TONI-NLP at PsyDefDetect: Defense Mechanism Detection via LLM-based Ensemble Methods

Durjoy C. Paul[1], Callum Chan[2], Arshitha Basavaraj[3], Veronica Perez-Rosas[1]
Diana Inkpen[2], Francisco Pereira[4], Juan Antonio Lossio-Ventura[4]
[1]Texas State University, USA, [2]University of Ottawa, Canada,
[3]International Institute of Information Technology, Bangalore, India,
[4]National Institute of Mental Health, National Institutes of Health, USA
{xei29,vperezr}@txstate.edu, {cchan073,diana.inkpen}@uottawa.ca,
arshitha.basavaraj@iiitb.ac.in, {francisco.pereira,juan.lossio}@nih.gov

Abstract

This system paper presents the approach of Team TONI-NLP to the PsyDefDetect 2026 shared task. The objective of the task was to classify utterances from helper–seeker conversations into nine categories: seven labels representing progressively higher levels of defensive maturity, one label indicating the absence of a defense mechanism, and one label for cases requiring additional information. We investigated several modern NLP approaches, including prompt engineering, fine-tuning, hierarchical modeling and classification using text embeddings derived from transformer-based models as well as classical embeddings such as TF-IDF. Our results show that ensemble methods performed best among our submitted systems, achieving a macro-F1 score of 0.320 and ranking 9th in the shared task out of 21 teams.

1 Introduction

The field of Natural Language Processing (NLP) has seen a surge in research dedicated to extracting mental health insights from textual data. While most of these efforts have leveraged large-scale social media data for psychological analysis (Garg, 2023; Skaik and Inkpen, 2020), recent work has begun to shift the focus to conversational data, particularly in the context of emotional and mental health support dialogues(Na et al., 2025).

This paper describes our team's participation in the PsyDefDetect 2026 shared task at BioNLP 2026. This task aims to advance the understanding of defense mechanisms expressed in emotional support dialogues. To address the complexities of identifying these mechanisms, we propose a multi-faceted approach utilizing Large Language Models (LLMs) and other traditional NLP methods.

2 Shared Task Description

The PsyDefDetect 2026 shared task focuses on identifying psychological defense mechanisms from multi-turn emotional support dialogues (Na et al., 2026a), involving two participants: a help-seeker (the person sharing their difficulties) and a supporter (also called helper). The task is grounded in the Defense Mechanism Rating Scales (DMRS) (Perry and Henry, 2004) and uses the PSYDEFCONV dataset of 200 seeker–supporter dialogues (Na et al., 2026b).

Dataset. For this task, 1,864 help-seeker turns were provided as training and 472 as test sets, totaling 2,336 turns. Hereafter, we use *utterance* to refer to a help-seeker turn. The defense labels include: No Defense, Action, Major Image-Distorting, Disavowal, Minor Image-Distorting, Neurotic, Obsessional, High-Adaptive, and Needs More Information. The training set is highly imbalanced and skewed towards the *High-Adaptive* label.

Train/Validation Split. Since no official validation set was provided, we partitioned the 1,864 training samples into training and validation sets at the conversation level to avoid data leakage: all turns from a given dialogue were assigned exclusively to one split.

From 200 unique dialogues, we selected 20 conversations (10%) for validation using a greedy scoring strategy that maximizes label diversity, approximates the global class distribution, and penalizes longer conversations. This yields 1,592 training samples (180 conversations) and 272 validation samples (20 conversations), an approximately 85/15 split. Table 1 reports the per-label distribution across the training and validation splits. Label proportions in the validation set closely mirror those in the full dataset, with validation percentages ranging from 11% (label 7, *High-Adaptive*) to 29% (label 8, *needs more information*), reflecting the inherent difficulty of balancing rare classes under conversation-level constraints. No dialogue leakage was detected between splits. Figure 1 provides a visual summary of these distributions.

Proceedings of the BioNLP 2026 (Shared Tasks), pages 132–140

Label	Defense Type	Train	Val	Total
0	No Defense	248 (83.8%)	48 (16.2%)	296
1	Action	89 (82.4%)	19 (17.6%)	108
2	Major Image-Distorting	49 (80.3%)	12 (19.7%)	61
3	Disavowal	77 (77.8%)	22 (22.2%)	99
4	Minor Image-Distorting	66 (78.6%)	18 (21.4%)	84
5	Neurotic	37 (77.1%)	11 (22.9%)	48
6	Obsessional	145 (84.3%)	27 (15.7%)	172
7	High-Adaptive	861 (89.0%)	107 (11.0%)	968
8	Needs More Information	20 (71.4%)	8 (28.6%)	28
Total		**1,592**	**272**	**1,864**

Table 1: Per-label sample counts and percentages across the training and validation splits. Percentages indicate the proportion of each label's total samples assigned to each split.

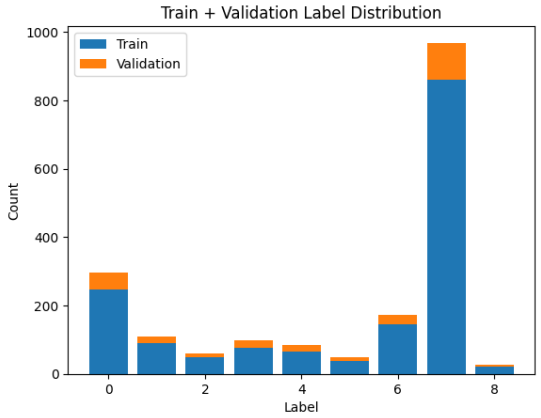


Figure 1: Label distribution across training and validation splits. The validation set closely mirrors the overall distribution, though the strong class imbalance is evident, particularly for the High-Adaptive class (label 7), which accounts for over half of all samples.

3 Methodology

We investigated four approaches for defense mechanism classification: prompting, hierarchical classification, supervised fine-tuning, and ensemble methods. These approaches were evaluated within a unified experimental setup using training and validation splits derived from the training data, as described in the previous section, enabling systematic comparison across methods.

3.1 Prompting

We conducted zero-shot and few-shot prompting experiments with instruction-tuned LLMs, following the prompt structure of Chan et al. (2025) and Lossio-Ventura et al. (2025). We evaluated prompt optimization, random few-shot prompting, and dynamic example retrieval across three model families: Llama (Llama-3.1-8B-Instruct, Llama-3.3-70B-Instruct), Claude (claude-opus-4-6 and claude-sonnet-4-6), and GPT (gpt-5.2).

3.1.1 Zero-shot experiments

We started with the zero-shot prompt provided by the task organizers and iteratively refined it based on classification performance. This process, informed by the DMRS annotation guidelines (Di Giuseppe and Perry, 2021), involved adding additional context, such as detailed label definitions, along with keysigns, workflow guidelines and disambiguation rules to capture nuanced distinctions between classes. All models were prompted to respond with a single digit (0–8) corresponding to the predicted defense level.

3.1.2 Few-shot experiments

We applied the context strategies described above to few-shot prompting and included labeled examples in the prompt. We experimented with few-shot prompting using randomly sampled examples (random few-shot) and also with examples selected based on semantic similarity (dynamic few-shot). For dynamic few-shot, we retrieved examples using cosine similarity over sentence embeddings, experimenting with $k \in \{5, 10, 20, 25\}$ examples per class to assess the effect of example coverage on classification performance.

Implementation details. All models received the same prompt structure, consisting of a task description, the DMRS annotation handbook, the label definitions, the full dialogue context, and the target utterance. We experimented with both the full and condensed handbook versions. Since Llama and Claude cap temperature at 1.0 while GPT allows up to 2.0, we sampled $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ and $\{0.4, 0.8, 1.2, 1.6, 2.0\}$ respectively, covering each model's temperature range. All models were queried with `top_p=0.7`.

3.2 Hierarchical Approach

We also performed experiments with a hierarchical classification approach inspired by the Defense Mechanism Rating Scales (DMRS) hierarchical organization of defensive categories (Di Giuseppe and Perry, 2021). Initially, we distinguished between four aggregated defense categories: Mature Defenses, Neurotic Defenses, Immature Defenses, and No Defense. We then performed a secondary categorization step using the fine-grained subcategories shown in Table 2. We used a Multinomial Logistic Regression (MLR) classifier with TF-IDF features combined with contextualized embeddings.

These embeddings are obtained using the simplified SBERT-WK (Wang and Kuo, 2020) pooling strategy to MentalBERT (Ji et al., 2022) by averaging the final four hidden layers.

Aggregated Category	Specific Defense Label
No Defense	0 – No Defense 8 – Needs More Information
Mature Defenses	7 – High-Adaptive
Neurotic Defenses	5 – Neurotic 6 – Obsessional
Immature Defenses	1 – Action 2 – Major Image-Distorting 3 – Disavowal 4 – Minor Image-Distorting

Table 2: Mapping between the four aggregated categories and their corresponding specific defense labels.

Similar to our prompting methods, we experimented with context augmentation strategies. We conducted two experiments: (1) using the target utterance only, and (2) using the target utterance plus context, consisting of the previous conversational turn[1]. For comparison, we also ran a 9-way classification task for all categories using the target utterance as the sole input.

3.3 Fine-tuning LLMs

We explored several fine-tuning strategies for the classification task, ranging from encoder-only discriminative models to decoder-only generative LLMs, with varying approaches to class imbalance mitigation, input representation, and parameter efficiency. We organized our fine-tuning experiments into three groups: **(i)** encoder-based discriminative models, **(ii)** decoder-only LLMs with a classification head, and **(iii)** generative LLMs trained in an instruction-following format. Table 3 summarizes the key design choices across all approaches.

3.3.1 Encoder-based Discriminative Models

We fine-tuned a set of encoder-only transformer models such as BERT (base and large), RoBERTa, and domain-specific mental variants, given the clinical nature of the task, such as MentalBERT and MentalRoBERTa (Ji et al., 2022) for direct sequence classification. All models operated on the *target utterance* only, with a maximum sequence length of 512 tokens. A grid search over learning rates $\{2e{-}5, 3e{-}5, 4e{-}5, 5e{-}5, 4e{-}4\}$, batch sizes $\{8, 16, 32\}$, and weight decay $\{0.01, 0.05, 0.1\}$ was conducted over 10 epochs. To address class imbalance, we explored two strategies:
(1) *Loss-based balancing:* inverse-frequency class weights were applied to the cross-entropy loss.
(2) *Hybrid balancing (data + loss):* majority classes were undersampled (e.g., label 7 capped at 200 samples, label 0 at 150), and class weights recomputed on the balanced dataset.

3.3.2 Decoder LLMs with Classification Head

We adapted causal LLMs for classification by attaching a linear head to the final token representation of a LoRA-adapted model. We evaluated several configurations differing in parameter efficiency, input context, and training strategy:
(1) *QLoRA (8B, 70B):* 4-bit quantized fine-tuning (Dettmers et al., 2023) with LoRA ($r = 16$, $\alpha = 32$) applied to attention projections. The input consisted of the target utterance only (512 tokens). The optimization used cosine scheduling with warmup and `paged_adamw_8bit`.
(2) *LoRA with dialogue context (8B):* full-precision BFloat16 training with LoRA applied to all projection layers. We evaluated configurations where $(r \in 8, 16, 32, 64)$ and (α) is set to $(2r)$. Inputs included full dialogue plus target utterance (up to 1,400 tokens).
(3) *LoRA with 5-fold cross-validation:* stratified folds were built on the initial training data, with class weights recomputed per fold. The configurations were $(r \in 8, 16, 32, 64)$ and $(\alpha = 2r)$.
(4) *One-vs-rest (OvR):* nine binary classifiers were trained with LoRA, combining class-weighted loss and minority-class oversampling. Predictions were aggregated via softmax renormalization.

3.3.3 Generative Instruction-Tuned LLMs

We reformulate the task as instruction-based generation, where the model predicts the label digit as the output token. The loss was computed only over the generated label token, with prompt tokens masked. We explored two fine-tuning strategies:
(1) *Full fine-tuning (8B):* using the llama-cookbook FSDP framework. Minority classes were oversampled to a minimum of 300 samples, yielding approximately 2,700 training instances. Training used learning rates of $\{5e{-}5, 4e{-}4\}$ for 10 epochs, with the checkpoint achieving the lowest validation loss selected for evaluation.
(2) *LoRA fine-tuning (70B):* using parameter-

[1]We experimented with other context sizes but did not observe a significant improvement

Group	Approach	Context	Imbalance Handling
Encoder	Weighted loss	Utterance	Loss-based (class weights)
	+ Undersampling	Utterance	Data + Loss (undersampling + weights)
Decoder + Head	QLoRA (8B, 70B)	Utterance	Loss-based (class weights)
	LoRA (8B)	Utterance / Dialogue	Loss-based (class weights)
	LoRA 5-Fold (8B)	Utterance / Dialogue	Loss-based (class weights)
	OvR (8B)	Utterance / Dialogue	OvR + Oversampling
Generative	Full FT (8B)	Dialogue	Oversampling
	LoRA (70B)	Dialogue	Oversampling

Table 3: Summary of fine-tuning approaches. "Context" indicates whether only the target utterance or the full dialogue was used as input. "Imbalance Handling" distinguishes between loss-based, data-level, and hybrid strategies. Note that the encoder group was based only on BERT/RoBERTa families. **8B** refers to Llama-3.1-8B-Instruct and **70B** refers to Llama-3.3-70B-Instruct.

efficient adaptation with LoRA ($r = 16$, $\alpha = 32$). Training used learning rates of $\{5e{-}5, 4e{-}4\}$. Despite scalability, validation loss analysis showed early overfitting, with the optimal checkpoint occurring at epoch 1.

3.4 Ensemble Methods

Our final approach used ensemble methods to combine predictions from multiple models. Ensemble methods improve performance by aggregating models, helping generalization to new data (Dietterich, 2000) and have proven effective across different NLP tasks (Zhang and Shafiq, 2024). After official results were released, we conducted post-hoc ensemble analyses over our submitted systems. We compared majority voting with weighted averaging using post-release system-level macro F1 scores. These analyses were not used to tune or select final submissions or guide model selection.

We created two different ensembles. The first (*Across approaches*) combined our best submissions from prompting, hierarchical classification, and fine-tuning. The second (*Best submissions*) combined our three strongest systems: *(Decoder) LoRA (8B) - utterance*, *(Decoder) LoRA (8B) - dialogue*, and *(Generative) Full FT (8B)*. For majority voting, ties were resolved using the Decoder LoRA (8B) utterance model, our best individual system. For weighted averaging, each system was weighted by its post-hoc macro F1 score-based performance.

4 Results

Table 4 presents results on the test data. The table reports prompt-engineering results, followed by hierarchical approaches, fine-tuned LLMs, and ensembles. We report macro-averaged precision, recall, and F1, following the shared-task metric. For prompting methods, dynamic few-shot prompting achieves the best macro F1 (0.197), followed by zero-shot prompting (0.193), while random few-shot prompting performs worse (0.078). The best configuration uses $k = 20$ examples per class retrieved via cosine similarity, with performance decreasing for both $k = 25$ and $k = 5$. For hierarchical methods, the best performance is obtained using a two-step configuration restricted to the utterance (macro F1 = 0.238). This setup outperforms both the single-step baseline (0.236) and the two-step variant using the previous conversational turn (0.217). For fine-tuned models, the decoder-only LoRA (8B) with utterance-only input gets the highest macro F1 (0.303), slightly exceeding the dialogue-context variant (0.300) and outperforming encoder-based models (MentalBERT: 0.252; MentalRoBERTa: 0.263). Overall, ensemble methods achieve the best performance. The best submission (voting) reaches a macro F1 of 0.320 by combining our three strongest systems, outperforming the across-approaches voting ensemble (0.300). Averaging-based ensembles perform poorly (0.212 and 0.105), suggesting that majority voting is more robust for this task. These results place our system 9th out of 21 teams on the leaderboard.

5 Discussion

In our prompting experiments, dynamic few-shot prompting using the top 20 most similar examples outperformed structured zero-shot prompting with the annotation handbook as additional context. In contrast, randomly selected few-shot examples performed worse than both approaches, likely due to poorer class coverage and lower example relevance. Performance also varied with the number of retrieved examples and the embedding model,

Approach	Method	Accuracy	Precision	Recall	Macro F1
Prompting	Zero-shot (Baseline)	0.364	0.500	0.134	0.193
	Random Few-shot	0.277	0.131	0.145	0.078
	Dynamic Few-shot	0.511	0.198	0.219	**0.197**
Hierarchical	Two-step (Utterance + Previous Turn)	0.532	0.213	0.224	0.217
	Two-step (Utterance)	0.534	0.241	0.242	**0.238**
	Single-step (Utterance)	0.519	0.225	0.267	0.236
LLM Fine-tuning	(Encoder) MentalBERT	0.532	0.254	0.264	0.252
	(Encoder) MentalRoBERTa	0.568	0.356	0.258	0.263
	(Decoder) LoRA (8B) (Utterance)	0.532	0.293	0.325	**0.303**
	(Decoder) LoRA (8B) (Dialogue)	0.657	0.367	0.278	0.300
	(Generative) Full FT (8B)	0.657	0.375	0.260	0.289
Ensemble	Across approaches (Voting)	0.591	0.299	0.304	0.300
	Across approaches (Averaging)	0.439	0.265	0.231	0.212
	Best submissions (Voting)	0.672	0.470	0.284	**0.320**
	Best submissions (Averaging)	0.193	0.098	0.176	0.105

Table 4: Results for team TONI-NLP on the PsyDefDetect 2026 shared task test set, grouped by approach. Within each group, **bold** denotes the best macro F1 and underline denotes the second best. **8B** refers to Llama-3.1-8B-Instruct. A simple voting over our best submissions achieved the highest score. Only systems included in the submitted/tested result set are shown.

indicating that both retrieval quality and class coverage are important. While $k = 25$ showed signs of prompt saturation, $k = 5$ provided insufficient coverage, with $k = 20$ the best balance.

The two-step hierarchical approach produced only marginal gains over the single-step baseline using the utterance alone, while adding one prior turn reduced test performance. The two-step design reduces the classification space by routing instances to a subset-specific classifier. However, this introduces error propagation from the first stage. In contrast, the single-step MLR shows a bias toward majority classes (Appendix Figure 2).

Across all fine-tuned models, using only the target utterance matched or outperformed full dialogue context, consistent with findings from the hierarchical approach. Among encoder models, MentalRoBERTa and MentalBERT outperformed the non-specialized BERT models. All encoder models were fully fine-tuned given their small size. Full fine-tuning of Llama-3.1-8B-Instruct was feasible but computationally expensive, while Llama-3.3-70B-Instruct could only be fine-tuned using LoRA due to GPU memory constraints. LoRA fine-tuning of the 8B model got the best result (0.303), suggesting that parameter-efficient adaptation is both more scalable and more effective for this task.

Voting outperformed averaging-based ensembles, suggesting that hard voting is more robust to poorly calibrated confidence scores. The best ensemble (0.320) combined decoder LoRA (utterance), decoder LoRA (dialogue), and generative full fine-tuning. Their diversity in input context and training strategy likely produced complementary error patterns and gains over any single model.

6 Conclusion and Future Work

We described TONI-NLP's participation in PsyDefDetect 2026, comparing prompting, hierarchical classification, fine-tuning, and ensemble strategies for defense mechanism detection. Our results showed that ensemble methods achieved the strongest overall performance, while target-utterance-only configurations performed well across both hierarchical and fine-tuned models. These findings suggest that reducing input complexity and combining complementary systems may be useful strategies for this task. In future work, we plan to further investigate hierarchical approaches using LLMs and explore data balancing strategies at each step of the classification pipeline.

7 Limitations

Our experiments were limited to the English mental health dataset released by the shared-task organizers. Given the dataset's small size and class imbalance, the generalizability of our findings should be interpreted with caution. The limited shared-task timeframe also restricted experimentation, preventing extensive ablation studies. Finally, the methods presented in this work were developed solely for research purposes within the shared-task setting and are not intended for clinical use.

8 Ethics

This work uses the annotated PSYDEFCONV dataset provided by the shared-task organizers. The conversational data was crowdsourced from consenting participants, and any future use of the dataset should continue to preserve participant anonymity. Because our approach relies on the PSYDEFCONV annotation scheme and the DMRS psychological framework, it may reproduce or amplify biases embedded in these theoretical and annotation frameworks. Therefore, any future application of this work should involve multidisciplinary collaboration with clinical psychologists and affected communities, particularly if the methods are considered for use beyond research settings.

Acknowledgments

Research reported in this publication was supported by the Natural Science and Engineering Research Council of Canada (Diana Inkpen). Juan Antonio Lossio-Ventura and Francisco Pereira were supported by the National Institute of Mental Health Intramural Research Program (ZICMH002968). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

Callum Chan, Sunveer Khunkhun, Diana Inkpen, and Juan Antonio Lossio-Ventura. 2025. Prompt engineering for capturing dynamic mental health self states from social media posts. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 256–267, Albuquerque, New Mexico. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Muskan Garg. 2023. Mental health analysis in social media posts: A survey. *Archives of Computational Methods in Engineering*, 30.

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.

Juan Antonio Lossio-Ventura, Callum Chan, Arshitha Basavaraj, Hugo Alatrista-Salas, Francisco Pereira, and Diana Inkpen. 2025. 5cNLP at BioLaySumm2025: Prompts, retrieval, and multimodal fusion. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 215–231, 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 and Melissa Henry. 2004. Chapter 9 - studying defense mechanisms in psychotherapy using the defense mechanism rating scales. In Uwe Hentschel, Gudmund Smith, Juris G. Draguns, and Wolfram Ehlers, editors, *Defense Mechanisms*, volume 136 of *Advances in Psychology*, pages 165–192. North-Holland.

Ruba Skaik and Diana Inkpen. 2020. Using social media for mental health surveillance: A review. *ACM Comput. Surv.*, 53(6).

Bin Wang and C-C Jay Kuo. 2020. Sbert-wk: A sentence embedding method by dissecting bert-based word models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2146–2157.

Hongzhi Zhang and M Omair Shafiq. 2024. Survey of transformers and towards ensemble learning using transformers for natural language processing. *Journal of big Data*, 11(1):25.

A Appendix

A.1 Additional Results for the Hierarchical Approach

We present additional results on the validation data held out from the training set. Table 5 shows the results of the hierarchical approach (2 steps), and compares them with the results of the 9-way classifier (Single step).

A.2 Aggregated Category Classification

In developing the hierarchical approach, we conducted a series of experiments on our validation set to optimize the initial categorization. We evaluated both MLR and Support Vector Machine (SVM) models across various feature representations, including multiple combinations of TF-IDF, standalone mental-bert, and SBERT-WK embeddings. Furthermore, we experimented with incorporating different context sizes (e.g., 0, 1, and 3 prior dialogue turns) alongside the target utterance.

This comparative evaluation was utilized to classify the three primary aggregated defense categories: Immature, Neurotic, and Mature, explicitly excluding the "No Defense" category. As detailed in Table 6, our results indicate that the target utterance only setting combined with an MLR classifier using concatenated SBERT-WK (mental-bert-base-uncased) and TF-IDF features yielded the best performance.

Features	**Method**	**3 Turns**		**1 Turn**		**0 Turns**	
		Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
TF-IDF	MLR	0.4676	0.3772	0.5417	0.4588	0.5093	0.4362
	SVM	0.4722	0.3725	0.5139	0.4143	0.5185	0.4380
MentalBERT	MLR	0.5046	0.4381	0.5231	0.4632	0.5463	0.4884
	SVM	0.4537	0.3885	0.5324	0.4587	0.5463	0.4666
MentalBERT + TF-IDF	MLR	0.4907	0.4160	0.5370	0.4581	0.5694	0.4997
	SVM	0.4769	0.3840	0.5000	0.4021	0.5648	0.4899
SBERT-WK (mental-bert-base-uncased) + TF-IDF	**MLR**	0.5417	0.4610	0.5694	0.4977	**0.5833**	**0.5189**
	SVM	0.5185	0.4390	0.5046	0.4106	0.5139	0.4478

Table 6: Performance comparison for the first-step aggregated classifier in different model configurations.

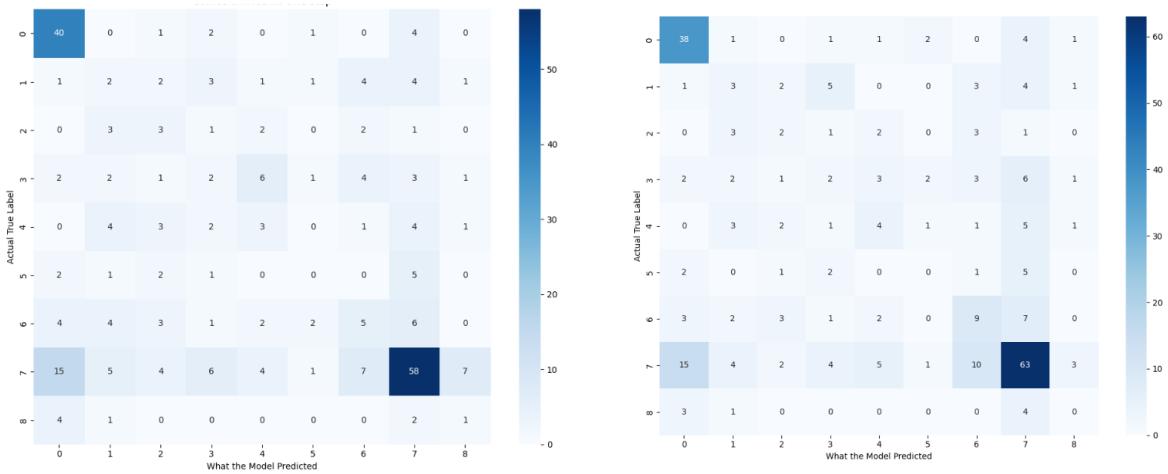
A comparison of the confusion matrices on Figure 2 reveals that, the single-step baseline struggled because it defaulted too often to the majority classes (Class 7 and Class 0). The hierarchical model improved this by routing data into aggregated categories first. This prevented mid-tier classes from being overshadowed by the majority.

Class (Defense Categ.)	Support (N)	Single-Step F1	Hier. F1	Δ F1	Key Error Dynamic / Impact of Hierarchical Routing
0 (No Defenses)	48	0.69	0.68	-0.01	Remains default prediction; reduced swamping of mid-tier classes.
1 (Action Defenses)	19	0.10	0.16	+0.06	Improved boundary resolution; less conflation with Class 6.
2 (Major Image-Distorting)	12	0.19	0.16	-0.03	Signal degradation; misclassified through sequential routing.
3 (Disavowal Defenses)	22	0.10	0.10	0.00	Persistent low discriminative power across both architectures.
4 (Minor Image-Distorting)	18	0.17	0.23	+0.06	Benefited from localized, secondary-level feature evaluation.
5 (Neurotic Defenses)	11	0.00	0.00	0.00	Fundamental data sparsity; unresolvable by structural changes.
6 (Obsessional Defenses)	27	0.20	**0.32**	**+0.12**	**Significant Gain:** Successfully shielded from Class 0 swamping.
7 (High-Adaptive Defenses)	107	0.60	0.61	+0.01	Marginal stabilization; remains the dominant statistical bias.
8 (Needs More Information)	8	0.11	**0.00**	**-0.11**	**Total Attrition:** Weak minority signal filtered out by two-step routing.
Global Metrics	**272**	**Acc: 41.9%**	**Acc: 44.5%**	**+2.6%**	Overall gains driven by localized resolution of mid-tier classes.

Note: Δ F1 denotes the absolute change in performance, calculated as (Hierarchical Macro F1 − Single-Step Macro F1). Positive values indicate an improvement utilizing the two-step approach.

Table 5: Comparative Performance and Error Dynamics (Single-Step vs. Two-Step Hierarchical Approach)

For example, true positives for Class 6 almost doubled (from 5 to 9) because the secondary classifier had a narrower, more focused task. However, the hierarchical approach has a noticeable drawback: if a mistake is made at the first step, it would propagate into the next step. This cascading error makes it difficult to correctly predict the lower frequency classes. For instance, the first-level classifier misclassified all Class 8 instances (mostly categorizing them into Classes 0 and 7), so the secondary classifier will not evaluate them. A similar case happened in Class 2, where true positives fell from 3 to 2. Finally, the confusion matrices show that some issues are strictly due to a lack of data. Class 5 has very few examples, and neither model could predict it correctly. In both approaches, nearly half of the Class 5 instances were mistakenly predicted as Class 7.



(a) Single-step Baseline

(b) Two-step Hierarchical Model

Figure 2: Confusion matrices comparing the final 9-way classification performance of (a) the Single-step baseline and (b) the two-step Hierarchical model on the validation set.