zzucs at PsyDefDetect: Bridging Long-Tail Imbalance and Clinical Rubrics for DMRS Defense-Level Detection

Bin Huang[1] **and Liuyuan Su**[1] **and Kaixuan Yuan**[1] **and Guanghui Zhao**[1] and **Shixin Zhang**[1] and **Kunli Zhang**[1,*]

[1]School of Computer and Artificial Intelligence, Zhengzhou University
Zhengzhou, China

{1084893712, 2805476399, 2285986836}@qq.com, suliuyuan@gs.zzu.edu.cn,
17837272557@163.com, ieklzhang@zzu.edu.cn

*Corresponding author

Abstract

Detecting DMRS defense levels in emotional support dialogues is challenging due to severe class imbalance and fine-grained clinical distinctions between adjacent levels, issues well documented in psychotherapy-oriented NLP surveys (Na et al., 2025). We present **zzucs** for PsyDefDetect at BioNLP 2026 (Na et al., 2026a), adopting a **data–supervision co-design** strategy. **SCCR** applies stratified resampling to balance support across nine defense levels. **CoR–QLoRA** encodes clinical rubrics, including task contracts, taxonomy definitions, and boundary cues, into static prompts for 8B model fine-tuning. Ablations show SCCR improves macro-F1 by **4.9 points** over random oversampling. Our system from team **zzucs**, submitted on CodaBench under the display name *sly_zzu* with submission ID 652647, achieves **0.3585 macro-F1** on the **official blind-test leaderboard LB1**. It ranks **6th of 21** registered teams with official submissions and surpasses all published 8B baselines by **4.4 F1 points** over the strongest 8B comparator, Ministral-8B. The code has been released at https://github.com/jackssdd/zzucs_psydefdetect_code.

1 Introduction

Psychological defenses, broadly defined as automatic strategies for managing distress, shape how individuals disclose emotions and respond to therapeutic support (Vaillant, 1992; Perry and Henry, 2004). The Defense Mechanism Rating Scales (DMRS) operationalize these constructs into a validated clinical taxonomy comprising three defensive categories, seven functional levels, and thirty specific mechanisms (Perry and Henry, 2004; Di Giuseppe and Perry, 2021). Despite their centrality in clinical theory, computational detection of defensive functioning in conversational settings remains largely unaddressed in current emotional support dialogue (ESD) systems (Liu et al., 2021; Rashkin et al., 2019).

The PSYDEFCONV corpus (Na et al., 2026b) introduces the first conversational dataset annotated with DMRS defense levels, establishing a nine-way classification task over help-seeker utterances. The PsyDefDetect shared task at BioNLP 2026 formalizes blind leaderboard evaluation on this setting (Na et al., 2026a). Recent surveys also contextualize LLMs in psychotherapy (Na et al., 2025). Although this appears to be a standard multi-class problem, it poses two concrete challenges that limit conventional supervised learning:

1. **Severe distributional imbalance:** Level 7 (high-adaptive) defenses dominate 51.8% of seeker turns, while Level 1 (action) and Level 5 (neurotic) each constitute merely 2.6–5.8% (Na et al., 2026b). Standard training is biased toward majority classes, leading to poor coverage of clinically important but rare defense levels.
2. **Clinical granularity:** Adjacent DMRS levels encode subtle distinctions. For example, *disavowal* at Level 3 and *neurotic* defenses at Level 5 both involve avoidance, but they differ in awareness and displacement patterns that remain hard to capture with vanilla classification objectives, even with strong pre-trained representations.

We argue that simply scaling model size is insufficient for this long-tailed, clinically nuanced setting. While instruction tuning broadly improves LLM task performance (Zhang et al., 2023), severely imbalanced clinical taxonomies still require targeted data and supervision design rather than a larger backbone (Na et al., 2025). We therefore adopt two complementary strategies:

- **Data-centric rebalancing:** stratified class-conditional resampling with replacement, which preserves within-class diversity while equalizing class support.

Proceedings of the BioNLP 2026 (Shared Tasks), pages 13–23

- **Rubric-based prompting:** encoding DMRS taxonomy definitions and pairwise boundary cues as a static input prefix, providing structured clinical supervision without inference-time generation.

Contributions. Our work makes three contributions: **(i) SCCR**, a stratified resampling method for class-imbalanced clinical dialogue data, which improves macro-F1 by 4.9 points over random oversampling and 9.5 points over no resampling (Table 1). **(ii) Chain-of-Rubric (CoR) prompting**, a structured prompting method that encodes DMRS taxonomy definitions and boundary cues directly in the input prefix. **(iii)** Results on the official LB1 blind-test leaderboard (Na et al., 2026a) showing our 8B QLoRA system outperforms all published fine-tuned 8B baselines. Due to the system paper page limit, **related work** is deferred to Appendix A.

2 Related Work

Na et al. (2026a) describe PsyDefDetect at BioNLP 2026: nine-way DMRS defense-level detection with blind leaderboard metrics over PSYDEFCONV (Na et al., 2026b). Na et al. (2025) survey LLM applications in psychotherapy and identify challenges around data sparsity and evaluation, which motivates structured clinical supervision for DMRS-level tasks. Most emotional support dialogue (ESD) resources emphasize supporter strategies and empathy ratings (Liu et al., 2021; Rashkin et al., 2019). In contrast, PSYDEFCONV targets *seeker-side* DMRS defense levels as a distinct clinical classification problem. Long-tailed text classification is tackled via reweighting, such as focal loss (Lin et al., 2017), resampling (Chawla et al., 2002), or quality-aware synthesis (Peng et al., 2024; Zhou et al., 2024). Our SCCR uses stratified resampling without generative augmentation to protect DMRS validity. Parameter-efficient adaptation (LoRA/QLoRA (Hu et al., 2022; Dettmers et al., 2023)) and structured or chain-style prompts (Wei et al., 2022; Wu et al., 2024) inform our CoR design. **Appendix A** provides an expanded discussion with additional positioning and citations.

3 Methodology

3.1 Task formulation

Each instance comprises dialogue history d, current seeker utterance u, and label $y \in \mathcal{Y} = \{0, \ldots, 8\}$ indexing DMRS levels plus *no defense* (0) and *needs more information* (8), as defined for PsyDefDetect (Na et al., 2026a,b). We predict $\hat{y}$ from contextualized encoding of composed input z.

3.2 Module I: Stratified Class-Conditional Resampling (SCCR)

Motivation. PSYDEFCONV exhibits extreme distributional skew (Appendix Fig. 2, left): Level 7 comprises 1,211 instances (51.8%), while Level 1 (action defenses) and Level 5 (neurotic) comprise merely 61–136 instances (2.6–5.8%) each. These rare classes represent clinically important defensive patterns (e.g., impulsive acting-out at Level 1 and repressed conflict at Level 5) that standard training tends to ignore due to majority-class bias.

Formalization. Partition training set $\mathcal{D}$ into per-class subsets $\mathcal{D}_c$ with sizes $n_c = |\mathcal{D}_c|$. The **class cardinality target** matches the largest class:

$$N^\star = \max_{c \in \mathcal{Y}} n_c \tag{1}$$

For each c, we build multiset $\tilde{\mathcal{D}}_c$ by **sampling with replacement** from $\mathcal{D}_c$ until $|\tilde{\mathcal{D}}_c| = N^\star$, then form $\tilde{\mathcal{D}} = \mathrm{Shuffle}(\bigcup_c \tilde{\mathcal{D}}_c)$. Pseudocode is given as Algorithm 1 in Appendix C.

Unlike *random oversampling*, which duplicates instances drawn uniformly from the union $\mathcal{D}$ until an aggregate balance heuristic is met, SCCR sets an explicit **per-class target** $N^\star$ equal to the majority count and resamples *within each stratum* $\mathcal{D}_c$ independently. This equalizes effective supervision across all nine labels while keeping duplicates drawn only from clinically homogeneous buckets.

The result is a balanced training set (Appendix Fig. 2, right) that preserves within-class diversity through replacement sampling. Importantly, we avoid synthetic utterance generation, which could distort DMRS clinical validity.

Validation protocol. Validation uses the same stratified holdout construction as training splits (at least five labeled examples per class), but we *do not* apply SCCR oversampling there: frequencies follow the natural validation distribution so macro-F1 tracks generalization rather than artificial balance. Checkpoint selection maximizes macro-F1 on this naturally distributed validation set.

3.3 Module II: Chain-of-Rubric (CoR) Prompting with QLoRA

Figure 1 summarizes the full pipeline. Appendix Fig. 3 illustrates the three-channel rubric layout in detail.

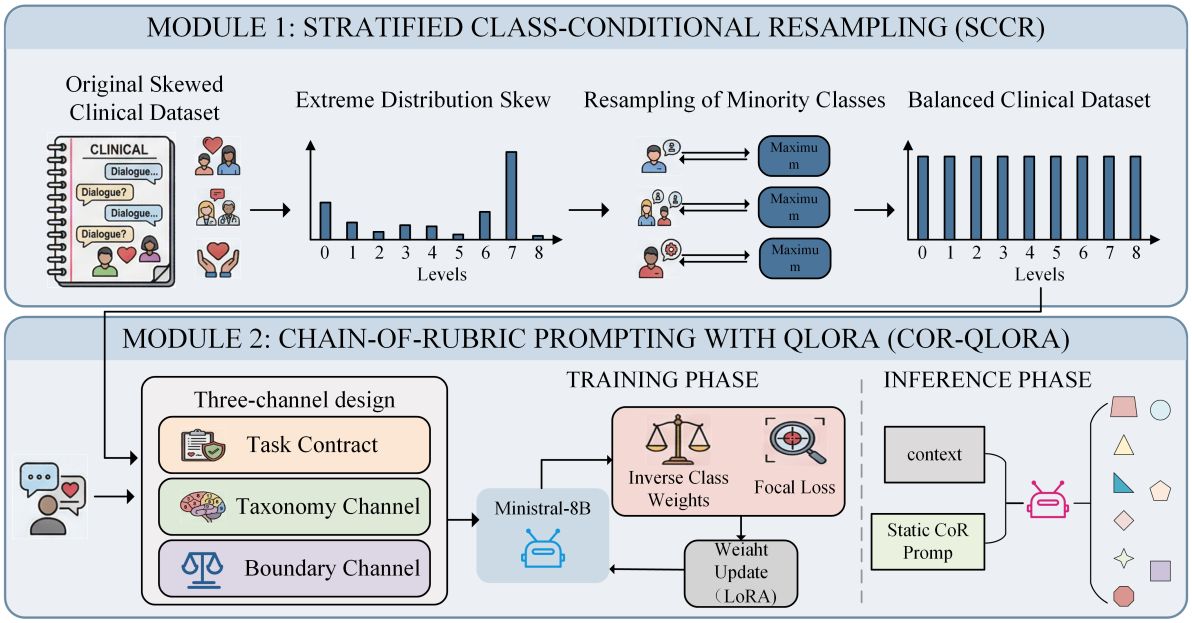

Figure 1: **System architecture. Module I** (SCCR) transforms imbalanced training data through stratified oversampling with replacement. **Module II** (CoR–QLoRA) constructs three-channel rubric prompts (task contract, taxonomy definitions, boundary cues) for injection into a quantized 8B backbone with LoRA adapters. Training employs imbalance-aware focal loss with inverse-frequency class weighting.

Three-channel rubric design. We construct instruction prefix π as **channel concatenation**:

1. **Task Contract (Channel a):** Output space $\mathcal{Y}$, role restriction (seeker-only), evidence boundary (context up to current turn).
2. **Taxonomy Channel (Channel b):** One-sentence definitions for each DMRS level (0–8) based on the clinical taxonomy, giving the model explicit conceptual anchors.
3. **Boundary Channel (Channel c):** Pairwise disambiguation cues for adjacent level pairs that commonly cause confusion:
   - Level 2 vs 4: extreme versus milder image distortion
   - Level 3 vs 5: defensive avoidance versus indirect neurotic displacement
   - Level 6 vs 7: rigid intellectualization versus flexible adaptive coping

Rationale. While chain-of-thought prompting (Wei et al., 2022) improves reasoning through intermediate generation, this incurs inference-time overhead and can introduce hallucinated reasoning chains. By encoding the rubric directly in the input prefix, we guide the model toward clinically meaningful distinctions without any inference-time generation cost.

Input serialization.

$$z = \underbrace{\pi}_{\text{rubric}} \,\|\, \underbrace{\texttt{META}(d,u)}_{\text{metadata}} \,\|\, \underbrace{\text{CTX}(d)}_{\text{context}} \,\|\, \underbrace{u}_{\text{target}} \tag{2}$$

QLoRA adaptation. We encode z with Ministral-8B-Instruct (4-bit NF4 quantization (Dettmers et al., 2023)), attach a 9-way classification head, and fine-tune LoRA adapters (r=16, α=32, dropout 0.05) on attention and MLP projection paths (Hu et al., 2022).

3.4 Imbalance-aware optimization

Combining inverse-frequency class weights $\alpha_y \propto 1/n_y$ with focal loss (Lin et al., 2017):

$$\mathcal{L}_{\text{focal}} = -\frac{1}{|\tilde{\mathcal{D}}|} \sum_{(z,y)\in\tilde{\mathcal{D}}} \alpha_y \big(1 - p_\theta(y|z)\big)^\gamma \times \log p_\theta(y|z) \tag{3}$$

where γ=1.5 and α_y are renormalized to sum to $|\mathcal{Y}|$. Checkpoints are selected by macro-F1 on stratified validation sets with class-wise evaluation.

4 Experiments

4.1 Experimental setup

Data We train on the official stratified PsyDefDetect training partition of PSYDEFCONV (Na et al.,

Configuration	Acc.	Macro-P	F1
Resampling:			
No resampling	.523	.301	.289
Random oversampling	.561	.342	.335
SCCR (stratified)	**.608**	**.468**	**.384**
Prompt design:			
CoR v1: minimal	.579	.422	.351
CoR v2: + taxonomy	.608	.468	.384
CoR v3: + boundary	**.612**	**.471**	**.388**
Loss functions:			
Standard CE	.589	.446	.371
Class weighting only	.598	.459	.380
Focal (γ=1.5) + weight	**.608**	**.468**	**.384**

Table 1: Controlled ablations on **stratified validation sets** (not blind test). All experiments use SCCR+focal+CoR v2 as base, varying one factor. Note: Validation F1 is higher than blind test (0.3585 in Table 2) due to distribution similarity with training.

System	Acc.	P	R	F1
Zero-shot LLMs (from (Na et al., 2026b)):				
GPT-5	.528	.276	.166	.195
Gemini 2.5 Pro	.564	.275	.261	.260
DeepSeek-V3.2	.557	.297	.275	.262
Fine-tuned 8B (from (Na et al., 2026b)):				
Llama3.1-8B	.629	.332	.301	.305
GLM-4-9B	.629	.301	.295	.286
Qwen3-8B	.614	.301	.289	.284
InternLM3-8B	.640	.335	.299	.305
Ministral-8B	.648	.340	.305	.315
Ours (zzucs)	**.644**	**.397**	**.352**	**.359**†

Table 2: Baseline numbers reproduced from Table 5 of Na et al. (2026b) (not our re-implementations). †**Official LB1** blind-test macro-F1 for team **zzucs** is 0.3585. Rank is **6/21** among registered teams with official submissions. Submission ID is 652647, and the CodaBench display name is *sly_zzu*.

2026b,a): 2,336 seeker utterances with 9-way labels. Validation follows the protocol in §3.2, using a stratified holdout with natural class frequencies and at least five examples per class.

Implementation. Base model: Ministral-8B-Instruct-2410. QLoRA uses 4-bit NF4. Max sequence length is 384. Training: 5 epochs, LR 2×10^{-5}, warmup 5%, batch size 1 with gradient accumulation (effective batch 4–8). See Appendix D for complete hyperparameters.

4.2 Ablation studies

Table 1 summarizes single-factor ablations on stratified validation. Expanded interpretation is in Appendix F. **Resampling:** SCCR outperforms both no resampling and random oversampling (+4.9–9.5 F1), consistent with stratified class-conditional replacement sampling. **Prompt design:** Taxonomy (v2) and boundary cues (v3) improve over v1. Version v3 attains the best F1 despite a higher eval loss than v2. **Loss:** Focal loss (γ=1.5) with inverse-frequency weighting beats standard CE and weighting-only variants.

4.3 Comparison to published baselines

Table 2 contextualizes our result against published benchmarks tabulated by Na et al. (2026b), following the PsyDefDetect evaluation framing and leaderboard metrics summarized by Na et al. (2026a). This aligns with observations in Na et al. (2025) that clinical NLP tasks benefit from structured supervision beyond generic LLM scaling. Our primary metric is **official LB1 macro-F1 = 0.3585** on the blind test (Na et al., 2026a). Table 1 shows development-set ablations. Our 8B QLoRA system achieves superior Macro-P (0.3969) and Macro-F1 (0.3585), outperforming the strongest listed 8B full fine-tuning baseline (Ministral-8B at 0.3148 F1) by **4.4 F1 points**.

4.4 Error analysis

On our stratified validation split, errors disproportionately map to Level 7, accounting for about 54% of mistakes. Adjacent-level confusion remains concentrated on 3 vs. 5 and 6 vs. 7 despite boundary cues. Rare Levels 1 and 5 also remain recall-limited after SCCR. Appendix E reports the full diagnostic breakdown and discussion.

5 Conclusion

We described a data and supervision co-design approach for DMRS defense-level detection, combining stratified resampling (SCCR) with rubric-based prompting (CoR) and quantized LoRA fine-tuning. SCCR improves macro-F1 by 4.9 points over random oversampling, and adding boundary cues to the prompt yields further gains despite a higher validation loss, suggesting that structured clinical knowledge acts as a useful regularizer. Our 8B QLoRA system outperforms the strongest published 8B fine-tuning baseline, showing that targeted data and prompt engineering can compensate for limited model scale on clinically fine-grained, long-tailed tasks.

Future work includes dynamic rubric retrieval for ambiguous utterances, quality-preserving syn-

thetic augmentation for rare classes, and cross-lingual transfer to Chinese and Spanish settings where DMRS frameworks may differ. **Ethical note**: Automated defense labels are research tools for analyzing supportive dialogue, not diagnostic substitutes.

Limitations

Language scope: English-only. DMRS taxonomy transfer to other languages requires validation.

Evaluation variance: Single submitted run. Seed and hyperparameter sensitivity remain uncharacterized.

Modality restriction: Text-only. Prosodic cues such as hesitation and affect are unavailable.

Rubric design: Hand-crafted boundary cues. Learned or retrieved rubric augmentation may improve adaptability.

Clinical validity: Synthetic rare-class augmentation must preserve DMRS clinical semantics, an open research challenge.

Acknowledgments

We thank the PsyDefDetect organizers for the benchmark and annotation guidelines. This work was supported by the Natural Science Foundation of Henan Province (Grant No. 252300421877).

References

Renad M. Alzghoul, Abdulrahman Tabaza, Aya Abdelhaq, and Ahmad Altamimi. 2024. CLD-MEC at MEDIQA-CORR 2024 task: GPT-4 multi-stage clinical chain of thought prompting for medical errors detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 537–556, Mexico City, Mexico. Association for Computational Linguistics.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

Mariagrazia Di Giuseppe and J. Christopher Perry. 2021. The hierarchy of defense mechanisms: Assessing defensive functioning with the defense mechanisms rating scales q-sort. *Frontiers in Psychology*, 12:718440.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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 (ICCV)*, 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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Yuhan Liu, Jun Gao, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. Empathetic response generation with state management. *Preprint*, arXiv:2205.03676.

Zeyu Liu, Souvik Kundu, Anni Li, Junrui Wan, Lianghao Jiang, and Peter A. Beerel. 2024. AFLoRA: Adaptive freezing of low rank adaptation in parameter efficient fine-tuning of large models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 161–167, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Hongbin Na, Zimu Wang, Zhaoming Chen, Peilin Zhou, Yining Hua, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, John Torous, Shaoxiong Ji, and Ling Chen. 2026b. You never know a person, you

only know their defenses: Detecting levels of psychological defense mechanisms in supportive conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, San Diego, USA. Association for Computational Linguistics.

Letian Peng, Yi Gu, Chengyu Dong, Zihan Wang, and Jingbo Shang. 2024. Text grafting: Near-distribution weak supervision for minority classes in text classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3741–3752, Miami, Florida, USA. Association for Computational Linguistics.

J. Christopher Perry and Melissa Henry. 2004. Studying defense mechanisms in psychotherapy using the defense mechanism rating scales. In Uwe Hentschel, Gudmund Smith, Juris G. Draguns, and W. Ehlers, editors, *Defense Mechanisms: Theoretical, Research and Clinical Perspectives*, volume 136 of *Advances in Psychology*, pages 165–192. Elsevier.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Pengjie Ren, Chengshun Shi, Shiguang Wu, Mengqi Zhang, Zhaochun Ren, Maarten de Rijke, Zhumin Chen, and Jiahuan Pei. 2024. MELoRA: Mini-ensemble low-rank adapters for parameter-efficient fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3052–3064, Bangkok, Thailand. Association for Computational Linguistics.

George E. Vaillant. 1992. *Ego Mechanisms of Defense: A Guide for Clinicians and Researchers*. American Psychiatric Press, Washington, DC.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35.

Zhaolong Wu, Abul Hasan, Jinge Wu, Yunsoo Kim, Jason P.Y. Cheung, Teng Zhang, and Honghan Wu. 2024. KnowLab_AIMed at MEDIQA-CORR 2024: Chain-of-though (CoT) prompting strategies for medical error detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 353–359, Mexico City, Mexico. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Yuhang Zhou, Jing Zhu, Paiheng Xu, Xiaoyu Liu, Xiyao Wang, Danai Koutra, Wei Ai, and Furong Huang. 2024. Multi-stage balanced distillation: Addressing long-tail challenges in sequence-level knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3315–3333, Miami, Florida, USA. Association for Computational Linguistics.

A Related Work (expanded)

Emotional support dialogue and defensive functioning. ESConv (Liu et al., 2021) established strategy-grounded supportive interaction corpora, with subsequent work extending multi-strategy turn planning and empathetic response evaluation (Rashkin et al., 2019; Liu et al., 2022). However, existing ESD benchmarks focus on supporter strategy selection or empathy ratings, and largely ignore seeker-side defensive functioning, i.e., how distressed individuals manage psychic pain through DMRS defense mechanisms (Na et al., 2026b). Our work addresses this gap by targeting PSYDEFCONV under the PsyDefDetect protocol (Na et al., 2026a). The challenges we observe are consistent with broader psychotherapy LLM issues surveyed by Na et al. (2025).

Long-tailed recognition in text classification. Natural class imbalance in real-world text corpora motivates diverse mitigation strategies. *Reweighting* approaches (focal loss (Lin et al., 2017), label-distribution-aware margins (Cao et al., 2019)) adjust loss contributions. *Resampling* approaches oversample minorities or undersample majorities, with naive random oversampling risking duplicate overfitting (Chawla et al., 2002). Recent work emphasizes *quality-aware* synthesis. Text Grafting (Peng et al., 2024) mines near-distribution templates for minority class augmentation. BalDistill (Zhou et al., 2024) dynamically selects representative head examples while synthesizing tail domain instances for knowledge distillation. SCCR differs by pursuing **pure resampling without synthesis**, leveraging the clinical specificity of PSYDEFCONV utterances where synthetic preservation of DMRS validity is challenging.

Parameter-efficient adaptation and structured prompting. LoRA and its quantized variant QLoRA enable efficient adaptation of large LMs with minimal trainable parameters (Hu et al., 2022; Dettmers et al., 2023). Concurrent work explores adaptive freezing (AFLoRA (Liu et al., 2024)) and mini-ensemble adapters (MELoRA (Ren et al., 2024)) for improved capacity–efficiency tradeoffs.

For classification under limited supervision, *chain-of-thought* prompting (Wei et al., 2022) exposes intermediate reasoning. Similar gains can also come from *schema-based* instructions that encode task structure without generation overhead. Clinical NLP applications demonstrate structured prompting efficacy: MEDIQA-CORR systems employ multi-stage clinical chain-of-thought for medical error detection (Wu et al., 2024; Alzghoul et al., 2024). We extend this line of work by encoding pairwise boundary cues for adjacent DMRS levels directly into the static prompt, providing clinical knowledge supervision without inference-time generation.

B Official benchmark portal

The PsyDefDetect organizers designate **LB1 (macro-F1 on blind test)** as the official shared-task metric (Na et al., 2026a). Numbers should be verified against the official results sheet[1].

Team **zzucs** achieves **LB1 macro-F1 = 0.3585**, ranking **6th of 21** registered teams that submitted to official evaluation.

The public **CodaBench** leaderboard (results tab) aggregates all uploaded entries. The organizers note that it contains 64 entries overall, while the official shared-task ranking is computed over the 21 registered teams with official submissions: `https://www.codabench.org/competitions/12124/#/results-tab`.

C Supplementary figures (SCCR and CoR)

For readability, high-resolution plots referenced from §3.2–3.3 are placed here.

D Implementation details

Hardware. NVIDIA RTX 3090 (24GB) and A100 (40GB) GPUs are used. Training time is approximately 45 minutes per run.

[1] `https://docs.google.com/spreadsheets/d/1fMDIwC4sisOguLKm3yMd2uMrC4_KjmAiHuWm_QQuld8/edit?gid=0#gid=0`

Algorithm 1 Stratified Class-Conditional Resampling (SCCR)

Require: Training set $\mathcal{D}$, label space $\mathcal{Y}$
Ensure: Balanced dataset $\tilde{\mathcal{D}}$

```
Partition D into per-class subsets D_c
N* ← max_c |D_c|
for each c ∈ Y do
    D̃_c ← {}
    while |D̃_c| < N* do
        x ~ Uniform(D_c)
        D̃_c ← D̃_c ∪ {x}
    end while
end for
D̃ ← Shuffle(⋃_c D̃_c)
return D̃
```

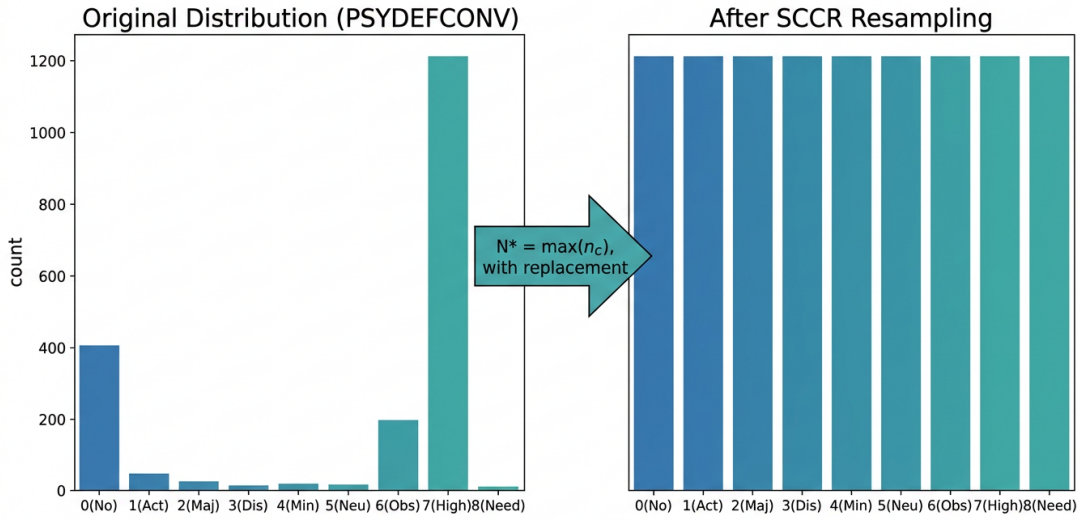

Figure 2: **SCCR.** *Left:* original training distribution (Level 7 majority). *Right:* after resampling, each class has $N^\star$=1,211 examples (replacement sampling).

Software stack. PyTorch 2.1.2, Transformers 4.36.2, PEFT 0.7.1, BitsAndBytes 0.41.3, LLaMA-Factory 0.6.0.

Hyperparameters.

- Base model: Ministral-8B-Instruct-2410 (vocab size 32000)
- Quantization: 4-bit NF4, double quantization enabled, compute dtype bfloat16, quant type nf4
- LoRA: $r = 16$, $\alpha = 32$, dropout 0.05, bias none, target modules q_proj,k_proj,v_proj,o_proj,gate_proj,up_proj,down_proj
- Training: 5 epochs, LR 2×10^{-5} constant with warmup, warmup ratio 5%, batch size 1, gradient accumulation steps 8 (effective batch 8)
- Optimizer: AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, weight decay 0.01, max grad norm 1.0
- Focal: $\gamma = 1.5$, class weights $\alpha_y \propto 1/n_y$ renormalized to sum to 9
- Early stopping: patience 5 on macro-F1, evaluation strategy steps, eval steps 80

SCCR implementation. Python/NumPy with fixed seed (42) for reproducibility. Training receives SCCR-balanced multisets as in Algorithm 1. For validation, we hold out a stratified subset with natural

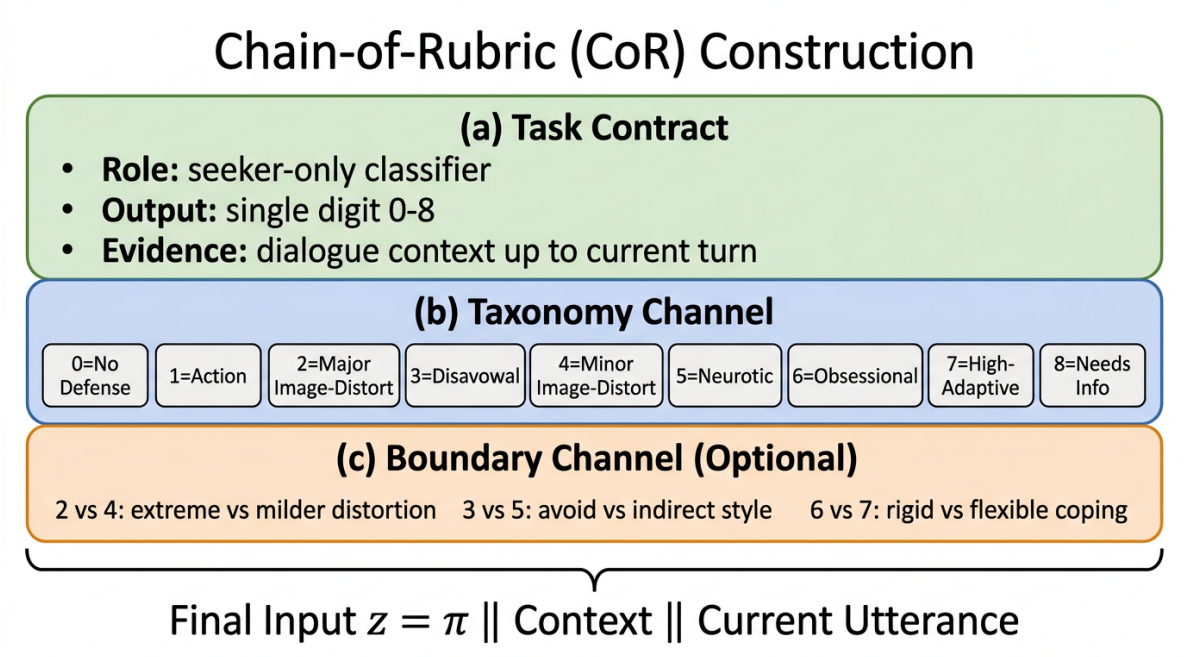

Figure 3: **Chain-of-Rubric (CoR) prompt.** Prefix π: task contract, nine-level taxonomy, pairwise boundary cues.

label frequencies and at least five examples per class where available. If $|n_c| < 5$ in the validation fold, we use all examples of class c without replacement.

E Error analysis (diagnostic)

Level 7 over-prediction. Even with SCCR, models conservatively predict Level 7 for ambiguous utterances (54.2% of errors). This suggests residual *majority bias* from pre-training on general-domain text where adaptive coping is linguistically dominant. Margin-based calibration or decoupled representation learning may further mitigate this.

Boundary confusion. Adjacent levels with subtle distinctions remain challenging despite boundary channel injection. Confusion between Level 3 and Level 5 constitutes 23.7% of errors, while confusion between Level 6 and Level 7 accounts for 18.4%. Manual inspection reveals contextual ambiguity (e.g., sarcasm masking disavowal) defeats explicit rubric cues, suggesting need for *dynamic* rubric conditioning on utterance context.

Rare class recall. Levels 1 (action) and 5 (neurotic) achieve 38.2% and 41.7% recall respectively even with SCCR. This indicates that *data quantity alone is insufficient*. These clinically critical but rare patterns may require specialized synthetic augmentation that preserves DMRS validity, or multi-task pretraining on related clinical constructs such as impulsivity and repression.

F Ablation discussion (expanded)

This section expands the interpretation of Table 1 referenced from §4.2.

Resampling. SCCR yields gains over both no resampling (+9.5 F1) and random oversampling (+4.9 F1). The key difference is per-class targeting to $N^\star$, rather than uniform random duplication, which equalizes class support while keeping each class's internal diversity intact.

Prompt design. Progressive rubric enrichment improves performance. Version v2 with the taxonomy channel gains +3.3 F1 over v1, and version v3 with boundary cues adds another +0.4 F1. Notably, v3's best validation loss is higher than v2 (0.6441 vs 0.6219), yet v3 achieves better F1. This suggests that the boundary cues act as a form of regularization, improving generalization even when the model has not fully minimized training loss.

Loss functions. Focal loss with γ=1.5 improves over standard CE (+1.3 F1) and pure class weighting (+0.4 F1), confirming hard-example mining benefits in this long-tailed setting.

G Complete CoR prompt template (v3)

```
PSYDEFCONV — DMRS Defense Level Classification for Help-Seeker Utterances

[TASK CONTRACT]
Role: Classify the help-seeker's CURRENT utterance only (not supporter).
Output: Single digit 0-8 representing defense level. No explanations.
Evidence boundary: Use ONLY provided dialogue context up to current turn.

[TAXONOMY CHANNEL — DMRS Level Definitions]
0 No Defense — Phatic/functional utterances with no psychological conflict.
1 Action Defenses — Distress acted impulsively on environment; impulsive,
   little reflection (e.g., acting out, passive aggression).
2 Major Image-Distorting — Extreme black-white splitting of self/other
   representations; gross distortion to manage intolerable anxiety.
3 Disavowal — Denial, rationalization, projection, autistic fantasy;
   refusing to acknowledge unacceptable aspects of reality.
4 Minor Image-Distorting — Milder self-esteem protection than Level 2;
   devaluation, idealization, omnipotence with less severity.
5 Neurotic — Repression, dissociation, reaction formation, displacement;
   managing conflict by keeping unacceptable wishes out of awareness.
6 Obsessional — Isolation of affect, intellectualization, undoing;
   feelings split from facts, rigid, excessive logic, flat affect.
7 High-Adaptive — Affiliation, altruism, anticipation, humor, sublimation;
   flexible, constructive coping integrating feelings with plans.
8 Needs More Information — Context insufficient for classification.

[BOUNDARY CHANNEL — Pairwise Disambiguation]
If uncertain between:
• Level 2 vs 4: 2 = extreme/gross distortion; 4 = milder self-esteem protection
• Level 3 vs 5: 3 = defensive avoidance/externalizing; 5 = indirect/displaced
  neurotic style with conflict kept unconscious
• Level 6 vs 7: 6 = rigid affect-intellect split, technical/detached;
  7 = flexible integration of feelings with constructive coping

[OUTPUT FORMAT]
Single digit: 0|1|2|3|4|5|6|7|8
```

H Per-class performance breakdown

Level	Without SCCR			With SCCR		
	Prec.	Recall	F1	Prec.	Recall	F1
0	0.72	0.58	0.64	0.74	0.68	0.71
1	0.12	0.08	0.10	0.38	0.38	0.38
2	0.21	0.15	0.17	0.42	0.41	0.41
3	0.28	0.22	0.24	0.45	0.44	0.44
4	0.25	0.18	0.21	0.41	0.39	0.40
5	0.14	0.09	0.11	0.42	0.42	0.42
6	0.48	0.38	0.42	0.52	0.51	0.51
7	0.68	0.82	0.74	0.64	0.71	0.67
8	0.35	0.28	0.31	0.48	0.46	0.47

Table 3: Per-class precision, recall, and F1 scores on validation set, comparing training without and with SCCR resampling. SCCR dramatically improves rare class (1,2,3,4,5) performance (+27–31 F1 points) while slightly reducing majority class (7) performance (-7 F1), indicating effective rebalancing.

I Negative results and abandoned directions

All macro-F1 figures: **validation**, classes 0–8 (not blind test).

Encoders. BERT-/RoBERTa-/ALBERT-base (CLS head) reached only **0.16–0.18**: short context, imbalance, and adjacent-level DMRS cues favor larger encoders.

Two-LLM few-shot augmentation. G uses label-conditioned few-shot ICL to propose $\tilde{x}$. S then scores the candidate with threshold τ. Accepted pairs $(\tilde{x}, c)$ are added to $\mathcal{D}$ (Alg. 2).

Algorithm 2 Few-shot LLM generate + score (abandoned)

Require: $\mathcal{D}$, $\mathcal{Y}_{\text{rare}}$, pools $\mathcal{E}_c^{\text{gen}}$, $\mathcal{E}^{\text{scr}}$, LLMs G, S, τ, K_c
Ensure: $\mathcal{D}'$

$\mathcal{D}' \leftarrow \mathcal{D}$
for $c \in \mathcal{Y}_{\text{rare}}$ **do**
  for $k = 1, \ldots, K_c$ **do**
    $\tilde{x} \leftarrow G(\textsc{GenPrompt}(\mathcal{E}_c^{\text{gen}}, c))$
    $s \leftarrow S(\textsc{ScorePrompt}(\mathcal{E}^{\text{scr}}, \tilde{x}))$
    if $s \geq \tau$ **then**
      $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{(\tilde{x}, c)\}$
    end if
  end for
end for
return $\mathcal{D}'$

Retraining reached ≈**0.26** macro-F1, which is below stronger non-augmented runs. We therefore use SCCR on real utterances instead (§3).

Ensemble. QLoRA on Qwen2.5-7B + Qwen3-8B with logits averaging reaches ≈**0.28**. This remains well below Ministral-8B + SCCR + Focal, and errors stay correlated on pairs such as 3/5 and 6/7.

Approach	Macro-F1 (val.)
BERT-base (CLS)	0.16
RoBERTa-base (CLS)	0.17
ALBERT-base (CLS)	0.18
Two-LLM few-shot + filter (Alg. 2)	≈0.26
Qwen2.5-7B + Qwen3-8B QLoRA ensemble	≈0.28

Table 4: Suboptimal explorations (validation macro-F1, classes 0–8).