

AlienAnnotators at PsyDefDetect: What Lies Between the Lines: Probing Lightweight Open-Source LLMs for Psychological Defense Mechanism Detection

Siam Rahman Karip and Nahid Hossain

United International University

Dhaka, Bangladesh

Abstract

Detecting psychological defense mechanisms in therapy dialogue is a clinically valuable but computationally underexplored task. We present our systematic analysis for PsyDefDetect, a shared task at BioNLP@ACL 2026, which frames defense detection as a nine-class utterance-level classification problem based on the Defense Mechanism Rating Scale (DMRS). We systematically evaluate six open-source, instruction-tuned small language models (SLMs, ≤ 9 B parameters) in zero-shot and fine-tuning settings, and compare a clinically-grounded prompt against the organizer-provided baseline. Our official submission achieved 59.96% accuracy and 16.28% Macro F1. Post-submission experiments show that fine-tuning combined with 5-fold cross-validation and logit averaging ensemble substantially improves performance, with the best configuration reaching 34.59% Macro F1 and 65.25% accuracy. We find that clinically-grounded prompts outperform bare label definitions, model scale does not consistently improve zero-shot performance, and fine-tuning dramatically recovers even collapsed zero-shot models. Certain defense tiers remain persistently difficult across all settings, pointing to clinical ambiguity at tier boundaries as a more fundamental bottleneck than data imbalance alone.

1 Introduction

Natural language processing is increasingly applied to understand different characteristics from client language in clinical interactions (Voultsiou and Moussiades, 2026; Na et al., 2025). However, many clinically meaningful constructs remain difficult to model computationally, such as, psychological defense mechanisms. Defense mechanisms are unconscious strategies that individuals use to manage internal conflict and distress. Accurate identification of defenses can improve case formulation

and therapeutic outcomes. The Defense Mechanism Rating Scale (DMRS) provides a structured framework for this task. It organizes defenses into hierarchical levels based on adaptiveness (Perry and Henry, 2004).

The PsyDefDetect shared task (Na et al., 2026a) operationalizes this problem as a nine-class classification task, based on the DMRS tiers. The task is challenging by design: the dataset is small, the label distribution is severely imbalanced (up to $34.6\times$ between majority and minority classes), and the tier boundaries require clinical judgment to distinguish. These challenges motivate a systematic study of how small-scale LLMs handle this task.

Our main contributions are as follows:

- A systematic zero-shot evaluation of six small-scale LLMs comparing a clinically-grounded prompt against the organizer-provided baseline, revealing that behavioral tier descriptions consistently outperform bare label definitions, and that larger model size does not consistently improve fine-grained clinical classification.
- A fine-tuning pipeline combining clinically-grounded DMRS prompt design, dialogue-grouped cross-validation, and ensemble strategies for robust low-resource clinical classification.
- A brief error analysis examining per-tier classification difficulty, model confusion patterns across clinically similar defense levels, and the effect of class imbalance on minority tier detection.

2 Task and Data

The task uses PSYDEFCONV (Na et al., 2026b), a dataset of emotional support dialogues annotated with DMRS defense levels. Each instance consists of a multi-turn dialogue and a target utterance produced by the help-seeker, which is assigned

one of eight DMRS defense levels, or flagged as requiring more information (L8: Needs More Information) when the dialogue context is insufficient for a confident classification. The official training set contains 1,864 samples, and the official test set contains 472 samples. The dataset exhibits severe class imbalance, with the "Highly Adaptive" tier (L7) comprising 51.9% of training samples, while the rarest class, "Needs More Information" (L8), accounts for only 1.5%, depicting an overall imbalance ratio of $34.6\times$. Table 1 presents the full label distribution.

ID	Label	Count	%
L0	No Defense	296	15.9
L1	Action	108	5.8
L2	Maj. Image-Distort	61	3.3
L3	Disavowal	99	5.3
L4	Min. Image-Distort	84	4.5
L5	Neurotic	48	2.6
L6	Obsessional	172	9.2
L7	High-Adaptive	968	51.9
L8	Needs More Info	28	1.5
Total		1,864	100.0

Table 1: Training set label distribution. Imbalance ratio is computed relative to the majority class (L7).

3 Methodology

3.1 Model Selection

We evaluate six open source, instruction tuned LLMs with at most 9B parameters, selected to cover a range of model families, scales, and pretraining objectives. From the Gemma family, we include Gemma3-1B-it and Gemma2-9B-it (Team et al., 2024). From the Llama family, we include Llama-3.1-8B-Instruct and Llama-3.2-1B-Instruct (Grattafiori et al., 2024). From the Qwen3 family, we include Qwen3-1.7B and Qwen3-8B (Yang et al., 2025). We primarily focus on general-purpose decoder-only models, as their instruction-following capability enables zero-shot evaluation without task-specific adaptation, which is central to our prompt design comparison.

3.2 Prompt Design

We compare two prompt variants. The **organizer-provided baseline prompt** (Variant A) presents the task instruction alongside bare label names and their constituent defense mechanisms. Our **clinically-grounded prompt** (Variant B) replaces these with behavioral descriptions derived from the DMRS manual (Perry and Henry, 2004), providing

the model with observable verbal cues for each tier. For example, the Disavowal tier is described as: *the speaker denies an obvious reality, externalizes blame, justifies behavior with plausible-sounding logic, or retreats into elaborate private fantasy*. Full prompt texts are provided in Appendix 5.

3.3 Zero-Shot Evaluation

We evaluate all six models in a zero-shot setting using both prompt variants. Each model receives the full dialogue context and target utterance, and is required to output a single digit (0–8) with no examples or additional guidance. We compare Variant A and Variant B across all models to assess the effect of clinically-grounded prompt design on zero-shot classification performance.

3.4 Finetuning

We fine-tune three models: Qwen3-1.7B, Gemma3-1B-IT and Llama-3.2-1B-Instruct. All models are trained using standard causal language modeling loss with the clinically-grounded prompt (Variant B) as the input format. To prevent data leakage across dialogue turns, we apply 5-fold stratified cross-validation grouped by dialogue ID, ensuring that all utterances from the same dialogue appear in the same fold. All models are trained with a learning rate of $5e-5$ with cosine decay to $5e-6$, a warmup ratio of 0.1, and a batch size of 4 (effective 16 with gradient accumulation). Training runs for up to 8 epochs with early stopping patience of 2 epochs based on validation Macro F1, and a weight decay of 0.01.

3.5 Ensemble Strategies

Given the 5-fold cross-validation setup, we explore three ensemble strategies over the fold checkpoints:

- **Logit Averaging:** Raw output logits are averaged across all five folds before taking the argmax.
- **Majority Vote:** The most frequent predicted label across folds is selected, with ties broken by logit confidence.
- **Best Single Fold:** The highest-performing individual fold checkpoint is used alone, serving as a non-ensemble baseline.

3.6 Evaluation Metrics

We report three metrics for all experiments. **Macro F1** is our primary metric, as it weights all classes equally regardless of support, directly capturing

Model	Variant A (Baseline)			Variant B (Clinical)		
	Acc	F1	MAE	Acc	F1	MAE
Qwen3-1.7B	25.85	12.17	2.32	34.32	14.57	2.77
Qwen3-8B	22.67	11.16	3.97	29.24	13.63	3.23
Gemma3-1B-it	18.97	8.01	3.45	21.17	12.10	3.01
Gemma2-9B-it	28.97	15.01	4.29	31.14	18.79	3.10
Llama-3.2-1B-Instruct	10.41	7.14	4.21	9.75	6.93	3.46
Llama-3.1-8B-Instruct	15.77	10.79	2.47	16.74	12.41	2.23

Table 2: Zero-shot results under Variant A (organizer-provided baseline) and Variant B (clinically-grounded) prompts on the official test set. Accuracy and Macro F1 (F1) are given in percentage(%) values. Best Macro F1 per model is **bolded**.

performance on minority tiers. **Accuracy** measures overall correctness but is susceptible to majority-class bias given the severe class imbalance. **Mean Absolute Error (MAE)** treats the DMRS tiers as an ordinal scale and penalizes predictions proportionally to their distance from the true tier, which is clinically meaningful given the hierarchical structure of the DMRS.

4 Results and Analysis

4.1 Zero-Shot Results

Table 2 presents the zero-shot performance of all six models under both prompt variants. Variant B (clinically-grounded) consistently outperforms Variant A (baseline) across most models in terms of Macro F1, demonstrating the benefit of behavioral descriptions over bare label names for this task.

4.2 Fine-Tuning and Ensemble Results

Table 3 presents the test set performance of the three fine-tuned models under three ensemble strategies and our leaderboard submission. Our official leaderboard submission was based on a fine-tuned Qwen3-1.7B model, achieving 59.96% accuracy and 16.28% Macro F1. Ensembling consistently improves over the best single fold across all models, with logit averaging and majority vote yielding comparable results. The results reported here reflect post-submission experiments with improved methodology. The official submission used the organizer-provided baseline prompt (Variant A) and was fine-tuned on a 75/15/15 train/validation/test split of the training set, as the official test set was not available during the submission period. Post-submission, we adopted the clinically-grounded prompt (Variant B), applied 5-fold cross-validation grouped by dialogue ID, and evaluated on the released test set, which accounts for the substantial improvement from 16.28% to 34.59% Macro F1.

Model	Strategy	Acc	F1	MAE
Qwen3-1.7B ¹	Logit Avg	65.25	34.59	0.94
	Maj. Vote	63.35	31.29	1.15
	Best Fold	58.90	33.47	1.28
Llama-3.2-1B ¹	Logit Avg	65.25	34.21	1.16
	Maj. Vote	63.14	33.74	1.25
	Best Fold	59.98	32.63	1.33
Gemma3-1B-it ¹	Logit Avg	59.75	27.85	1.34
	Maj. Vote	56.89	26.51	1.38
	Best Fold	59.30	29.46	1.35
Qwen3-1.7B ²	Official Sub.	59.96	16.28	2.45

Table 3: MAE = Mean Absolute Error. Accuracy and Macro F1 (F1) are given in percentage(%) values.¹Post-submission Fine-tuning and ensemble experiment results on the official test set. ²Official leaderboard submission.

4.3 Analysis

Effect of Model Scale in Zero-Shot setting:

Larger models do not consistently outperform smaller ones in zero-shot. Within the Qwen3 family, the 1.7B model (F1=14.57%) outperforms the 8B model (F1=13.63%). In the Gemma and Llama families, larger models do improve over their smaller counterparts, but only modestly (+6.69pp and +5.48pp respectively). This reflects the highly specialized nature of the task: DMRS tier classification requires expert clinical reasoning that is not well-represented in general pretraining data, limiting the benefit of additional parameters.

Per-Tier Difficulty and Class Learnability:

Across all models and settings, L2 (Major Image-Distorting), L5 (Neurotic), and L6 (Obsessional) remain the hardest tiers, with near-zero F1 in zero-shot and modest improvement after fine-tuning, as shown in Figure 1. These tiers are both low-resource and clinically subtle, their boundaries require distinguishing between cognitively similar defensive patterns (e.g., intellectualization vs. ratio-

nalization), which current small models struggle to capture reliably. L0 (No Defense), by contrast, is consistently well-learned across all fine-tuned models with F1 ranging from 77% to 91%, despite not being the majority class. L7 (High-Adaptive), despite comprising 51.9% of training samples, shows more variable F1 (74%–81%) and is frequently confused with L0, as evidenced by consistent off-diagonal mass in the L7 row of the confusion matrices (Appendix, Figure 17). L8 (Needs More Info) shows partial recovery under fine-tuning, reaching up to 33% F1 with logit averaging, likely because its defining characteristic (insufficient context) is more lexically identifiable than subtle tier distinctions. Figure 17 further illustrates the per-tier improvement from zero-shot to fine-tuning. These patterns suggest that class frequency alone does not determine learnability, and that clinical ambiguity at tier boundaries is a more significant bottleneck than data imbalance.

Ordinal Proximity of Predictions: Despite modest Macro F1 scores, models show meaningful ordinal awareness after fine-tuning. Qwen3-1.7B with logit averaging achieves an MAE of 0.9428, the only configuration to fall below 1.0 (Table 3). This indicates predictions are on average less than one DMRS tier away from the true label. Llama-3.2-1B with logit averaging follows closely with MAE = 1.1695. In contrast, zero-shot models show substantially higher MAE, with Llama-3.2-1B-Instruct reaching 3.4597, indicating near-random tier assignment. This gap confirms that fine-tuning not only improves exact classification but also brings predictions significantly closer to the correct tier on the ordinal scale.

Fine-Tuning Rescues Collapsed Zero-Shot Models: Llama-3.2-1B-Instruct nearly collapses in zero-shot evaluation, achieving only 6.93% Macro F1, effectively predicting low-tier labels for almost all inputs. Similarly, Gemma3-1B-it achieves only 12.10% Macro F1 in zero-shot, with L2, L4, L5, and L8 all at zero F1. Despite these failures, fine-tuning brings Llama-3.2-1B to 32.63% Macro F1 (second best overall) and Gemma3-1B-it to 29.46%, representing improvements of $4.7\times$ and $2.43\times$ respectively. Qwen3-1.7B, which already performs best in zero-shot among the fine-tuned models (F1=14.57%), also benefits the most in absolute terms, reaching 33.47% after fine-tuning. These results demonstrate that zero-shot performance is a poor predictor of fine-tuning potential, and that

even severely collapsed models can be effectively adapted with task-specific training. Tables 2 and 3 show the detailed results.

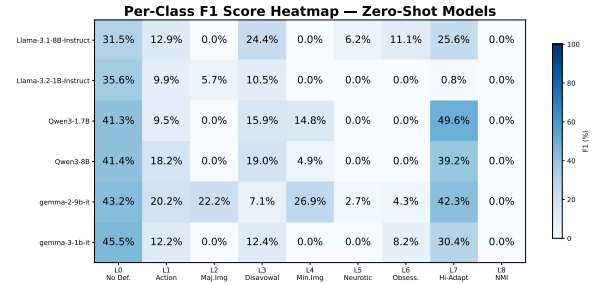


Figure 1: Per-class F1 heatmap for all six zero-shot models. L2 (Major Image-Distorting), L5 (Neurotic), and L6 (Obsessional) show near-zero F1 across all models, reflecting both low support and high clinical ambiguity.

Ensemble Effect: Logit averaging consistently outperforms majority vote and best single fold across fine-tuned models. For Qwen3-1.7B, logit averaging achieves F1=34.59% vs. 32.29% for majority vote and 33.47% for best single fold, confirming that aggregating probability distributions across folds is more effective than hard-label voting. Llama-3.2-1B also showed competitive performance by reaching 34.21% Macro F1, with similar results for Majority Voting. On the other hand, Gemma performed very inconsistent with its best fold performance (29.46% F1) being better than ensembling. Table 3 shows detailed results.

5 Conclusion

We presented a systematic study of small language models for psychological defense mechanism classification under the DMRS framework. Across six models in zero-shot and fine-tuning settings, we find that task-specific fine-tuning combined with logit averaging ensemble is the dominant factor in performance, reaching 34.59% Macro F1 (Qwen3-1.7B), nearly doubling the best zero-shot result of 18.79% (Gemma2-9B-IT). Model scale does not consistently improve zero-shot performance, suggesting that the task’s clinical specificity limits the benefit of additional parameters. Certain defense tiers, particularly L2, L5, and L6, remain persistently difficult across all settings, pointing to clinical ambiguity at tier boundaries as a fundamental challenge beyond data imbalance.

Limitations

Due to GPU constraints, fine-tuning is restricted to models at or below 1.7B parameters. Ensemble strategies are limited to fold-level aggregation within each model, and all evaluations use a single dataset, leaving generalization untested. A broader exploration of prompt strategies, including few-shot and chain-of-thought prompting, remains for future work.

References

- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- 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.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Evdokia Voultsiou and Lefteris Moussiades. 2026. [A systematic review of large language models in mental health: Opportunities, challenges, and future directions](#). *Electronics*, 15(3).
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Appendix

Variant A: Organizer-Provided Baseline Prompt

You are a Defense Mechanism Rating Scale (DMRS) specialist. Examine the dialogue carefully and select the single most appropriate defense tier. When multiple defenses seem plausible, choose the tier with the strongest supporting evidence; if evidence is weak or contradictory, default to '0' (No defense).

Dialogue context: {conversation}
Target utterance: {current_text}

Labels:

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

Return the label digit (0-8) ONLY. No additional content is allowed.

Variant B: Clinically-Grounded Prompt

You are a clinician trained in the Defense Mechanism Rating Scale (DMRS; Perry, 1990). Analyze the dialogue and identify the defense mechanism in the target utterance based strictly on observable verbal behavior.

Dialogue context: {conversation}
Target utterance: {current_text}

- 0 = No defense: The speaker communicates directly with no defensive distortion.
- 1 = Action: Acts on impulse, expresses hostility indirectly, or repeatedly seeks yet rejects help.
- 2 = Major Image-distorting: Rigidly splits others into all-good/all-bad, or projects own feelings onto others.
- 3 = Disavowal: Denies reality, externalizes blame, justifies behavior with plausible logic, or retreats into fantasy.
- 4 = Minor Image-distorting: Subtly belittles self/others, idealizes unrealistically, or expresses special invulnerability.
- 5 = Neurotic: Redirects emotion onto safer target, shows emotional blankness, expresses opposite of what is felt.
- 6 = Obsessional: Detaches via abstract reasoning, focuses on trivial details to avoid affect, or attempts to undo thoughts.
- 7 = Highly Adaptive: Seeks support, acts for others, plans ahead, uses humor constructively, asserts needs calmly.
- 8 = Need More Information: Defense is suspected but insufficient evidence to confirm any tier.

Return the digit (0-8) ONLY. No additional content is allowed.

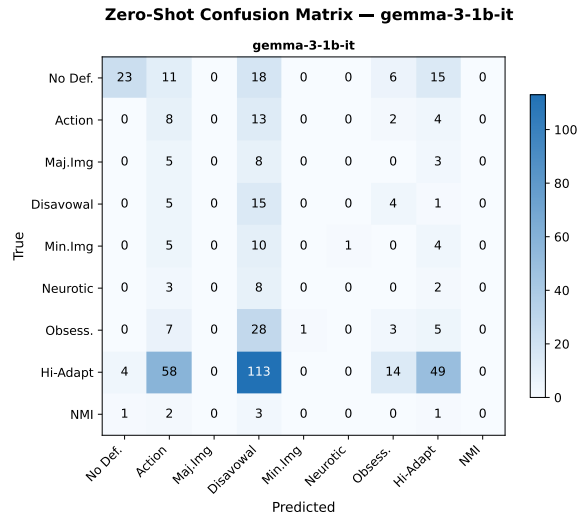


Figure 2: Confusion matrix for Gemma3-1B-IT (Zero-shot).

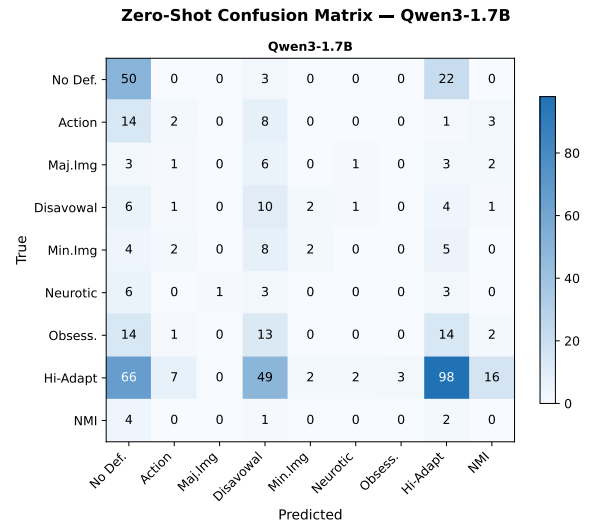


Figure 4: Confusion matrix for Qwen3-1.7B (Zero-shot).

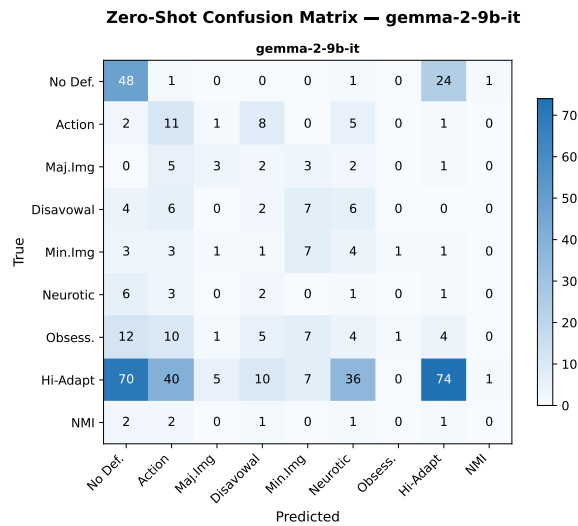


Figure 3: Confusion matrix for Gemma2-9B-IT (Zero-shot).

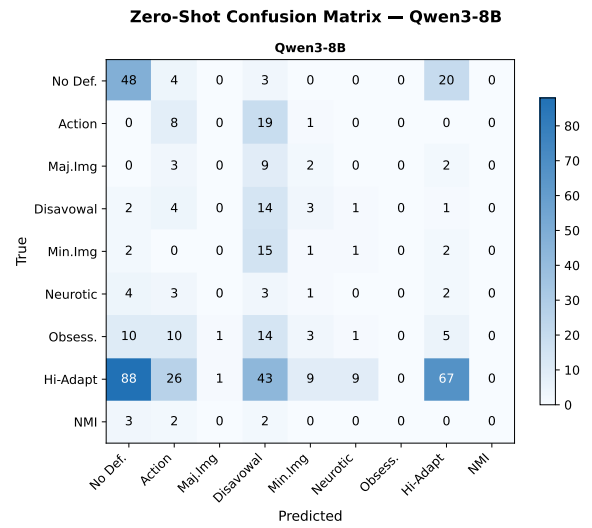


Figure 5: Confusion matrix for Qwen3-8B (Zero-shot).

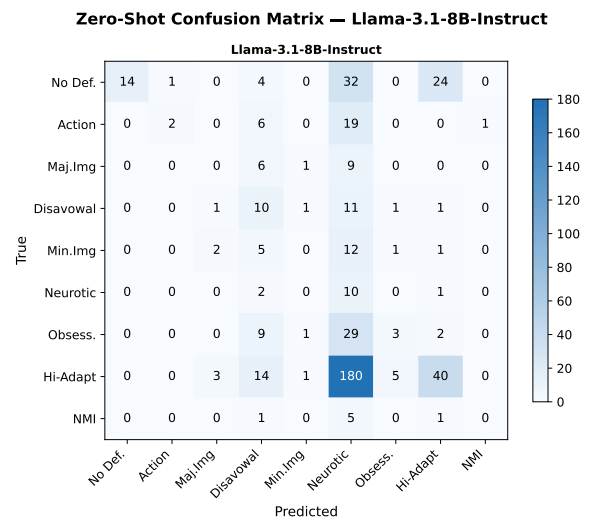


Figure 6: Confusion matrix for Llama-3.1-8B-Instruct (Zero-shot).

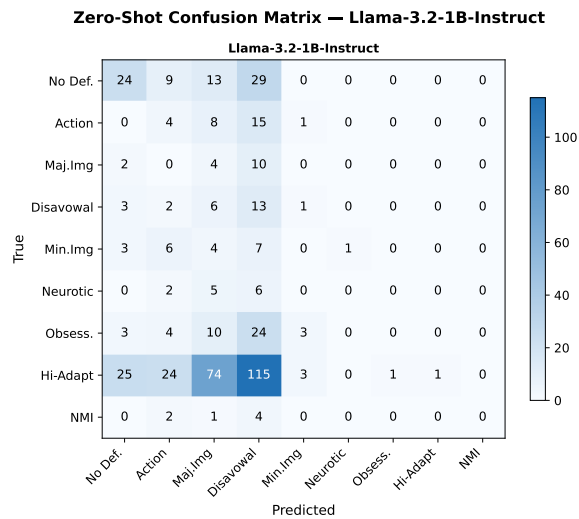


Figure 7: Confusion matrix for Llama-3.2-1B-Instruct (Zeroshot)

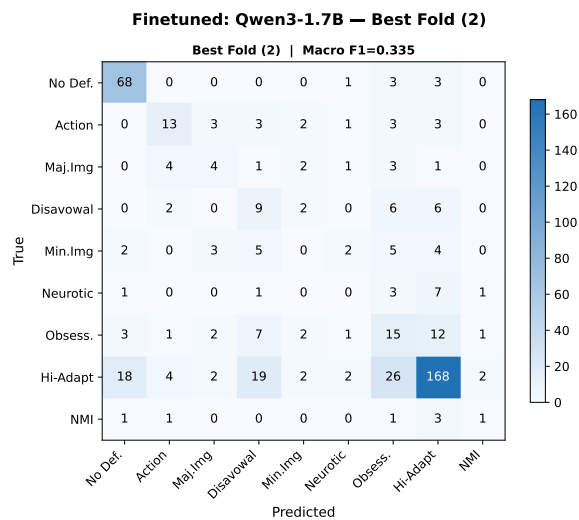


Figure 8: Confusion matrix for Qwen3-1.7B (Best Fold).

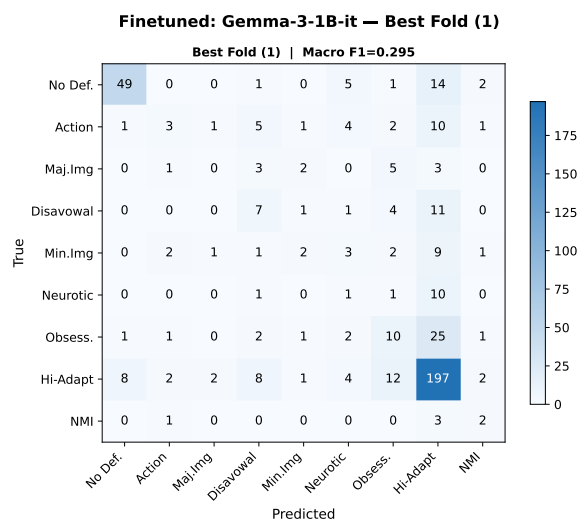


Figure 9: Confusion matrix Gemma-3-1B-it (Best Fold).

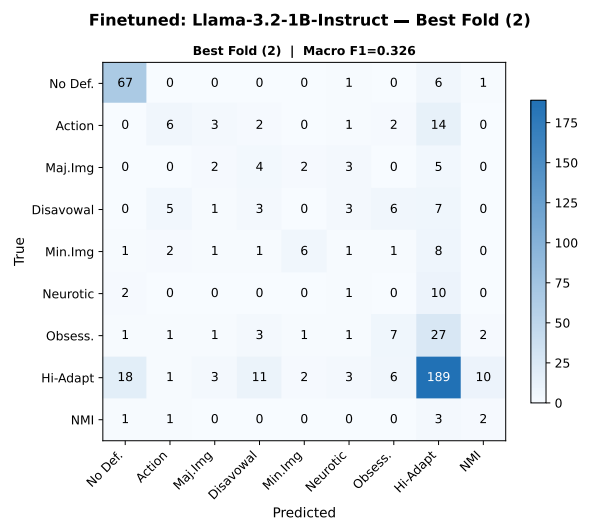


Figure 10: Confusion matrix for Llama-3.2-1B-Instruct (Best Fold).

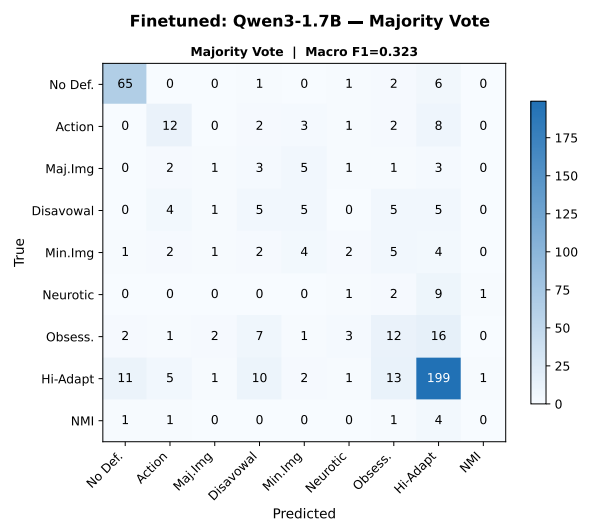


Figure 11: Confusion matrix for Qwen3-1.7B (Ensembling - Majority Voting).

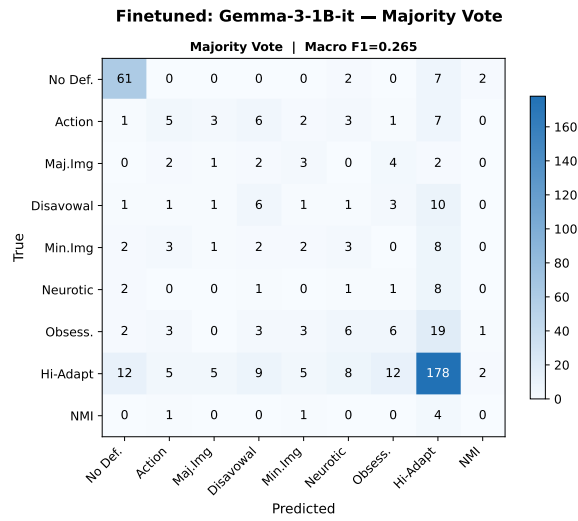


Figure 12: Confusion matrix for Gemma-3-1B-it (Ensembling - Majority Voting).

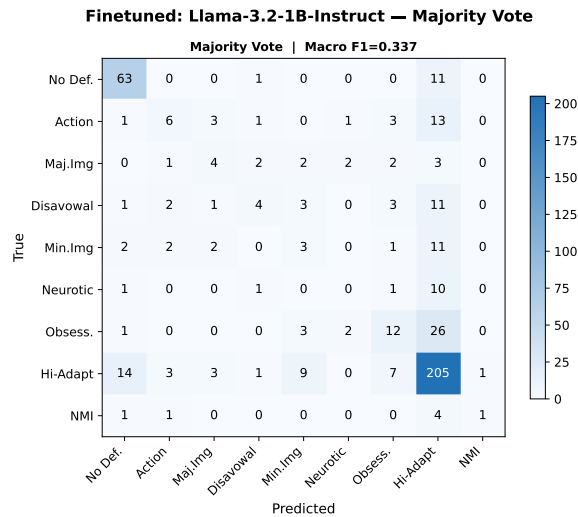


Figure 13: Confusion matrix for Llama-3.2-1B-Instruct (Ensembling - Majority Voting).

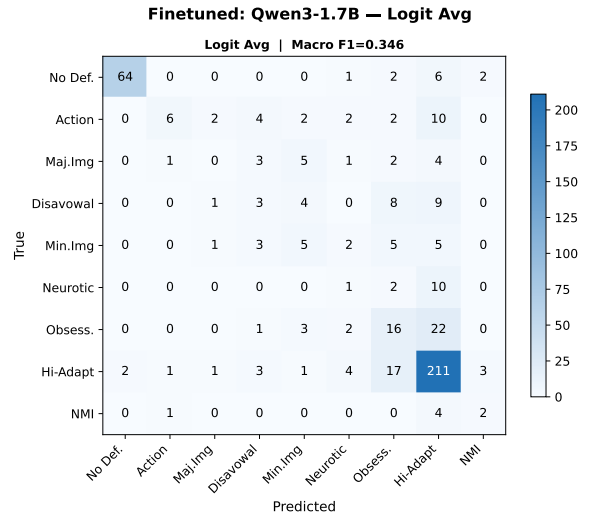


Figure 14: Confusion matrix for Qwen3-1.7B (Ensembling - Logit Averaging).

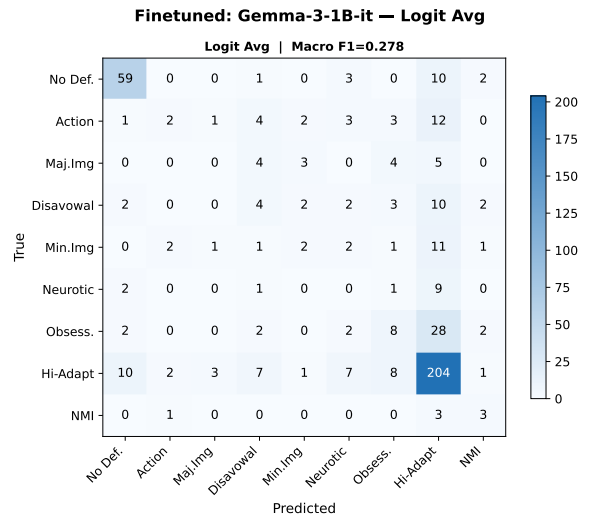


Figure 15: Confusion matrix for Gemma-3-1B-it (Ensembling - Logit Averaging).

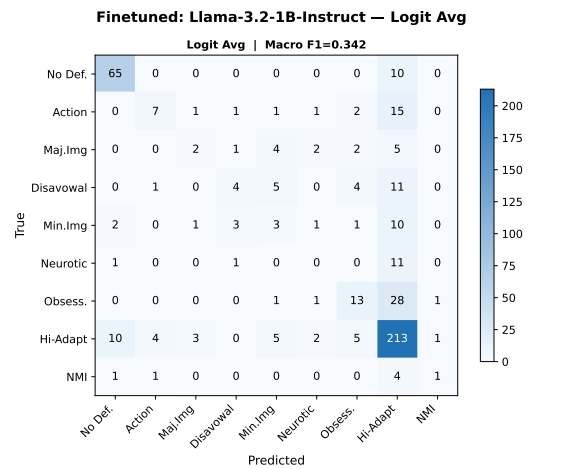


Figure 16: Confusion matrix for Llama-3.2-1B-Instruct (Ensembling - Logit Averaging).

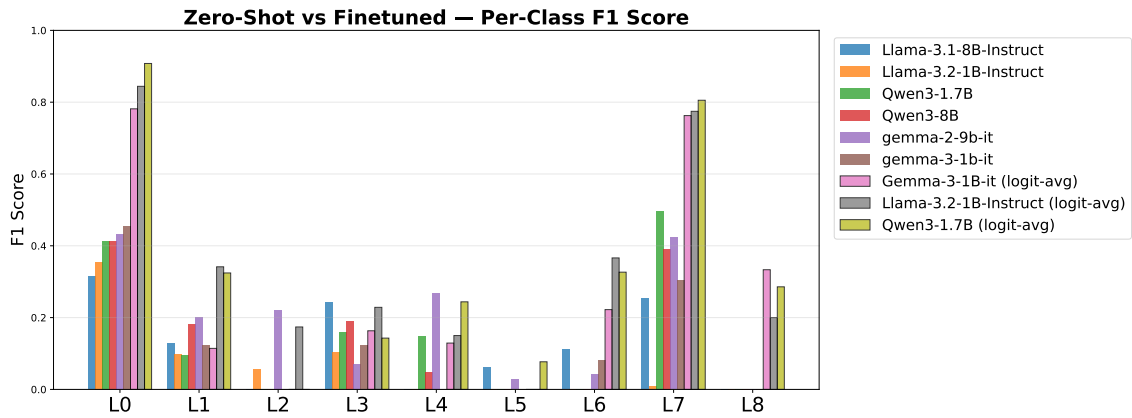


Figure 17: Per-class F1 comparison between the best zero-shot model (Gemma2-9B-IT) and the best fine-tuned model (Qwen3-1.7B, logit averaging). Fine-tuning improves most tiers, though L2, L5, and L6 remain difficult.

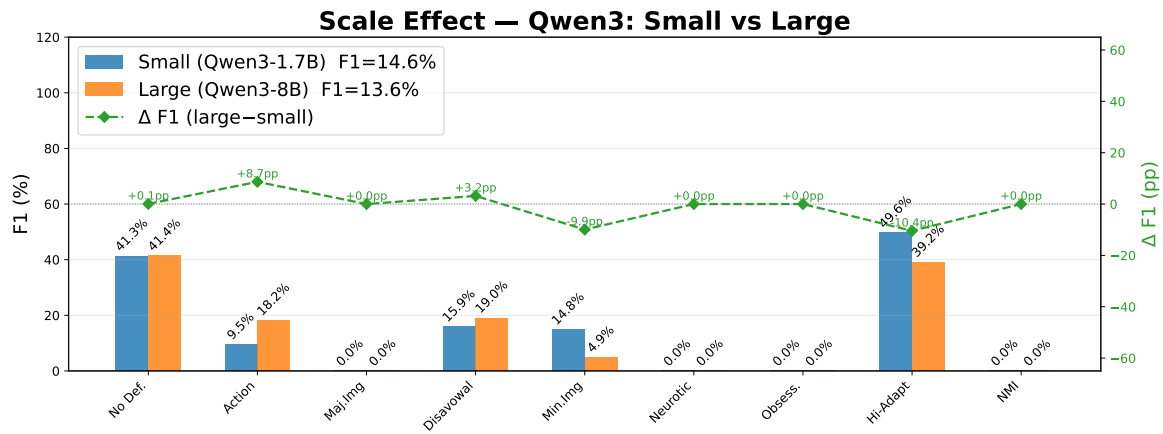


Figure 18: Per-class F1 comparison between Qwen3-1.7B and Qwen3-8B in zero-shot. The larger model underperforms on several tiers.

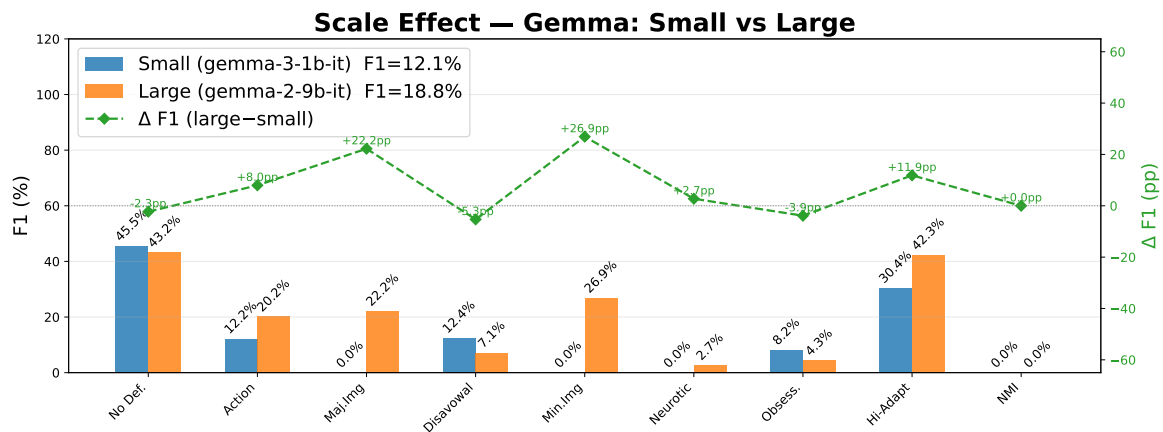


Figure 19: Per-class F1 comparison between Gemma3-1B-IT and Gemma2-9B-IT in zero-shot.

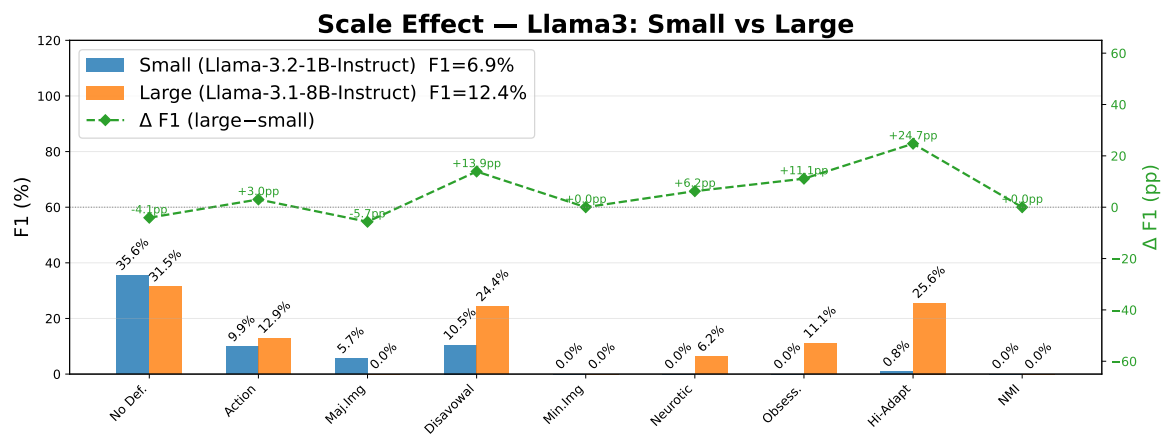


Figure 20: Per-class F1 comparison between Llama-3.2-1B-Instruct and Llama-3.1-8B-Instruct in zero-shot.