# VISHC AT PSYDEFDETECT: Mitigating Data Scarcity in Psychological Defense Classification with Context-Aware Synthetic Augmentation

**Hoang-Thuy-Duong Vu[1,2], Quoc-Cuong Pham[1,2], Huy-Hieu Pham[1,2,3]***

[1]College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam
[2]VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam
[3]Center for Innovations in Health Sciences, VinUniversity, Hanoi, Vietnam
{26duong.vht, 24cuong.pq, hieu.ph}@vinuni.edu.vn

## Abstract

Psychological defense mechanisms (PDMs) are unconscious cognitive processes that modulate how individuals perceive and respond to emotional distress. Automatically classifying PDMs from text is clinically valuable but severely hindered by data scarcity and class imbalance, challenges which generative augmentation alone cannot resolve without psychological grounding. In this work, we address these challenges in the PsyDefDetect shared task (BioNLP@ACL 2026) by proposing a context-aware synthetic augmentation framework combined with a hybrid classification model. Our hybrid model integrates contextual language representations with basic clinical features, along with 150 annotated defense items. Experiments demonstrate that definition quality in prompting directly governs generation fidelity and downstream performance. Our method surpasses DMRS CO-PILOT, reaching an accuracy of 58.26% (+40.25%) and a macro-F1 of 24.62% (+15.99%), thereby establishing a strong baseline for psychologically grounded defense mechanism classification in low-resource settings. Source code is available at: https://github.com/htdgv/CASA-PDC.

## 1 Introduction

Psychological Defense Mechanisms (PDMs) present a unique challenge for Natural Language Processing (NLP) field, particularly, they are unconscious, context dependent processes that appear through subtle cues such as narrative inconsistency, shifts in emotional framing, and distorted attribution, rather than clear lexical markers (Vaillant, 1994; Cramer, 1987; Bond et al., 1983). This implicit nature creates semantic ambiguity in which identical surface text may reflect distinct defensive processes, depending on underlying intent and psychological context, leading standard token or sentence level models to conflate adaptive coping with maladaptive defenses.

Data scarcity and class imbalance further compound these difficulties. Synthetic augmentation via Large Language Models (LLMs) offers a natural remedy, yet without psychologically grounded constraints, generative models produce fluent but theoretically invalid text, creating hallucinating defenses that introduce noise and erode model reliability (Ji et al., 2023; Na et al., 2025; Anaby-Tavor et al., 2020; Kumar et al., 2020). A key point is that the PsyDefDetect shared task (Na et al., 2026a) on the PSYDEFCONV dataset (Na et al., 2026b), based on the ESCONV dataset (Liu et al., 2021), introduces two auxiliary labels, *No Defense* (Level 0) and *Need More Information* (Level 8), that carry no corresponding clinical defense items (Di Giuseppe and Perry, 2021). These labels violate standard multi-class assumptions and produce skew distributions, making defense-item-based feature extraction underspecified, demanding a principled reformulation of the task.

We address these challenges in the PsyDefDetect shared-task through context-aware synthetic augmentation paired with a Hybrid Feature Fusion architecture. Specifically, our contributions are:

- **Psychologically grounded augmentation.** Stressor-anchored, theory-driven prompts with class-specific definitions from the Defense Mechanisms Rating Scales (DMRS) for synthetic augmentation to ensure generating high-fidelity examples, demonstrating that definition quality in prompting governs downstream performance.

- **Clinical feature engineering.** Structured features from all 150 defense items, along with basic clinical features, are fused with contextual language representations, bridging clinical theory and neural classification.

*Corresponding author: hieu.ph@vinuni.edu.vn



*Proceedings of the BioNLP 2026 (Shared Tasks)*, pages 77–86

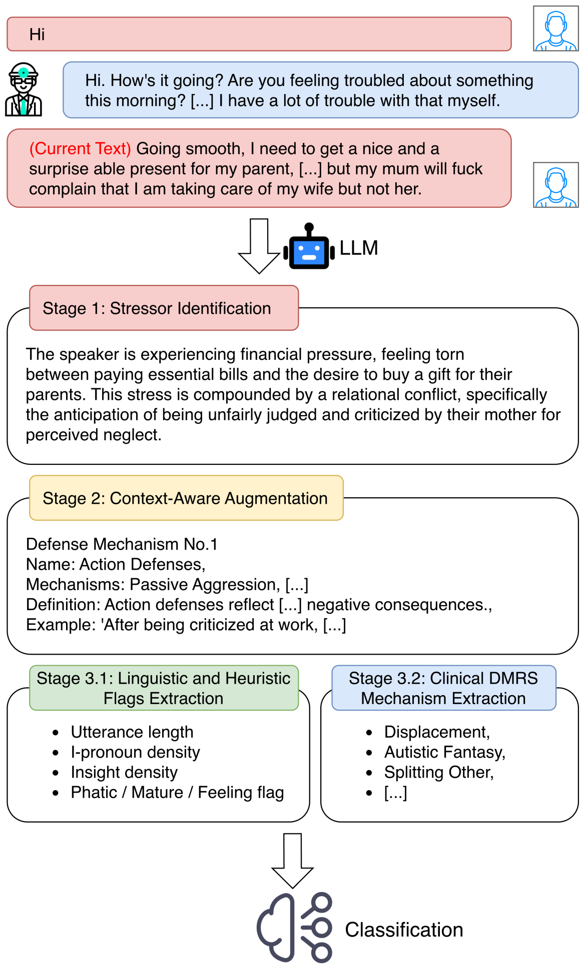



Figure 1: Overview of the multi-stage research pipeline. The process begins with (1) LLM-based stressor identification to establish contextual grounding; (2) context-aware synthetic data augmentation to address class imbalance; and (3) a dual-domain feature extraction stage targeting linguistic heuristics and clinical DMRS mechanism indicators; followed by the final classification.

- **Strong low-resource baseline.** Using Llama-3-8B-Instruct as data generator, our system improves accuracy (18.01% to 58.26%) and macro-F1 (8.63% to 24.62%) on the PsyDefConv blind-test set, establishing a competitive foundation for PDM classification.

## 2 Methodology

We address PDM classification under data scarcity through three stages: (1) a context aware synthetic data augmentation pipeline grounded in clinical theory; (2) a dual-domain feature extraction stage, combining linguistic heuristics with DMRS derived defense profiles; and (3) a hybrid fusion architecture that integrates contextual language representations with structured clinical features (Figure 1).

| Class | Label | N | SB↓ | SA↑ |
|---|---|---|---|---|
| No Defense | 0 | 500 | 0.399 | 0.481 |
| Action | 1 | 500 | 0.496 | 0.585 |
| Major Image | 2 | 488 | 0.440 | 0.559 |
| Disavowal | 3 | 500 | 0.453 | 0.597 |
| Minor Image | 4 | 500 | 0.416 | 0.592 |
| Neurotic | 5 | 384 | 0.392 | 0.525 |
| Obsessional | 6 | 500 | 0.429 | **0.619** |
| Needs Info | 8 | 224 | **0.413** | 0.601 |
| *Avg.* | | | **0.430** | **0.570** |

Table 1: Synthetic data quality per class. **SB**: Self-BLEU (Zhu et al., 2018) (lower = greater lexical diversity); **SA**: Semantic Adherence via Natural Language Inference (NLI) entailment (higher = stronger label alignment). Class 7 (High-Adaptive) is excluded from augmentation due to sufficient original samples; all other classes are capped at $N$=500 synthetic instances.

### 2.1 Context-Aware Data Augmentation

Standard augmentation methods such as paraphrasing and back translation (Wei and Zou, 2019) introduce diversity in phrasing but do not retain the functional role of defense mechanisms. We present a Synthetic Data Augmentation (SDA) pipeline based on Llama-3-8B-Instruct that captures the psychological conditions that give rise to defenses, with emphasis on function instead of form.

**Stressor-Anchored Generation.** Defense mechanisms emerge in response to perceived stressors rather than in isolation. Each prompt is anchored in a key stressor identified from the dialogue, such as interpersonal conflict, job loss, or social rejection. This approach promotes responses that reflect realistic defensive behavior instead of generic expressions of emotion.

**Theory-Driven Prompting.** To control semantic drift and reduce label inconsistency, each prompt defines the target defense level using structured clinical details from the DMRS framework (Di Giuseppe and Perry, 2021). These details include the defense name, its formal definition, and common linguistic and behavioral patterns. Paired with few-shot examples (Appendix A.3), this design steers generation toward samples that express the theoretical role of each defense rather than relying on surface plausibility (Brown et al., 2020).

### 2.2 Data Quality Control

Uncontrolled generation may introduce label noise and artifacts. To mitigate this, we use two quality control steps. First, a soft balancing scheme restricts each minority class to 500 total samples (real plus synthetic), which helps reduce overfitting to generation specific patterns. We examine five augmentation settings: $\times 1$, $\times 2$, $\times 5$, $\times 8$, and $\times 10$, together with the 500 cap variant. Second, a machine as annotator filter applies a secondary classifier to assign labels to generated batches; Only batches achieving a Cohen's Kappa of $\kappa \geq 0.60$ (Cohen, 1960), reflecting substantial agreement, are retained (Table 1).

### 2.3 Feature Extraction

Each seeker utterance is represented by two complementary feature sets: (i) lightweight linguistic heuristics capturing surface-level cues, and (ii) clinically grounded DMRS-derived features encoding latent defensive functioning.

**Linguistic and Heuristic Features.** We define six lightweight features to distinguish non-defensive (Label 0) from defensive responses, which are often conflated: *Utterance Length* (narrative elaboration proxy), *I-Pronoun Density* (self-focus), *Insight Density* (reflective reasoning), *Phatic Flag* (short filler utterances), *Mature Coping Flag* (triggered by length $> 12$, high insight, and elevated I-pronouns), and *Emotion Intensity* (model confidence in non-neutral predictions).

**DMRS Defense Profile.** We approximate latent defensive functioning using a four-step indicator inference procedure:

1. *Indicator Scoring:* An NLI model estimates entailment probability $P(T \Rightarrow I_j)$ for each of 150 DMRS indicators given utterance $T$.

2. *Mechanism Aggregation:* Indicator scores are grouped into 30 defense mechanisms and normalized to form mechanism scores $S(M_k)$.

3. *Profile Construction:* The resulting 30-dimensional vector defines the Defense Profile of the utterance.

4. *Level Mapping:* Mechanism scores are aggregated by DMRS level to obtain the predicted defense level: $\hat{y} = \arg\max_\ell \sum_{M_k \in \ell} S(M_k)$.

### 2.4 Hybrid Feature Fusion Architecture

Our system integrates contextual language representations with structured clinical features using a late fusion approach, as described below:

1. **Textual Encoder:** MentalRoBERTa (Ji et al., 2022) encodes each instance formatted as `[Stressor:S|Turn:T]`, conditioning the representation on both the triggering context and the response, yielding a 768-dim embedding.

2. **Feature Encoders:** The heuristic (7-dim) and DMRS-derived (30-dim) features are each passed through a dedicated Multilayer Perceptron (MLP) with the following structure: 64 $\rightarrow$ Batch Normalization $\rightarrow$ ReLU $\rightarrow$ Dropout (p = 0.3) $\rightarrow$ 32, producing two 32-dim vectors.

3. **Fusion and Classification:** The three representations are concatenated into an 832-dim vector (768 + 32 + 32) and passed through two fully connected layers (256 $\rightarrow$ 128, ReLU + Dropout ($p$=0.4)) and a final linear layer producing a probability distribution over 9 labels (Kiela et al., 2020).

## 3 Experiment & Results

### 3.1 Experimental Setup

**Data Pre-processing.** The training corpus combines human-annotated dialogues with synthetic samples generated by Llama-3-8B-Instruct under theory-driven prompting. We evaluate six augmentation scales, $\times 1$, $\times 2$, $\times 5$, $\times 8$, $\times 10$, and a hard cap of $N$=500 per class, yielding corpora ranging from approximately 1,800 to 5,100 instances. Class 7 is excluded from augmentation given its already substantial representation. The baseline corresponds to the organizer-provided code rerun with Llama-3-8B-Instruct and no augmentation. [1]

**Implementation Details.** The model is implemented in PyTorch using the Hugging Face Transformers library. MentalRoBERTa (`mental-roberta-base`) (Ji et al., 2022) serves as the textual encoder; its parameters are fine-tuned end-to-end with a learning rate of $1{\times}10^{-6}$. Task-specific layers (MLPs, fusion head, and classifier) use a higher learning rate of $1{\times}10^{-4}$, optimized with AdamW (Loshchilov and Hutter, 2019). Training runs for up to 20 epochs with batch size 16, early stopping on validation macro-F1, weight decay of $1{\times}10^{-2}$, and label smoothing ($\epsilon = 0.1$) to mitigate noise introduced by synthetic samples.

**Evaluation Metrics.** We follow the same protocol employed in Na et al. (2026a) and report all results on both the PsyDefConv development and blind test sets using macro-averaged Precision,

[1] Per-class metrics and DMRS activation patterns are detailed in Appendix A.

| Setting | Acc↑ | P↑ | R↑ | F1↑ |
|---|---|---|---|---|
| *Baseline* | | | | |
| DMRS CO-PILOT* | 0.1801 | 0.1904 | 0.1715 | 0.0863 |
| *Our System* | | | | |
| $\times 1$ | 0.5508 | 0.2555 | 0.2601 | 0.2543 |
| $\times 2$ | 0.5508 | 0.2789 | 0.2882 | **0.2799** |
| $\times 5$ | 0.5487 | 0.2764 | 0.2821 | 0.2783 |
| $\times 8^{\dagger}$ | **0.5826** | **0.2588** | **0.2503** | **0.2462** |
| $\times 10$ | 0.5254 | 0.2237 | 0.2289 | 0.2238 |
| $N$=500 | 0.5275 | 0.2659 | 0.2654 | 0.2628 |

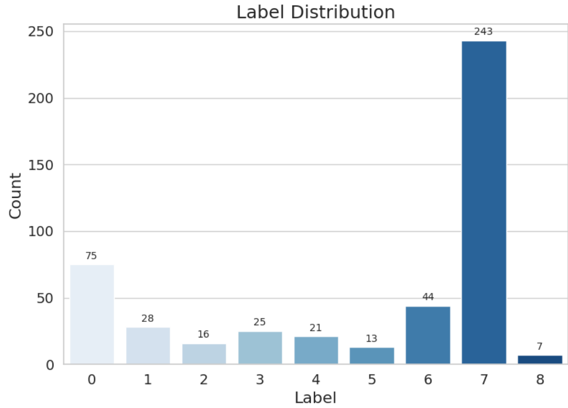
Table 2: Classification performance on the PSYDEFCONV official test set. *Baseline rerun with Llama-3-8B-Instruct on the released test set; original results in Na et al. 2026b use a different backbone. $\times k$: each minority class expanded to $k$ times its original size. $N$=500: hard cap of 500 instances per class. Metrics are macro-averaged. †Official leaderboard submission; all other rows are post-hoc evaluations on the released test set. Best results per column in **bold**.

Recall, and F1, as well as overall Accuracy, to account for class imbalance.

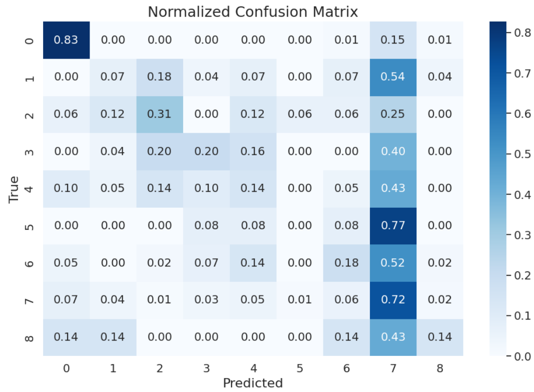
### 3.2 Results Analysis

**Classification performance across settings.** Our submission ranked 13 out of 21 registered teams in the official evaluation. Table 2 reports results across all six augmentation configurations on the official blind-test set. Every augmented variant substantially outperforms DMRS CO-PILOT in accuracy (+40.25 pp) and macro-F1 (+15.99 pp), confirming that theory-driven augmentation delivers robust gains over a prompt-only LLM baseline in this low-resource setting. Performance improves at lower augmentation scales but deteriorates as augmentation becomes more aggressive. The $\times 2$ configuration yields the highest macro-F1 (27.99%), indicating an effective balance between expanded class coverage and synthetic generation noise. Further scaling leads to a steady decline in macro-F1, which falls to 22.38% at $\times 10$, consistent with noise accumulation in heavily augmented corpora (Kumar et al., 2020). While $\times 8$ records the highest accuracy (58.26%), its macro-F1 remains 3.37 pp below $\times 2$, revealing that overall accuracy is disproportionately influenced by dominant Label 7 predictions at the expense of minority-class recall.

**The Label 7 sink effect and class imbalance.** The confusion matrix (Figure 2b) confirms Label 7 as a universal prediction sink. The most severe case is Label 5 (Neurotic), where 77% of instances are misclassified as Label 7, rendering this class unlearnable. Per-class F1 scores shows that Labels 0 and 7 exceed F1 $> 0.70$, while all remaining classes fall below 0.30, with four classes below 0.15. This implies that the accuracy (0.55-0.58) substantially overstates practical utility. The Label 5/7 confusion is semantically meaningful: both involve reflective discourse, but differ in whether anxiety is intellectualized or channeled.




(a) Label distribution of the PSYDEFCONV official test set.




(b) Row-normalized confusion matrix of our official leaderboard submission (PSYDEFCONV test set, $\times 8$).

Figure 2: (a) The PSYDEFCONV official test set label distribution and (b) row-normalized confusion matrix of our official leaderboard submission ($\times 8$). Label 7 dominates both the distribution (243/472 instances) and predictions, absorbing errors from all other classes.

**The primacy of definition quality.** A key finding is the sensitivity of model performance to prompt design. Compared to the baseline of Na et al. (2026b), which uses shallow class descriptions and achieves a Macro-F1 of 8.63%, our best setting ($\times 2$) improves by 19.36 pp. We attribute this to the richer definitional context provided by our DMRS-based definitions derived from Di Giuseppe and Perry (2021), which better disambiguate overlapping classes and reduce label ambiguity during generation.

## 4 Conclusion

Clinical specificity of class definitions is the primary driver of synthetic augmentation effectiveness in PDM classification. Our hybrid system, combining MentalRoBERTa with DMRS-derived features and stressor-anchored generation, achieves substantial gains over DMRS CO-PILOT (Accuracy: 18.01%→58.26%; macro-F1: 8.63%→24.62%). However, the Label 7 sink effect and resulting bimodality indicate that augmentation alone cannot overcome majority-class bias and clinical proximity. Future works should consider including constraint-based decoding, human-in-the-loop validation, and dialogue-level modeling to address temporal volatility identified in our analysis.

## Limitations

**Majority-class dominance and augmentation ceiling.** The most critical limitation is the Label 7 sink effect identified in Section 3.2. Despite augmenting minority classes to $N = 500$, the model's decision boundary remains heavily biased toward Label 7 (243/472 development instances), and macro-F1 performance on six of eight classes remains below 0.30. This suggests that naive count-balancing is insufficient when the majority class also exhibits high linguistic surface overlap with adjacent classes. Addressing this will require loss re-weighting strategies (e.g., focal loss (Lin et al., 2017)), hard-negative mining during augmentation, or explicit contrastive learning objectives that sharpen inter-class boundaries rather than simply expanding minority class size.

**Turn-level modeling and temporal blindness.** Our proposed architecture operates on isolated seeker utterances, each formatted with only the preceding stressor context. However, our dataset analysis (Appendix A.1) demonstrates that defense levels are unstable, frequently undergoing abrupt transitions across turns, and that larger defensive shifts tend to occur rapidly (Figure 5). A turn-level classifier observes only the outcome of a defensive transition, not the trajectory that produced it. This architectural limitation is especially problematic for clinically adjacent classes (e.g., Labels 6 and 7), whose distinction may reside in discourse-level patterns spanning multiple turns rather than in any single utterance.

**Synthetic data validity and clinical reliability.** Our quality control pipeline filters for inter-annotator agreement ($\kappa \geq 0.60$) using a secondary classifier, providing a practical proxy for label consistency. However, this process does not guarantee clinical validity. A generated utterance may receive consistent classification by both the primary LLM and the secondary classifier while still failing to instantiate the functional psychological role of the target defense. Without human expert validation of a representative sample of synthetic instances, we cannot bound the rate of theoretically invalid but classifier-plausible samples in our training data. This is a fundamental limitation of machine-as-annotator pipelines in psychologically grounded domains, and future work should incorporate systematic clinician review.

**Label underspecification and task formulation ambiguity.** Labels 0 (No Defense) and 8 (Needs More Information) do not correspond to clinical defense mechanisms and thus lack the DMRS indicator structure used to construct our Defense Profile features. For Label 8, the classification signal must be derived from the textual encoder and heuristic features, with the DMRS branch contributing noise rather than discriminative signal. Meanwhile, Label 0 is defined by the absence of a positive mechanism, making it harder to synthesize and harder for the NLI-based indicator scoring to characterize. A principled resolution, treating Labels 0 and 8 as a prior detection stage (defensive and non-defensive and ambiguous) before running the eight-class classifier, is deferred to future work.

## Ethical considerations

This work relies on publicly released datasets (PSYDEFCONV and ESCONV) collected under informed consent and Institutional Review Board (IRB) oversight, with no new human data collection. Synthetic utterances simulating psychological distress are used exclusively for model training; clinical validity is not guaranteed, and expert review is required before any real-world deployment. The system is a research prototype and must not be used as a diagnostic tool.

## Acknowledgments

We would like to thank the organizers of the BioNLP 2026 PsyDefDetect shared task and acknowledge the PSYDEFCONV dataset as the foundation of this work.

## References


Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? Deep learning to the rescue! *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7383–7390.

Michael Bond, Susan Gardner, J. Christian, and J. Sigal. 1983. An empirical study of self-rated defense style. *Archives of General Psychiatry*, 40:333–338.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Phebe Cramer. 1987. The development of defense mechanisms. *Journal of Personality*, 55(4):597–614.

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.

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*, pages 7184–7190, Marseille, France. European Language Resources Association.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2020. Supervised multimodal bitransformers for classifying images and text. *Preprint*, arXiv:1909.02950.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

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 ACL-IJCNLP 2021*, pages 3469–3483. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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.

George E. Vaillant. 1994. Ego mechanisms of defense and personality psychopathology. *Journal of Abnormal Psychology*, 103(1):44–50.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

# A Appendix

## A.1 Dataset Analysis

We conduct an exploratory analysis of PSYDEF-CONV to characterize its structural and temporal properties. Class imbalance is also reported in (Na et al., 2026b).


Figure 3: Class distribution across defense levels in the development set. Level 7 (High-Adaptive) dominates with 968 instances; Level 8 (Needs Info) contains only 28, motivating soft-balancing augmentation.

**Temporal Volatility of Defense States.** Defense levels are not stable within a dialogue (Fig. 4). Frequent transitions across levels, including abrupt shifts between adaptive and disavowal patterns within a single session, indicate that classification cannot rely on static turn-level features alone and must account for broader discourse context.

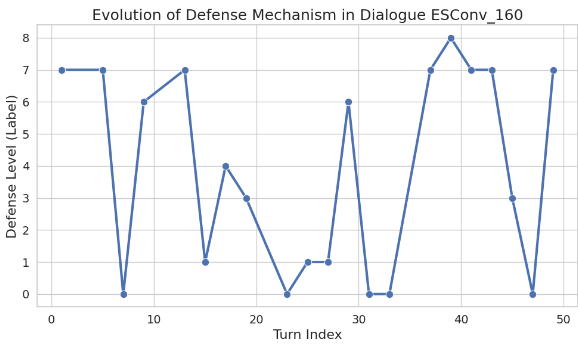



Figure 4: Defense level trajectory across turns in dialogue ESConv_016. Frequent transitions, including abrupt shifts between Level 7 and Level 0, indicate that defense states are temporally unstable and cannot be modeled from isolated turns.

**Defense Volatility: Magnitude and Speed of Change.** The scatter plot (Fig. 5) shows a positive correlation between the magnitude and speed of defense-level transitions: larger shifts in defensive functioning tend to occur over fewer turns.

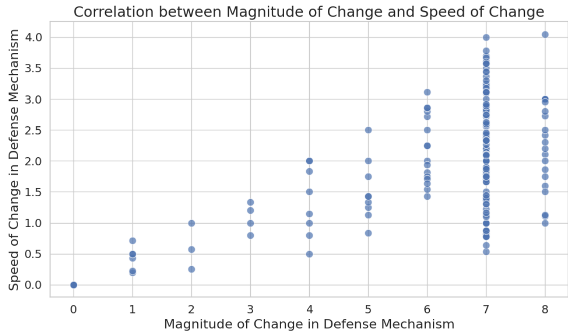



Figure 5: Correlation between magnitude and speed of defense level change across dialogue turns. Larger shifts in defense level tend to occur more rapidly, suggesting that defensive transitions are abrupt rather than gradual, a property that static turn-level classifiers are structurally unable to capture.

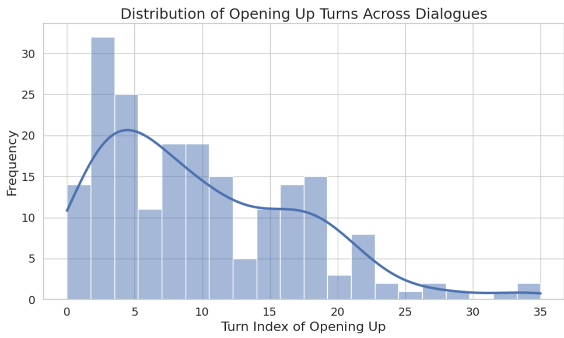



Figure 6: Distribution of turns at which seekers exhibit increased openness. The modal opening-up turn is 3–4, indicating early disclosure before defensive consolidation.

This is a key empirical finding, it implies that when a seeker's defense changes, it changes quickly and dramatically, rather than gradually. This property motivates dialogue-level or sequential modeling as a future direction, as turn-level classifiers observe only the outcome of a transition, not its dynamics.

**Disclosure Dynamics.** Analysis of the CDI (Fig. 7) reveals a consistent disclosure peak around the 10-20% mark, followed by gradual stabilization. The opening-up distribution (Fig. 6) confirms that seekers tend to disclose early (modal turn $\approx$ 3-4), suggesting that defensive activation intensifies *after* initial vulnerability rather than preceding it.

**Response Latency as a Defensive Signal.** Seeker response time varies across defense levels (Fig. 8). Label 0 (No Defense) shows higher latency variance, while defense-related responses cluster more tightly. This supports the use of temporal features as auxiliary signals.

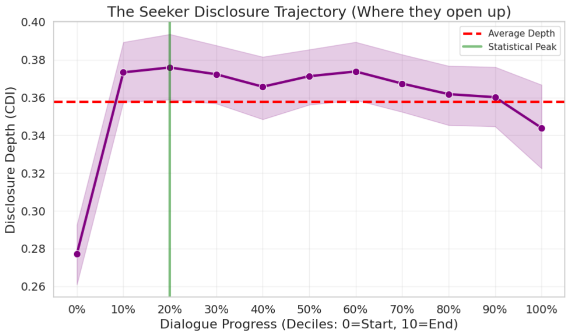



Figure 7: Composite Disclosure Index (CDI) across normalized dialogue progression. Disclosure peaks at the 10–20% mark then stabilizes, suggesting defensive activation intensifies after initial vulnerability.

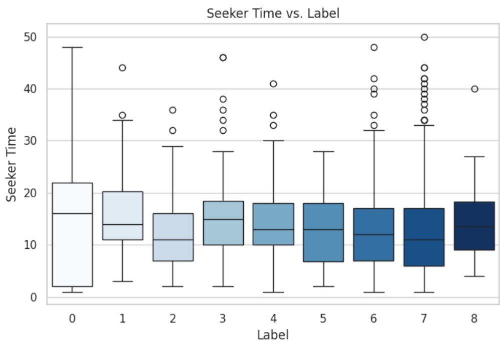



Figure 8: Seeker response time per defense label. Label 0 (No Defense) shows notably higher latency variance; defense-related labels cluster in tighter distributions, supporting temporal features as auxiliary classification signals.

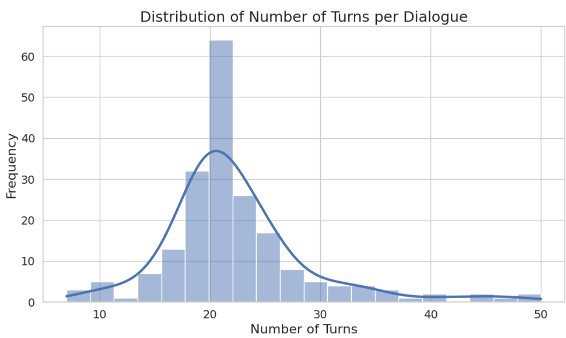



Figure 9: Distribution of number of turns per dialogue. Dialogues average approximately 20 turns, confirming the multi-turn nature of the classification task.

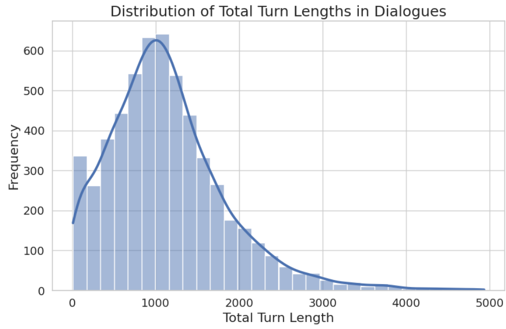



Figure 10: Distribution of total turn lengths per dialogue. The right-skewed distribution peaks around 1,000 tokens, with a long tail of extended sessions up to 5,000 tokens.

**Corpus Structure.** Dialogues average 20 turns and 1,000 tokens in total turn length (Figs. 9 and 10), confirming the multi-turn nature of the task and the need for context-aware modeling beyond single utterances.

## A.2 Extended Result Analysis

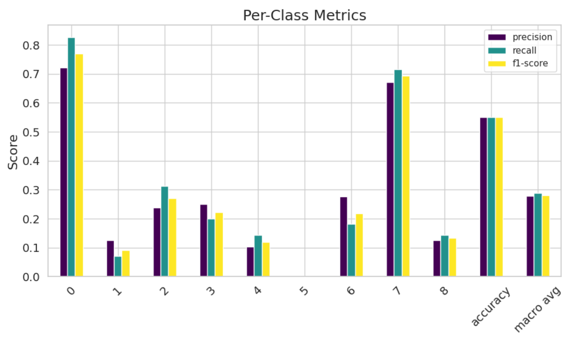



Figure 11: Per-class Precision, Recall, and F1 under the best setting (×2). Labels 1, 4, 5, and 8 remain below F1 = 0.15; Label 5 (Neurotic) achieves zero precision and recall, consistent with its severe underrepresentation (13 dev instances).

**Per-Class Metrics.** Labels 1, 4, 5, and 8 each yield F1 below 0.15. Label 5 (Neurotic) is never predicted, consistent with only 13 dev instances and chronic underrepresentation across all augmentation scales.

**DMRS Mechanism Activation Patterns.** Activation values are uniformly low (range: -0.8 to -1.7), reflecting the implicit nature of defensive language. Despite this, differential patterns across classes, particularly on *Autistic Fantasy*, *Undoing*, and *Affiliation*, confirm that the Defense Profile carries discriminative signal that complements the contextual encoder in the fusion architecture.

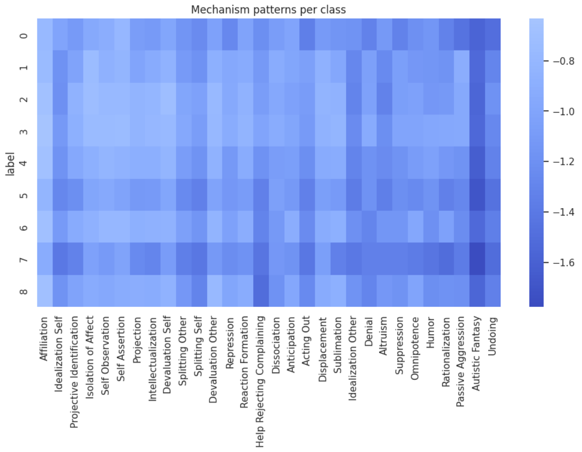



Figure 12: Mean NLI-inferred DMRS mechanism activation per defense class (log-entailment scores). All values are negative due to log-probability scaling. Differential gradients on *Autistic Fantasy*, *Undoing*, and *Affiliation* provide discriminative signal for the hybrid fusion model despite uniformly low absolute scores.

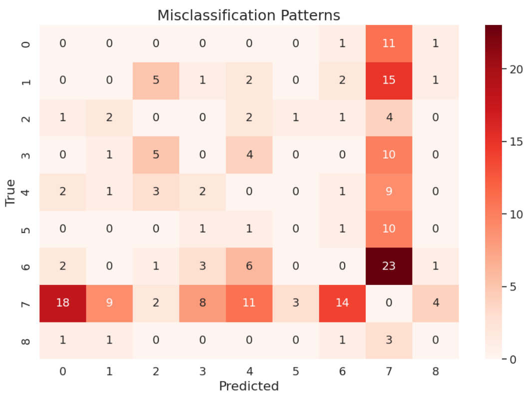



Figure 13: Off-diagonal misclassification counts (best setting, $\times 2$). Label 7 acts as a prediction sink across all classes. The Label 6 $\rightarrow$ 7 confusion (23 errors) is the largest single off-diagonal cell, reflecting clinical proximity between obsessional and high-adaptive defenses.

**Misclassification Patterns.** The off-diagonal error analysis (Fig. 13) reveals that Label 7 is the dominant prediction sink: it absorbs the largest share of errors from every other class, accounting for 11, 15, 10, 9, 10, 23, and 3 misclassified samples from Labels 0-8 respectively. This is not random confusion but a systematic bias toward the majority class. Notably, Label 6 (Obsessional) is misclassified as Label 7 in 23 of 44 cases (52%), suggesting high clinical proximity between obsessional and high-adaptive functioning, a distinction that requires deeper contextual modeling to resolve.

### A.3 Prompt Template for Synthetic Data Generation

We employ a theory-driven prompting strategy to generate synthetic utterances conditioned on stressors, dialogue history, and clinical defense mechanisms. The template used for generation is shown below:

#### Llama3 Prompt

```
prompt = f"""
    ### TASK: Generate Synthetic Psychological Defense Examples
    You are simulating a seeker in a mental health support chat.

    ### CONTEXTUAL GROUNDING:
    STRESSOR: {stressor}
    DIALOGUE HISTORY:
    {history}

    ### DEFENSE TO SIMULATE:
    Mechanism: {mechanism_name} (Level {level})
    Definition: {definition}
    Pattern: {pattern_description}

    ### REFERENCE STYLE (Few-Shot):
    1. "{example_1}"
    2. "{example_2}"
    3. "{example_3}"

    ### GOAL:
    Generate 5 NEW seeker utterances for the NEXT TURN using the {mechanism_name} defense.
    Ensure they follow the history and react to the stressor.

    ### OUTPUT FORMAT:
    1 string.
    No explanation, no markdown, no code fences.
    """
```

### A.4 Prompt Template for Stressor Identification

#### Llama3 Prompt

```
prompt = f"""
    ### TASK: Clinical Stressor Identification
    Identify the "Salient Stressor" causing psychological conflict in the Target Utterance.

    ### DIALOGUE CONTEXT:
    {history}

    ### TARGET UTTERANCE:
    "{target_turn}"

    ### OUTPUT FORMAT:
    1. Stressor Category: (e.g., Interpersonal Conflict, Self-Esteem Threat, External Crisis)
    2. Description: (One sentence explaining the threat)
    """
```