UTS at PsyDefDetect: Multi-Agent Councils and Absence-Based Reasoning for Defense Mechanism Classification

Dima Galat and **Marian-Andrei Rizoiu**
University of Technology Sydney

Abstract

This paper describes our system for classifying psychological defense mechanisms in emotional support dialogues using the Defense Mechanism Rating Scales (DMRS), placing second (F1 0.406) among 64 teams.[1] A central insight is that defense mechanisms are defined by what is *absent*: missing affect, blocked cognition, denied reality. We encode this as an *affect-cognition integration spectrum* in prompt-level clinical rules, which account for the largest single gain (+11.4pp F1).

Our architecture is a multi-phase *deliberative* council of Gemini 2.5 agents where class-specific advocates rate evidence strength rather than voting, achieving F1 0.382 with no fine-tuning—a top-5 result on its own. We find, however, that the council is *confidently wrong* about minority classes: 59–80% of stable minority predictions are incorrect, driven by a systematic "L7 attractor" in which emotional content defaults to the majority class. A targeted override ensemble from three fine-tuned Qwen3.5 models applies 16 overrides (+2.4pp), selected by a structured multi-agent system (builder, critic, regression guard) that produced a larger F1 gain in one iteration than 8 prior attempts combined.

1 Introduction

The BioNLP 2026 shared task (Na et al., 2026a,b) requires classifying target utterances in emotional support dialogues into 9 levels of the Defense Mechanism Rating Scales (DMRS; Perry 1990), a hierarchical clinical instrument ranging from action-based defenses (Level 1) to highly adaptive coping (Level 7). The difficulty is that mechanisms are defined by psychological *function*, not linguistic form: the same surface expression can indicate denial (Level 3), intellectualization (Level 6), or adaptive coping (Level 7) depending on context (illustrated in §2).

[1]Code available at https://github.com/dimagalat/bionlp2026

The task exhibits severe class imbalance (Table 1): Level 7 (Highly Adaptive) comprises 51.9% of training data, while Level 5 and Level 8 account for only 2.6% and 1.5%. We call this the *L7 attractor effect*: LLMs over-predict the majority class because emotional engagement in therapeutic dialogue looks like "adaptive coping."

Our system uses a multi-phase deliberative council (§3.1), built on the Gemini 2.5 API (Gemini Team, Google, 2023): three specialist agents classify in parallel, class-specific advocates rate evidence strength, and a resolution stage adjudicates. We frame the DMRS hierarchy as an affect-cognition integration spectrum: many defense mechanisms are defined by what is absent (missing affect, blocked cognition, denied reality), which requires reasoning about what should be present but is not. A targeted override ensemble from three fine-tuned models applies 16 overrides to the council's predictions, achieving macro-F1 0.406 (2nd out of 21 registered teams, or 64 CodaBench entries). We also flag a retrieval-leakage risk: same-dialogue exemplars in few-shot prompts inflate validation accuracy from 65% to 97.7% (§3.2).

2 Task and Data

The dataset contains 1,864 training samples from 200 dialogues and 472 test samples from 189 dialogues. All 189 test dialogue IDs overlap with training dialogues (different utterances from the same conversations), creating a retrieval leakage risk addressed in §3.2. The official metric is macro-averaged F1; we write "F1" throughout to mean macro-F1.

The core challenge is illustrated by this training example: a speaker responds to "How are you today?" with *"I'm OK. Just dealing with a lot of unknowns."* This reads like Level 7 (*Suppression*, consciously managing distress). The ground

Proceedings of the BioNLP 2026 (Shared Tasks), pages 38–46
July 3-4, 2026 ©2026 Association for Computational Linguistics

Level	Name	Train	%
0	No Defense / Neutral	296	15.9
1	Action Defense	108	5.8
2	Major Image-Distorting	61	3.3
3	Disavowal	99	5.3
4	Minor Image-Distorting	84	4.5
5	Neurotic Defense	48	2.6
6	Obsessional Defense	172	9.2
7	Highly Adaptive	968	51.9
8	Needs More Information	28	1.5

Table 1: DMRS class distribution in training data. Level 7 accounts for over half of all samples; the five lower-level defense classes (L1–L5) together comprise only 21.5%.

truth is Level 6 (*Isolation of Affect*): the speaker acknowledges difficulty cognitively ("a lot of unknowns") but the expected emotional response is absent ("I'm OK"). The distinction turns on what is absent from the utterance, not what is present.

3 System Architecture

3.1 Multi-Phase Deliberative Council

Most LLM ensembles aggregate votes. A *deliberative* council instead evaluates *evidence strength per candidate* through structured advocacy. In our architecture, Phase 1 agents propose candidates with alternatives (not just top-1 predictions). Phase 2 spawns class-specific advocates that rate fit as STRONG, MODERATE, or WEAK; each argues *for* its assigned class using retrieved exemplars. Phase 3 resolves via evidence quality (unique STRONG wins immediately; ties require pairwise comparison), not vote count. Majority voting loses minority classes because L7 always outnumbers them; evidence-based resolution can select a minority class that receives STRONG even when the majority favors L7.

Figure 1 illustrates the pipeline, which uses the Gemini 2.5 API (primarily Flash, with Pro for resolution) and requires 3–10 LLM calls per sample depending on consensus.

Formally, given a dialogue d with target utterance u and label space $\mathcal{Y} = \{0, \ldots, 8\}$, the council proceeds in three stages. Three agents a_1, a_2, a_3 each produce a candidate set $C_i = \{(y_i, y_i', p_i)\}$ (primary label, alternative, confidence). Let $\mathcal{C} = \bigcup_i \{y_i, y_i'\}$ be the candidate pool. If all primaries agree ($y_1 = y_2 = y_3 = y^*$) with $\sum_i \mathbf{1}[p_i \geq \tau] \geq 2$, the council returns y^* immediately. Otherwise, for each unique candidate $c \in \mathcal{C}$, a class-specific advocate A_c produces a strength rating

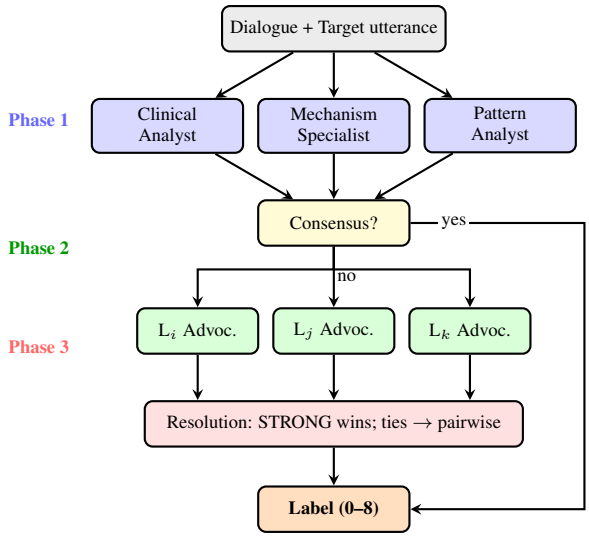

Figure 1: Council pipeline (3–10 LLM calls per sample). Phase 1: three specialist agents classify in parallel. If unanimous with high confidence, the pipeline exits early (3 calls). Otherwise, Phase 2 spawns class-specific advocates rating fit as STRONG/MODERATE/WEAK (2–5 calls). Phase 3 resolves via priority hierarchy (0–1 call).

$s(c) \in \{\text{STRONG}, \text{MODERATE}, \text{WEAK}\}$. A resolution function R selects the final label via priority ordering: unique STRONG wins; ties are resolved by pairwise head-to-head comparison. Phase details are in Appendix A.

Clinical Knowledge Encoding. We encode the DMRS hierarchy (Perry, 1990) as an *affect-cognition integration spectrum*. The discriminative question at each level is: what is the relationship between what the speaker knows and what they feel? Cognition present but affect drained → L6 (Obsessional); affect present but cognition blocked → L5 (Neurotic); cognition distorts reality to manage affect → L2–4; affect and cognition integrated → L7 (Adaptive). The most impactful single test is **Reporting vs. Processing** (L6/L7): describing painful facts without proportional emotion is Isolation of Affect (L6), not adaptive coping (L7). We complement this with five prompt-level rules: (1) 60+ DMRS-Q behavioral indicators per mechanism (Perry, 1990); (2) *Emotion ≠ Defense*: "I feel sad" is not a defense without distortion or transformation; (3) prefer lower-level (less mature) defenses when ambiguous; (4) a watchlist of 8 high-confusion class pairs with discriminative tests; and (5) an L7 verification gate requiring a named adaptive mechanism before permitting L7. In ablation, the Gemini 2.5 Pro council without these clinical rules achieves F1 0.268 (Table 4); adding them

raises F1 to 0.382 (+11.4pp).

Phase-Level Bottleneck. Even with clinical rules, the council's residual errors concentrate at *resolution* rather than detection: the correct label enters Phase 2 as a candidate in 96% of errors and receives a MODERATE+ rating in 76%, but in 94% of errors a wrong label (typically L7) also receives STRONG. The minority-class signal exists; what is missing is a way to prevent L7 from winning the head-to-head, which motivates the override ensemble (§3.3).

3.2 Retrieval and Fine-Tuned Models

Few-shot examples are retrieved via TF-IDF with MMR (Carbonell and Goldstein, 1998) for Phase 1 diversity and semantic re-ranking for Phase 2 within-class exemplars (Lewis et al., 2020). Dialogue-ID exclusion prevents same-conversation leakage; without it, council validation accuracy inflates from 65% to 97.7%.

We train three models via LoRA (Hu et al., 2022) in 4-bit quantization (Dettmers et al., 2023) using Unsloth (Unsloth AI, 2024) and TRL (von Werra et al., 2024): Qwen3.5-9B (65.1% val acc, strongest on L6/L1), Qwen3.5-35B-A3B MoE (61.7%, strongest on L2/L3), and Qwen3.5-9B f1_boost (62.5%, strongest on L1/L2) (Qwen Team, 2026). All use **completion-only loss** (`train_on_responses_only`): training on the full sequence wastes >99% of gradient updates on dialogue auto-completion, and this single change improves accuracy from 25–55% to 59–65%. Self-consistency inference (Wang et al., 2023) (multiple runs at temperature=0.3) provides per-sample confidence scores. A separate pairwise L6/L7 resolver (Qwen2.5-7B, 97.8% val accuracy) handles the dominant confusion pair. Per-model details are in Appendix B.

3.3 Ensemble Strategy

The ensemble applies minimal, high-confidence corrections to the council's predictions (Figure 2).

Type A (L7→minority, high risk): we override when a fine-tuned model achieves ≥80% self-consistency for a minority class and a council rerun or pairwise resolver corroborates. If wrong, we lose a true L7 and add a false minority, a double penalty. **Type B** (minority→minority, lower risk): requires ≥3 of 6 independent sources to agree, with FOR > AGAINST. A credibility gate discards models with <15% val recall on the target class. Our final submission applies 7 Type A and 9 Type B corrections (3.4% of predictions).

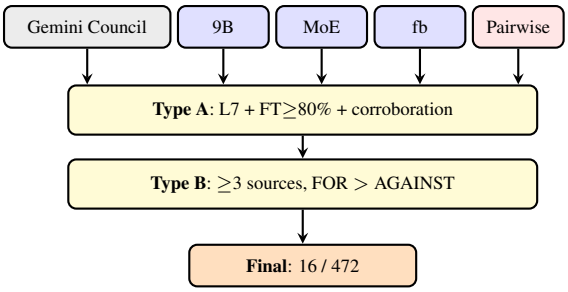

Figure 2: Override ensemble. The council's predictions are checked against three fine-tuned models (20 self-consistency runs each) and a pairwise L6/L7 resolver. Type A corrects L7 over-predictions (7 overrides); Type B corrects minority confusions (9 overrides).

Agentic Override Selection. The override search space is combinatorial: 472 × 8 = 3,776 candidates. For the final submission, we decomposed override selection into three formal roles: (1) parallel *builder agents* scanning every sample against all sources with credibility gates; (2) an independent *critic agent* verifying every claim with separate data access; and (3) a programmatic *regression guard* hard-rejecting submissions below the evidence threshold. This structured approach found 5 corrections that ad-hoc exploration had missed, pushing F1 from 0.393 to 0.406, a larger gain than the preceding 8 iterations combined. The lesson: the value is in formal role decomposition, not automation; agents all the way down, but with structure at every level.

Propose–Verify–Decide. Both our council (§3.1) and the override system follow a three-stage *propose–verify–decide* pattern (Figure 3), differing in the verifier's stance: council advocates argue FOR each candidate class while the override critic argues AGAINST every proposal, and only candidates surviving its scrutiny pass the programmatic regression guard. This mirrors the distinction between the generator–verifier paradigm (Cobbe et al., 2021) and adversarial debate (Irving et al., 2018). Separating proposer and verifier into distinct agents with independent data access prevents the confirmation bias of single-agent self-refine loops (Madaan et al., 2023).

4 Key Findings

L7 Attractor Effect. The council's L7 advocate rates STRONG 73% of the time versus 32–43% for

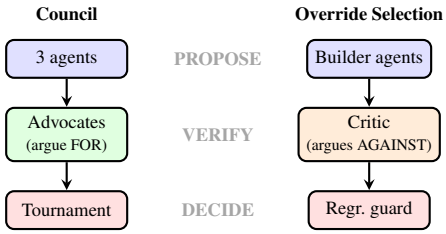

Figure 3: Both multi-agent systems follow a propose–verify–decide pattern. The council's verifiers are advocates (each argues for one class); the override system's verifier is an adversary (argues against all candidates).

True → Predicted	Count	% Errors
L6 → L7	66	16.8
L0 → L7	37	9.4
L4 → L7	34	8.7
L3 → L7	32	8.2
L7 → L0	27	6.9
L1 → L3	18	4.6
Any → L7	*213*	*54.3*

Table 2: Top 6 stable error confusions across 3 council runs (392 total errors). Over half are incorrect L7 predictions; L6→L7 alone accounts for 17% of all errors.

other classes, causing 54% of stable errors to be incorrect L7 predictions (Table 2). The effect intensifies with dialogue length (15.0 turns for errors vs. 12.6 for correct) and minority class accuracy decays sharply with position: L1 drops from 100% (early turns) to 16% (late turns).

Confidently Wrong. Three identical council runs show 22.2% prediction instability, driven by asynchronous execution order in the multi-agent architecture even at temperature=0. For minority classes L3, L4, L6, and L8, unstable predictions are paradoxically more accurate than stable ones (Table 7 in App. C); between 59% and 80% of stable minority predictions are wrong.

Prompt Overfitting. Every prompt modification that improved training-set F1 degraded test performance (Table 3). We hypothesize this reflects prompt overfitting: iteratively tuning prompts against training-set metrics acts as implicit gradient descent with no generalization check.

Additional findings. TF-IDF with MMR retrieves 38% L7 examples versus 42–46% for all semantic variants, because its lack of semantic understanding prevents emotional-content clustering, a useful property for diversity-dependent classification. Completion-only loss is the most impactful fine-tuning intervention: it improves accuracy from 25–55% to 59–65%.

Modification	Δ Train	Δ Test
L7 advocate rigor	+1.2pp	−3.6pp
Dialogue position metadata	+0.8pp	−2.1pp
Training exemplars in prompts	+1.5pp	−3.6pp
Few-shot k=3 → 5	+1.5pp	−1.4pp
Length bias warning	+0.5pp	−2.0pp

Table 3: The overfitting paradox: modifications improving training F1 consistently degrade test F1.

System	Acc	P	R	F1
Council (no clin. rules)	.653	.395	.249	.268
Council baseline	.663	.473	.364	.382
+ 7 Type A overrides	.670	.477	.377	.391
+ 9 Type B overrides	**.674**	**.461**	**.388**	**.406**

Table 4: Test set results. Clinical knowledge rules account for +11.4pp F1 over the unconfigured council; the override ensemble adds +2.4pp through 16 overrides.

Override Selection Lessons. Beyond architecture, our 9 submissions yielded replicable patterns for ensemble correction on imbalanced classification, four of which are not derivable from individual model accuracy. **Prediction volume signals per-sample reliability.** Under-predictors (9B predicts L2 for 11 samples vs. ∼16 expected) carry higher precision; over-predictors (f1_boost: 31 vs. ∼16) inflate false positives, so we weight votes by this volume discount. **Architectural independence is necessary but not sufficient.** A Gemma4-26B-A4B (Gemma Team, Google DeepMind, 2026) agreeing with the Qwen3.5 ensemble only 69–74% of the time (the most independent source we trained) never improved test F1, because per-class accuracy was too low for its disagreements to carry signal. **Override count vs. F1 is sharply non-monotonic.** 75 overrides → F1 .367; 16 → .406; 21 → .399; the peak lives at a narrow intersection of evidence quality and quantity. **Regression guards must be programmatic.** Advisory warnings ("this override has weak evidence") were ignored by human operators and LLM agents alike; only hard-rejecting submissions failing structural checks prevented regressions.

5 Results

Table 4 shows the ablated contribution of each component. Clinical rules account for the largest gain (+11.4pp). The 7 Type A overrides add +0.9pp F1; the 9 Type B overrides add +1.5pp (recall domi-

Team	Acc	P	R	F1
Nürnberg NLP	**.701**	.451	**.404**	**.420**
UTS (ours)	.674	**.461**	.388	.406
PerceptionLab	.674	.426	.409	.396
LinguIUTics	.642	.400	.396	.392
LDI Lab	.636	.377	.389	.371

Table 5: Top 5 on the shared task leaderboard (21 registered teams). Our system has the highest precision.

nates despite −1.6pp precision). On the shared task leaderboard (Table 5), our system achieves the highest precision among all teams (0.461), reflecting the conservative override strategy. The development progression (Table 8 in Appendix D) shows that the council's minority predictions are ∼47% correct, so only overrides backed by overwhelming evidence (≥80% FT confidence, ≥3 independent sources, zero opposition) reliably improve F1. Post-hoc analysis with the released test labels: 9 of 16 overrides individually corrected council errors (4 Type A, 5 Type B), 4 regressed correct predictions, and 3 were lateral, yielding the +5 net correct predictions behind the +2.4pp F1 gain. The 56% override precision sits just above the 47% minority baseline, validating the conservative gating threshold.

6 Related Work

Our deliberative council builds on multi-agent debate (Du et al., 2023) and specialized medical prompting (Nori et al., 2023). The DMRS framework (Perry, 1990) provides the theoretical foundation; the shared task dataset (Na et al., 2026b) enables the first large-scale computational study of defense mechanisms in naturalistic dialogue, and a recent survey (Na et al., 2025) situates this within the broader LLM-psychotherapy landscape. Prior computational approaches to defense mechanisms have been limited to rule-based systems on structured clinical notes; our work is among the first to apply LLMs to this task. Our override framework relates to the generator–verifier paradigm (Cobbe et al., 2021) and adversarial debate (Irving et al., 2018); we adapt self-consistency (Wang et al., 2023) for classification confidence and find that TF-IDF diversity (Carbonell and Goldstein, 1998) outperforms semantic retrieval (Lewis et al., 2020) for few-shot selection under class imbalance.

7 Limitations

Our system depends on the Gemini API, limiting reproducibility to researchers with equivalent access. The validation set is a single GroupKFold split (373 samples); cross-validation was infeasible given API costs. All experiments use English emotional support dialogues from a single cultural context; generalization to other languages or therapeutic traditions is untested. The affect-cognition spectrum is our operationalization of clinical theory (Perry, 1990), not a validated clinical instrument. We address absence-as-signal heuristically; explicit counterfactual reasoning remains open.

8 Conclusion

Our council-ensemble system achieves 2nd place on DMRS defense mechanism classification. The core finding is that defense mechanisms are defined by what is absent, and encoding this insight in clinical rules produces the largest single gain (+11.4pp), more than any model or architectural choice. Formalizing override selection into builder–critic–guard roles then added +2.4pp in a single submission, more than 8 prior iterations combined. Future work could explore hierarchical classification, contrastive training from council error logs, and zero-shot classification via frontier embedding models.

Ethics Statement

This system classifies psychological defense mechanisms (constructs describing internal psychological states) and is a *research tool*, not a clinical diagnostic instrument. It should not be used to label individuals without clinical oversight, as misclassification could distort therapeutic understanding. The training data comes from the shared task organizers (Na et al., 2026a) who ensured appropriate consent and anonymization. Following Strubell et al. (2019), the system's total compute footprint is ∼22 kWh (council API at ∼$75; fine-tuning 13 kWh; self-consistency inference 1.2 kWh; PUE 1.2), corresponding to ∼8.5 kg CO_2eq on the US grid (4.3 kg renewable).

References

Anthropic. 2026. Claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering doc-

uments and producing summaries. In *Proceedings of SIGIR*, pages 335–336.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized language models. *arXiv preprint arXiv:2305.14314*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Google DeepMind. 2026. Gemma 4. https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shanen Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899*.

Chankyu Lee, Rajarshi Roy, Menber Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Han. 2024. NV-Embed: Improved techniques for training LLMs as generalist embedding and retrieval models. *arXiv preprint arXiv:2405.17428*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. In *Proceedings of NeurIPS*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of EACL*, pages 2014–2037.

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.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI. 2024. The GPT model family. *OpenAI Technical Reports*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

J. Christopher Perry. 1990. Defense Mechanism Rating Scales. Technical report, Cambridge Hospital, Harvard Medical School.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of ACL*, pages 3645–3650.

Unsloth AI. 2024. Unsloth: Fast language model finetuning. https://github.com/unslothai/unsloth.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, and Nathan Lambert. 2024. TRL: Transformer reinforcement learning. https://github.com/huggingface/trl.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of ICLR*.

A Council Phase Details

Phase 1: Initial Assessment (3 parallel calls). Three specialist agents independently classify each utterance: (a) a *Clinical Analyst* applying psychodynamic reasoning (stressor → function → mechanism → level), (b) a *Mechanism Specialist* screening all 9 levels using DMRS-Q behavioral indicators (Perry, 1990), and (c) a *Pattern Analyst* performing analogical reasoning from TF-IDF-retrieved few-shot examples (k=3). Each agent outputs a primary label, alternative label, confidence score, and identified mechanism. If all three agents agree on the same label with $\geq$2 having high confidence, the pipeline exits early (3 calls total).

Phase 2: Differential Diagnosis (2–5 calls). For each unique candidate label from Phase 1 (primaries and alternatives), a class-specific advocate evaluates fit as STRONG, MODERATE, or WEAK using class-representative exemplars retrieved via semantic similarity. Advocate criteria are calibrated: STRONG requires "clear, specific evidence; functions similarly to examples"; MODERATE requires "partial or suggestive evidence"; WEAK indicates "little evidence; functions differently." A minority class screening step injects at least one underrepresented class as a candidate. The L7 advocate rates STRONG 73% of the time (vs. 32–43% for other classes), creating the attractor effect.

Phase 3: Smart Resolution (0–1 call). Resolution follows a priority hierarchy: single STRONG advocate → pick immediately (0 calls); multiple STRONG → pairwise head-to-head comparison (1 call); multiple MODERATE → pairwise comparison; single MODERATE → pick; all WEAK → deliberation moderator synthesis. The pairwise resolver compares two candidates by studying class-representative examples for each, then determining which candidate the target utterance is more *functionally* similar to.

B Model Details

Qwen3.5-9B (Qwen Team, 2026) (LoRA r=64, attention + MLP targets): 65.1% val accuracy. Strongest on L6 (43% recall) and L1 (36%).

Qwen3.5-35B-A3B MoE (LoRA r=32, attention-only[2]): 61.7% val accuracy. Strongest on L2 (60% recall) and L3 (32%).

Qwen3.5-9B f1_boost (variant training recipe): 62.5% val accuracy. Strongest on L1 (46%) and L2 (60%).

All models use: (1) **completion-only loss** via `train_on_responses_only` (loss computed only on the label token, not the dialogue); (2) GroupKFold by dialogue_id for zero-leakage validation (Pedregosa et al., 2011); (3) balanced sampling (L7 capped at 300, minorities oversampled to 80).

Self-Consistency Inference (Wang et al., 2023). Each model runs multiple times at temperature=0.3 per test sample; the majority vote serves as the prediction and the agreement fraction as a confidence score. Since our output is a single classification token, logit probabilities from a single forward pass could serve as an alternative confidence measure; we used sampling for implementation convenience.

Pairwise Differential Resolver. A separate Qwen2.5-7B model fine-tuned on 942 pairwise comparison examples serves as an L6-vs-L7 specialist. Given two candidate levels and a dialogue, it determines which better fits the target utterance. Per-pair val accuracy: L6/L7 97.8%, L3/L7 100%, L1/L7 100%, L0/L7 52.3%.

C Detailed Findings

C.1 The L7 Attractor Effect

Beyond the headline numbers in §4, the attractor effect intensifies with utterance verbosity: incorrectly classified samples average 23.0 words versus 17.1 for correct ones, and L5 accuracy decays from 100% (early turns) to 18% (late turns), mirroring the L1 pattern. The model uses utterance elaboration as a proxy for defense sophistication, inverting

[2]MoE expert layers are excluded from LoRA targets due to an Unsloth adapter reload bug. The model leverages pretrained expert routing.

Class	9B	MoE	fb
L1 (Action)	36	18	**46**
L2 (Image-Dist.)	40	**60**	**60**
L3 (Disavowal)	21	**32**	5
L4 (Image-Dist.)	7	**29**	14
L5 (Neurotic)	14	14	**29**
L6 (Obsessional)	**43**	23	23

Table 6: Per-class recall (%) on validation data for each fine-tuned model. Bold indicates the best model for each class. No model dominates; each is the best or only credible source for $\geq$1 class.

Class	Disagree Rate	Stable Acc	Stable Wrong	Unstable Acc
L0	17.9%	84%	16%	36%
L1	40.7%	41%	59%	32%
L2	37.7%	21%	79%	13%
L3	31.3%	32%	68%	**48%**†
L4	29.8%	24%	76%	**40%**†
L5	27.1%	23%	77%	31%
L6	35.5%	25%	75%	**36%**†
L7	15.5%	91%	9%	47%
L8	46.4%	20%	80%	**38%**†

Table 7: Prediction stability across 3 identical council runs. † marks classes where unstable predictions are more accurate than stable ones; the system is confidently wrong on these minority classes.

the clinical truth: long, detailed descriptions of hardship without emotional processing indicate L6, not L7.

C.2 Prediction Instability and the Confident-Wrong Problem

On the full training set, stable predictions achieve 73.0% accuracy versus 39.2% for unstable ones. However, for minority classes L3, L4, L6, and L8, unstable predictions are paradoxically more accurate than stable ones (Table 7). Between 59% and 80% of stable minority predictions are wrong. The system is not uncertain on hard examples; it is confidently incorrect.

C.3 The Overfitting Paradox in Prompt Engineering

Beyond the table in §4, the mechanism appears to be that iteratively tuning prompts against training-set metrics acts as implicit gradient descent with no generalization check, causing the prompt to memorize training distribution artifacts rather than capture the true classification signal.

#	Configuration	Ov.	F1
1	All FT overrides (aggressive)	75	.367
2	L6/L2 only	42	.375
3	Double corroborated	10	.385
4	Triple corroborated	4	.387
5	+ pairwise L6 resolver	7	.391
6	10 overrides variant	10	.391
7	8 overrides variant	8	.391
8	+ 4 minority→minority (Type B)	11	.393
9	**+ 5 more Type B**	**16**	**.406**

Table 8: Development progression across 9 test submissions. All entries are official competition submissions; the final row (#9, F1 .406) was our selected leaderboard entry. The key insight: moving from 75 aggressive overrides (F1 .367) to 16 surgical ones (F1 .406). Fewer, higher-confidence corrections consistently outperform larger override sets.

C.4 Retrieval: TF-IDF Beats Semantic for Diversity

We compared five retrieval strategies for Phase 1 few-shot selection. All four semantic variants (Gemini embedding-001, enriched, task-framed, and focused TF-IDF on last 3 turns) produced 42–46% L7 in retrieved examples. TF-IDF with MMR achieved 38%, well below the 52% base rate, because its lack of semantic understanding prevents emotional-content clustering. For Phase 2 within-class retrieval, however, semantic re-ranking finds better functional matches: embeddings capture functional similarity when the class label constrains the search space. This phase-dependent pattern is worth noting for future work: frontier embedding models trained for zero-shot classification (Lee et al., 2024; Wang et al., 2022; Muennighoff et al., 2023) perform a similar constraint implicitly.

D Development Progression and Failed Approaches

Failed Approaches.

GPT-5.4 (OpenAI, 2024). Standalone F1 of 0.265 with heavy L0 bias (30.5% of predictions). Near-zero L1 detection.

Claude Opus 4.6 agent council (Anthropic, 2026). F1 0.261 without retrieval augmentation, with 69.3% L7 over-prediction.

Aggressive class balancing. Oversampling minorities to 150 (vs. 80) and capping L7 at 200 (vs. 300) caused minority over-prediction, reducing val accuracy from 59.0% to 51.5%.

Chain-of-thought fine-tuning. Template reasoning ("The speaker is managing internal state...") taught the model to parrot templates rather than classify. Removing CoT and training on label-only output was strictly better.

Gemma4 MoE (Gemma Team, Google DeepMind, 2026). Attention-only LoRA on Gemma4-26B-A4B achieved 63.5% val accuracy with the best L4 recall of any model (36%). However, its test-set predictions failed to improve F1: the model's L4 signal was contradicted by all Qwen models, and its L1 predictions were heavily overpredicted (47 vs ∼27 expected). Despite genuine architectural independence from Qwen (69–74% agreement), this independence did not translate to useful override evidence, a cautionary result for cross-architecture ensembling. An earlier attempt applying LoRA to MoE expert layers (not just attention) failed due to an Unsloth adapter serialization bug: the model trained correctly but collapsed to all-L7 when reloaded from checkpoint.