.

J. Christopher Perry. 1990. *Defense Mechanism Rating Scales (DMRS)*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the*

*55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. *Preprint*, arXiv:1507.04808.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *Preprint*, arXiv:1901.08149.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis : A survey. *Preprint*, arXiv:1801.07883.

## A Hyperparameters

Table 3 details the hyperparameters for reproducibility, including the LoRA adapters and the 4-bit quantization setup.

Table 3: Detailed training and model configurations for reproducibility.

| **Hyperparameter** | **DeBERTa-v3-base** | **Gemma-2-2B + QLoRA** |
|---|---|---|
| Learning rate | 2e-5 | 1e-4 |
| Effective batch size | 8 | 16 |
| Epochs | 5 | 3 |
| Weight decay | 0.01 | 0.01 |
| Max sequence length | 512 | 1024 |
| Quantization | — | 4-bit NF4 |
| LoRA rank | — | 16 |
| LoRA alpha | — | 32 |
| LoRA dropout | — | 0.05 |
| Warmup ratio | — | 0.05 |
| Grad clip norm | — | 1.0 |
| Oversampling target | — | 400/class |
| Precision | fp32 | bf16/fp16 |
| Random seed | 42 | 42 |

## B Prompt Template

The following template is used to construct the input for both training and inference. During training, the assistant turn contains the gold label digit; during inference, the generation prompt cue is appended and the assistant turn is left empty for the model to complete.

```
[User]:
You are a clinical psychology expert specializing in defense mechanism analysis.
Classify the psychological defense level of the FINAL seeker utterance based on the conversation history provided.

Defense Level Reference:
 0 = No Defenses
 1 = Action Defenses (Passive Aggression, Help-Rejecting Complaining, Acting Out)
 2 = Major Image-Distorting (Splitting, Projective Identification)
 3 = Disavowal Defenses (Denial, Rationalization, Projection, Autistic Fantasy)
 4 = Minor Image-Distorting (Devaluation, Idealization, Omnipotence)
 5 = Neurotic Defenses (Repression, Dissociation, Reaction Formation, Displacement)
 6 = Obsessional Defenses (Isolation of Affect, Intellectualization, Undoing)
 7 = High-Adaptive Defenses (Affiliation, Altruism, Anticipation, Humor, Self-Assertion, Sublimation, Suppression)
 8 = Needs More Information

Rules:
- Read the full conversation for context.
- Focus on the TARGET utterance's defensive style, not the supporter's.
- Respond with ONLY a single digit (0-8). No explanation, no punctuation.

— Conversation History —
[Speaker]: [utterance]
[Speaker]: [utterance]
...

— Target Utterance [Seeker] —
[target utterance text]

What is the defense level? Respond with ONLY a single digit (0-8):

[Assistant]:
{label digit}
```