Nürnberg NLP at PsyDefDetect: Multi-Axis Voter Ensembles for Psychological Defence Mechanism Classification

Philipp Steigerwald, Eric Rudolph and Jens Albrecht
Technische Hochschule Nürnberg Georg Simon Ohm
{philipp.steigerwald,eric.rudolph,jens.albrecht}@th-nuernberg.de

Abstract

Detecting levels of psychological defence mechanisms in supportive conversations is inherently ambiguous. In the PsyDefDetect shared task at BioNLP 2026 the eight positive defence categories share surface language and differ only in pragmatic function and trained raters reach only moderate inter-annotator agreement. On such a task the decisive lever is not a stronger single model but error independence, since any single representation will waver on the overlapping defence boundaries. We translate this insight into a 9-voter ensemble spanning three orthogonal axes: class granularity (all nine classes for the gatekeeper, only the eight defence classes for the specialists), training method (generative and discriminative) and base model. The system reaches $F1_{test}$=.420 on the hidden test set, placing first among 21 registered teams.

1 Introduction

The PsyDefDetect shared task (Na et al., 2026a) asks a model to classify each seeker utterance in an emotional-support conversation by its level of psychological defence. The PSYDEFCONV corpus (Na et al., 2026b) pairs ESConv (Liu et al., 2021), a corpus of crowdsourced support dialogues, with the Defense Mechanism Rating Scale (DMRS; Perry, 1990)—a clinical taxonomy of eight hierarchical defence levels plus a "No Defence" category—and is evaluated on macro-F1 over classes 1–8. The task is clinically motivated (Perry, 2014) but difficult: trained raters reach only a moderate Cohen's κ=.639 and the corpus is heavily imbalanced (C7 covers 52%, the three rarest classes together 7.4%). Several defence categories share surface-level language and differ only in pragmatic function, so the semantic boundaries are inherently fuzzy.

Given these fuzzy boundaries, voting across diverse voters was our starting point. Rather than chasing a stronger single model, we sought voters with uncorrelated errors that arbitrate the ambiguity. We tested different training methods (generative and discriminative), several base LLMs and different class granularities. A geometric analysis of the QLoRA-adapted hidden states indicates that only the no-defence class is reliably separable, motivating a *generalist-specialist* split (9-class generalist + 8-class specialists). To counter the heavy imbalance, we additionally augmented the minority defence classes with GPT-5.2 synthetic dialogues.

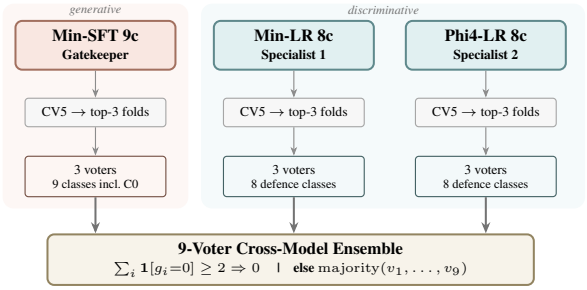

Figure 1: Architecture of our 9-voter cross-model ensemble.

All training uses 5-fold cross-validation (CV5) as both a voter pool and an internal performance estimate ($F1_{cv}$, mean macro-F1 over classes 1–8) since test labels were hidden. Comparing voters across folds, methods and base models, each axis produces systematically different errors: where some voters get confused on the fuzzy defence boundaries, others succeed and majority voting sharpens those boundaries—the error independence (Dietterich, 2000) we sought, realised in a 9-voter cross-model cross-method ensemble (Figure 1) that reaches $F1_{test}$=.420 (+33.4% over the baseline).

Our contributions: (i) the winning 9-voter cross-model cross-method ensemble; and (ii) an embedding-level analysis quantifying the defence-class semantic overlap that drives task difficulty. Additionally, we release our class-imbalance synthetic dialogues for replication.

Proceedings of the BioNLP 2026 (Shared Tasks), pages 59–65

2 System

We build our 9-voter ensemble (Figure 1) step by step, adding one voter voice at a time. Each step posed a design choice—which method, which base model, which folds to trust—which we settled by the CV5 signal, guided by the principle that a diversity of voices sharpens fuzzy class boundaries better than any single strong voice.

2.1 Data Augmentation

The PSYDEFCONV training set is heavily imbalanced (C7 covers 52%, the three rarest classes together 7.4%). Our first step was to replicate the organisers' baseline on a dialog-stratified 80/20 split, and the resulting 1,520-sample training split was augmented with up to $\min(200-n_c,\ 3 \cdot n_c)$ GPT-5.2 synthetic dialogues per class—the first term targets 200, the second caps synthetic at 75%. Classes 0 and 7 are excluded as already well-represented; this yields 738 synthetic dialogues. When we later moved to CV5 for the voter pool, we reused the same 738 synthetic dialogues unchanged across all five folds (per-class counts in Appendix Table 4). Validation and test sets (472 samples) remain original human-annotated data only.

2.2 Voting

On this augmented data, voting was our first step toward the voter diversity we kept extending throughout the system. Given the dataset's fuzzy class boundaries, the decisive lever is error independence between voters rather than a stronger single model; where one voter wavers, another may be more confident on the same sample and majority voting sharpens the joint decision. CV5 provides both a voter pool (five trained models per configuration) and an internal performance estimate ($F1_{cv}$, mean macro-F1 over classes 1–8) since test labels were hidden, with the majority across the five fold-models giving the ensemble prediction. Our first ensemble ran the organisers' baseline approach (Ministral-8B with generative supervised fine-tuning on all 9 classes of the augmented data, Min-SFT 9c) as 5-fold majority voting, lifting $F1_{test}$ from .315 to .373 (Table 1)—already a substantial gain with no architectural diversity yet present.

2.3 Training Axis

We tested three adaptation methods for fine-tuning a base LLM. **SFT** (supervised fine-tuning) fine-tuned the LLM end-to-end with QLoRA on a generative objective, learning to emit the class digit as text. **ClsHead** (classification head) attached a randomly initialised head to the base LLM and jointly fine-tuned both with QLoRA and focal loss. **LR** (logistic regression) froze the ClsHead-adapted LLM, discarded the trained classification head and fitted a new linear head on the frozen last-token hidden states—architecturally identical to the discarded one but retrained as an L2-regularised logistic regression on frozen features rather than jointly with the backbone. Because LR reused the ClsHead-adapted backbone (extraction in minutes, fit in seconds), it added essentially no compute and let us screen many base-model and class-mode combinations cheaply. All three shared the same input prompt (see appendix).

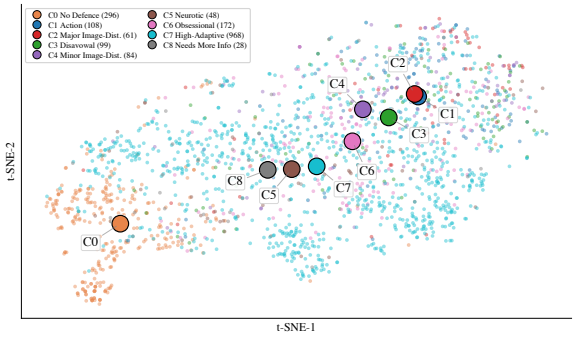

Figure 2: Out-of-fold per-class t-SNE of SFT QLoRA-adapted 9-class Ministral-8B hidden states on the 1,864 original training utterances, with C0 No Defence forming the only well-separated cluster.

2.4 Class Granularity Axis

To strengthen the 5-voter baseline we faced two design questions, which classes are reliably distinguishable in the hidden-state space and which configurations to combine for uncorrelated errors.

A t-SNE of the SFT QLoRA-adapted 9-class Ministral-8B hidden states (Figure 2) indicates that C0 No Defence forms the most separable cluster, while the eight defence classes overlap substantially. This motivates a class granularity split. The *gatekeeper* keeps all nine classes, using C0 predictions for the no-defence override and C1–C8 predictions for the defence vote. The 8-class *specialists* focus entirely on the overlapping defences.

For the 8-class specialist we tested seven base models with ClsHead and LR (Appendix Table 3). LR matched or outperformed ClsHead in most cells and achieved the highest $F1_{cv}$ on the majority of base models, becoming the specialist default.

For the gatekeeper, although LR 9c slightly outperforms SFT 9c on $F1_{cv}$, we deliberately chose generative SFT, expecting two different methods to disagree on a different subset of samples than two LR branches would and trading a small per-voter $F1_{cv}$ loss for a larger gain in ensemble error independence. A post-hoc ablation supports this. Pairing the Min-SFT 9c gatekeeper with another SFT specialist (Min-SFT 8c, top-3 folds, t=2) yields $F1_{test}$=.373, no improvement over the 5V baseline, while pairing with the discriminative Min-LR 8c lifts it to .391 (Table 1).

We paired the Min-SFT 9c gatekeeper with the Min-LR 8c specialist, keeping only the top-3 folds by $F1_{cv}$ per branch to drop each branch's most uncertain folds and save 40% inference cost. The resulting 6-voter ensemble fortuitously covers all five folds (Min-SFT 9c $\{f0, f1, f4\}$, Min-LR 8c $\{f0, f2, f3\}$).

With a dedicated gatekeeper, the simple majority vote extends to a two-stage rule. The gatekeeper voters first decide whether the sample is C0 and the remaining defence classification is settled by majority across all voters. Letting $g_1, \ldots, g_G$ denote the G gatekeeper predictions and $v_1, \ldots, v_V$ all V voter predictions, the ensemble decision is

$$\hat{y} = \begin{cases} 0 \text{ if } \sum_i^G \mathbf{1}[g_i{=}0] \geq (G{+}1)/2 \\ \text{argmax}_c \sum_j^V \mathbf{1}[v_j{=}c] \text{ else} \end{cases} \quad (1)$$

with ties broken in favour of class 7 (the majority class). The gatekeeper voters participate in both branches of Equation 1—triggering the C0-override when a majority of gatekeepers predicts C0, otherwise voting on defence classes alongside the LR specialists and adding method diversity since SFT and LR fail on different subsets of ambiguous samples.

2.5 Model Axis

To extend the diversity principle to the third (model) axis, we tested three additional 8-class LR variants—Phi-4-14B (Phi4-LR 8c), Llama-3.1-8B (Llama-LR 8c) and PsychoCounsel-Llama3-8B (PCounsel-LR 8c, a counselling-domain Llama3-8B finetune)—and ranked them by per-fold Pearson correlation with Min-LR 8c's $F1_{cv}$ profile (negative values indicate anti-aligned per-fold strengths, contributing independent voter signal). We selected Phi4-LR 8c on the most anti-aligned per-fold profile (r=−.544; Llama-LR 8c +.06, PCounsel-LR 8c −.09), completing the three-axis 9-voter ensemble. With only n=5 folds, the gap between Phi4-LR (r=−.544) and Llama-LR (r=+.06) is indicative rather than statistically decisive; consistent with this, all three candidates land within .006 of each other on the test set (Table 1).

System	$F1_{test}$
Baseline (Na et al., 2026b) (Min-SFT 9c, no-aug)	.315
Min-SFT 9c full-train, augmented (single model)	.307
Voting baseline (no axes)	
5V Min-SFT 9c (5 folds)	.373
Class + training axis	
6V Min-SFT 9c + Min-LR 8c	.391
Class + training + model axis	
6V Min-SFT 9c + Phi4-LR 8c	.391
6V Min-SFT 9c + Llama-LR 8c	.392
9V Min-SFT 9c + Min-LR 8c + PCounsel-LR 8c	.414
9V Min-SFT 9c + Min-LR 8c + Llama-LR 8c	.417
9V Min-SFT 9c + Min-LR 8c + Phi4-LR 8c	**.420**

Table 1: Hidden-test-set scores ($F1_{test}$, classes 1–8) for our submitted systems (all trained on augmented data), grouped by which diversity axes are active.

3 Results and Analysis

The exploratory design above, guided by the $F1_{cv}$ signal across methods, base models and folds, produced the configurations in Table 1. The rest of this section asks three questions of the winning 9V—where do its per-class errors concentrate, what does the third specialist arbitrate and how much of the gain depends on augmentation.

3.1 Per-Class Analysis

On the hidden test set (n=472), the winning 9V performs strongly on surface-identifiable defences (C0 No Defence $F1_{test}$=.899, C7 High-Adaptive .833) but struggles where categories overlap semantically (Figure 3, Appendix Table 5). The C0-override (Equation 1) fires on 17.6% of test samples, close to the 15.9% training prevalence.

Two error patterns dominate. C6 and C7 are swapped on 28 samples (16 C6→C7, 12 C7→C6) and 7 of 13 C5 Neurotic samples (54%) are labelled C7 High-Adaptive—the highest relative confusion rate. All three classes produce measured, reflective language and distinguishing them needs intent or longitudinal context rather than a single utterance (Perry, 2014). The model therefore defaults to C7—the clinically costly direction, where a neurotic defence read as mature coping misses the signal for intervention.

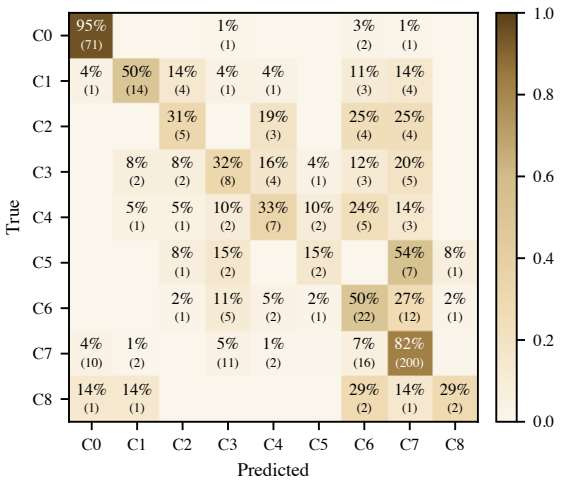

Figure 3: Row-normalised confusion matrix of the winning 9V system on the hidden test set. C7 High-Adaptive absorbs most misclassifications from mid-hierarchy classes.

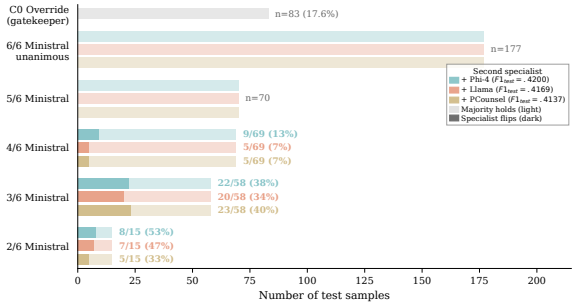

Figure 4: Each bar is one 9V system (6 Ministral + 3 specialist voters); y-axis groups samples by how many Ministral voters agreed—the specialist can flip the Ministral majority only from 4/6 downwards—and dark portions mark the actual flips.

3.2 Voter Diversity: Flipping and Arbitration

In the 9V the third specialist cannot overrule a confident Ministral majority since 5/6 and 6/6 are mathematically locked against 3 specialist votes, so it can only intervene on the n=142 samples where Ministral is itself split (Figure 4). The question is whether the specialist flips Ministral's wrong calls (helpful arbitration) or its correct ones (harmful noise). Adding Phi4-LR 8c lowers the system Krippendorff's α from .451 (6V Min-SFT 9c + Min-LR 8c) to .397 (9V) and the drop sits across branches rather than within them (within-branch Min-SFT .617, Min-LR .630, Phi4-LR .464; lowest cross-pair Min-SFT×Phi4-LR .382). Such cross-branch disagreement among accurate base models is the well-known prerequisite for ensemble gains beyond the strongest member (Dietterich, 2000). But low Krippendorff's α only proves the voters disagree, not whether they disagree where it matters, so we trace the actual flips.

Phi4-LR flips Ministral on 39 of those 142 samples and 33 (85%) touch the C6/C7 boundary that dominates the per-class confusion matrix (Figure 3). C7 is the most common flip source (12 of 39, 31%), reflecting Min's tendency to over-call the majority class, but the redistribution is heterogeneous with no single direction dominating (top flip C7→C3 is 6 of 39). Swapping Phi4-LR for Llama-LR or PCounsel-LR yields .417 and .414, so the gain reproduces across all three candidates.

3.3 Augmentation Ablation

Without GPT-5.2 augmentation (*no-aug*; Table 2), the 9V loses only .042. Augmentation alone hurts a single model (−.008), but voting averages over the synthetic noise and turns augmentation into a +.042 lift on top of voting's +.063—they are interlocked, not additive.

System	aug	no-aug
Min-SFT 9c (single, no voting)	.307	.315
5V Min-SFT 9c	.373	.319
+ Min-LR 8c (6V)	.391	.369
+ Phi4-LR 8c (9V)	**.420**	.378

Table 2: Augmentation ablation ($F1_{test}$, classes 1–8); augmentation hurts a single model but helps once voting averages the noise.

4 Conclusion

Our 9-voter ensemble reaches $F1_{test}$=.420 (+33.4% over baseline), driven by voter error independence. C0 forms a well-separated cluster while the defence-class clusters are less distinct—exactly the fuzzy boundaries our approach disentangles. We got there by stacking diversity axes (class granularity, training method, base model) and following the $F1_{cv}$ signal. GPT-5.2 augmentation hurts a single model alone but lifts the 9V by +.042. Post-hoc analysis reveals aug × no-aug as another diversity axis—a 6V combining aug Min-SFT 9c and no-aug PCounsel-LR 8c clears .40 (Appendix Table 6), supporting our hypothesis that independent errors are the lever. Voter diversity tends to help but does not scale arbitrarily. Disentangling these clusters with richer signal and matching the 9V more cheaply remain open.

Limitations

PSYDEFCONV provides only 1,864 training samples, so the $+.029$ gain from the best non-cross-model 6V (.391) to the 9V (.420) is a single hidden-test observation and what generalises is the complementary-model selection principle, not the exact 9V ranking. Several design choices—specialist selection by an n=5 Pearson correlation, top-3 fold selection without cross-validation and the C0-override threshold—rest on limited statistical support, so the submitted configuration is one of several plausible winners. Moderate annotator agreement (κ=.639) bounds the macro-F1 target and places the rare clinical classes C2, C5 and C8 inside the annotator-disagreement band. All evaluation is on PSYDEFCONV/ESConv (English) with 738 GPT-5.2 synthetic dialogues that carry generator-specific artefacts. The 9V needs distillation for real-time deployment, though the 5V single-branch already captures more than half of the gain at one third of the inference cost.

Ethics Statement

At $F1_{test}$=.420 the system misclassifies the majority of defence-bearing utterances with an adaptive-skew bias that under-flags exactly the patients who most warrant clinical attention. Assigning defence labels to a person's utterances is itself a psychological intervention and should not occur outside supervised clinical workflows with informed consent (Steigerwald et al., 2026; Na et al., 2025). Like other LLM-based tools in mental health, such systems should augment, not replace, the human practitioner (Steigerwald et al., 2025); outputs are categorical labels only and could be misused to pathologise individuals in adversarial contexts, mirroring known privacy, bias and accountability risks of mental-health LLMs (Steigerwald and Albrecht, 2026). The DMRS taxonomy reflects Western, English-language therapeutic traditions and PSYDEFCONV (Liu et al., 2021) is simulated rather than clinical data whose crowdworkers consented to support-dialogue collection, not to psychodynamic re-annotation.

Data Availability

The 738 GPT-5.2 synthetic dialogues, generation prompt and parameters are released under CC BY-NC 4.0 at `https://github.com/th-nuernberg/nuernberg-nlp-psydefdetect`.

References

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *MCS 2000, Lecture Notes in Computer Science*, pages 1–15. Springer.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 3469–3483.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Yining Hua, Rena Gao, Kailai Yang, Ling Chen, Wei Wang, Shaoxiong Ji, John Torous, and Sophia Ananiadou. 2026a. Overview of the PsyDefDetect shared task at BioNLP 2026: Detecting levels of psychological defense mechanisms in supportive conversations. In *Proceedings of the 25th Workshop on Biomedical Language Processing*, San Diego, USA. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.

J Christopher Perry. 1990. *Defense Mechanism Rating Scales (DMRS)*, 5th edition. Cambridge, MA.

J Christopher Perry. 2014. Anomalies and specific functions in the clinical identification of defense mechanisms. *Journal of Clinical Psychology*, 70(5):406–418.

Philipp Steigerwald and Jens Albrecht. 2026. From "Help" to helpful: A hierarchical assessment of LLMs in mental e-health applications.

Philipp Steigerwald, Nico Bienlein, Jennifer Burghardt, Mara Stieler, Robert Lehmann, and Jens Albrecht. 2025. CAIA in practice: Field evaluation of an AI-assisted support system for text-based online counselling. In *2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE.

Philipp Steigerwald, Jennifer Burghardt, Eric Rudolph, and Jens Albrecht. 2026. AI systems in text-based online counselling: Ethical considerations across three implementation approaches.

A Hyperparameter Details

All fine-tuning uses 4-bit NF4 QLoRA on all linear projections (dropout 0.05, cosine schedule with 10% warm-up, 10 epochs, effective batch size 8, max sequence length 4,096 tokens). SFT uses LoRA rank 32 (α=64), learning rate 10^{-4}, cross-entropy on the label digit. ClsHead uses rank 16 (α=32), learning rate 2×10^{-5}, focal loss (Lin et al., 2017) (γ=2) with inverse-frequency class weights $w_c = N/(K\,n_c)$. LR is L2-regularised multinomial logistic regression on frozen last-token hidden states ($C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ swept per fold; `class_weight="balanced"`; `scikit-learn` defaults). Augmented samples enter only training.

SFT Prompt Template

System: You are an expert psychologist specialising in the Defense Mechanism Rating Scale (DMRS). You analyse emotional support conversations and classify the psychological defence mechanisms used by the help-seeker in their utterances. You always respond with exactly one line in the format 'label: <number>' where <number> is 1–8.

User: Below is an emotional support conversation between a SEEKER and a SUPPORTER. Your task is to classify the TARGET utterance according to the Defense Mechanism Rating Scale (DMRS).

Conversation
[Full dialogue history with SEEKER and SUPPORTER turns]

Target Utterance
[The seeker utterance to be classified]

DMRS Defence Mechanism Categories
[Definitions for all 8 (or 9) DMRS levels, e.g. "1: Action: The speaker uses action-oriented defences such as acting out, passive aggression, or help-rejecting complaining…"; full descriptions for every level]

Examine the dialogue carefully and select the single most appropriate defence tier. When multiple defences seem plausible, choose the tier with the strongest supporting evidence. Every utterance contains a defence mechanism. Return exactly one line: `label: <1-8>`.

Assistant: `label: {class_id}`

B Multi-Model CV5 Comparison

Table 3 reports $F1_{cv}$ for all candidate base models, methods and class modes. The LR column dominates across models and Ministral-8B and Phi-4-14B top the 8-class LR column—the pairing adopted in the winning system.

C Data Augmentation and Class Balancing

GPT-5.2 (temperature 0.9) generated 738 synthetic dialogues against the 80/20 dialog-stratified train split (1,520 originals) using a few-shot prompt with the full DMRS taxonomy and five randomly sampled originals of the target class; each dialogue is a 2–6 turn emotional support exchange ending in a target seeker utterance that demonstrates the specified defence level. Table 4 reports the per-class budget against this 80/20 split. The same 738 synthetic dialogues underpin both the "Min-SFT 9c full-train, augmented (single model)" baseline in Table 1 (trained on all 1,864 originals plus the 738 synthetic) and every CV5 voter (each trained on its fold-train split of ∼1,493 originals plus the same 738 synthetic). Synthetic data enters only training splits, while validation and test sets remain exclusively original human-annotated data.

	SFT		ClsHead		LR	
Model	8c	9c	8c	9c	8c	9c
Ministral-8B	.321	.306	.333	.311	**.342**	.315
Phi-4-14B	–	.293	.337	–	**.337**	–
Llama-3.1-8B	.251	.279	.246	.284	.312	.284
Qwen2.5-7B	.266	.256	.302	.268	.307	.283
PsychoCounsel-8B	–	–	.316	–	.301	–
PsyLLM-8B	–	–	.295	–	.289	–
GPT-OSS-20B	.212	.183	.278	–	.292	–

Table 3: $F1_{cv}$ for all candidate base models (mean over n=5 folds, classes 1–8; $\sigma\in[.019, .045]$).

Data Augmentation Prompt (Abridged)

System: You are a psychology expert generating training data. Output only valid JSON.

User: You are an expert psychologist specialising in psychological defence mechanisms.

[Full DMRS taxonomy with definitions and example markers for all 9 levels]

Your Task: Generate {n} NEW and DIVERSE examples of **Level {ℓ}: {name}** in emotional support conversations.

Requirements:
1. Create REALISTIC dialogues between a help-seeker and emotional supporter
2. The TARGET UTTERANCE must clearly demonstrate Level {ℓ}
3. VARY the topics: work stress, relationships, health anxiety, family conflict, finances, grief, …
4. Use natural, conversational English
5. The SEEKER uses the defence mechanism (not the supporter)
6. Dialogue context: 2–6 turns before the target

Few-Shot Examples:
[5 randomly sampled original training examples of the target class, each showing dialogue context + target utterance + label]

Output Format: Generate exactly {n} examples as JSON objects, each with a dialogue (list of speaker/text turns) and a target utterance.

D Embedding Geometry after 8-Class Specialist Training

Figure 5 shows per-class t-SNEs for the two LR 8c specialists. 8-class training does not separate the

ID	Defence Level	Orig.	+Aug	Total
0	No Defence	244	0	244
1	Action	88	112	200
2	Major Image-Dist.	54	146	200
3	Disavowal	83	117	200
4	Minor Image-Dist.	67	133	200
5	Neurotic	34	102	136
6	Obsessional	135	65	200
7	High-Adaptive	794	0	794
8	Needs More Info	21	63	84
	Total	*1,520*	*738*	*2,258*

Table 4: Per-class composition of the 80/20 train split against which the augmentation budget was computed; C0 and C7 are excluded from augmentation but remain in training.

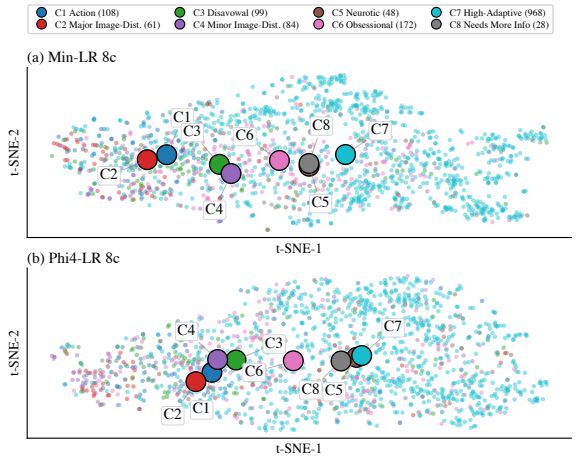

Figure 5: Per-class t-SNE of the LR 8c specialist hidden states (n=1,568, C0 excluded).

defence clusters in either model—C6 and C7 remain the dominant overlap (cf. Figure 3)—yet the two models produce visibly different local geometries, consistent with our finding that error independence drives the ensemble.

E Per-Class Results

Table 5 reports per-class scores of the 9V ensemble, with C3 and C5 below .300 and confirming the C7 absorption seen in Figure 3.

F Post-Hoc Re-Voting Search

The configurations in this appendix are not part of our shared-task submission and were never uploaded to the leaderboard. After the test labels were released we re-voted over our 18 cached per-fold prediction sets (12 aug + 6 no-aug) without retraining; an 18-branch search takes seconds on cached predictions. Fold selection is post-hoc, so these scores are oracle upper bounds rather than blind submissions.

ID	Defence Level	$F1$	P	R	n
0	No Defence	.899	.855	.947	75
1	Action	.583	.700	.500	28
2	Major Image-Dist.	.333	.357	.312	16
3	Disavowal	.291	.267	.320	25
4	Minor Image-Dist.	.350	.368	.333	21
5	Neurotic	.200	.286	.154	13
6	Obsessional	.436	.386	.500	44
7	High-Adaptive	.833	.844	.823	243
8	Needs More Info	.333	.400	.286	7

Table 5: Per-class $F1_{test}$ of the winning 9V ensemble on the hidden test set (n=472).

$F1_{test}$	t	Configuration (gatekeeper + specialists)
6V		
.402	2	Min-SFT 9c + PCounsel-LR 8c (n)
.396	2	Min-SFT 9c + Phi4-LR 8c (n)
.395	2	Min-SFT 9c + Min-LR 8c
9V		
.452	1	Min-SFT 9c + Min-SFTinit-LR 8c + Phi4-LR 8c (n)
.449	2	Min-SFT 9c + Min-SFTinit-LR 8c + Phi4-LR 8c (n)
.445	3	Min-SFT 9c + Min-SFTinit-LR 8c + Phi4-LR 8c (n)
12V		
.471	3	Phi4-SFT 9c + Min-Cls 9c + Min-LR 8c (n) + Phi4-LR 8c (n)
.464	3	Phi4-SFT 9c + Llama-LR 8c + Min-LR 8c (n) + Phi4-LR 8c (n)
.456	3	Phi4-SFT 9c + Min-SFTinit-LR 8c + Min-LR 8c (n) + Phi4-LR 8c (n)

Table 6: Top-3 post-hoc voter combinations on the test set per ensemble size; (n) marks no-aug branches (unmarked = augmented), t is the C0-override threshold; Min-SFTinit-LR is a Min-LR 8c specialist initialised from the SFT 9c adapter rather than ClsHead.

Three patterns emerge (Table 6). First, mixing aug with no-aug branches is itself a diversity axis: 9V ensembles combining aug and no-aug specialists average $F1_{test}$=.391 vs. .372 for aug-only pairs and every top-3 entry from 9V upward draws on at least one no-aug branch. Second, for 6V the best gatekeeper + specialist pairing is the no-aug PCounsel-LR 8c (.402), narrowly ahead of no-aug Phi4-LR 8c (.396); the cross-architecture LR specialists dominate the top of the 6V leaderboard. Third, ensemble gain grows from 9V to 12V (best .452 $\rightarrow$ .471, +.019) and begins to plateau beyond 12V, suggesting that adding more branches from a saturated voter pool re-introduces correlated errors faster than independent signal.