# Self-State Identification with Retrieved In-Context Examples and Open-Weight LLMs

**Alina Ponomareva*** **and Nina Stekacheva Sancho*** **and Karina Litvinova***
University of Zurich
{alina.ponomareva,nina.stekachevasancho,karina.litvinova}@uzh.ch

## Abstract

This paper describes the StateOfMIND team submission to Task 1 of the CLPsych 2026 shared task on identifying adaptive and maladaptive psychological self-states in social media posts. For each post, the task requires systems to predict which fine-grained subelements of a self-state are expressed (Task 1.1) and to rate how strongly adaptive and maladaptive states are present (Task 1.2). We address both subtasks with a retrieval-augmented in-context learning ensemble of two open-weight LLMs, Qwen3.5-27B and Mistral-Small-3.2-24B-Instruct. Our approach combines label-aware retrieval of in-context examples, a three-prompt decomposition targeting adaptive content and the Affect element, and subtask-specific ensemble rules that favor precision for subelement classification and recall for presence rating. Our best submissions rank 2nd of 17 teams on Task 1.1 (Macro F1 = 0.441) and 5th of 17 teams on Task 1.2 (RMSE = 0.994), with ablations showing retrieval and ensembling as the largest drivers of performance.

## 1 Introduction

Large language models are increasingly used to support mental health-related NLP, particularly for interpreting psychological signals in user-generated text (Guo et al., 2024; Malgaroli et al., 2025). Yet clinically meaningful annotation often requires mapping nuanced language onto specialized psychological constructs rather than detecting broad sentiment or distress. The CLPsych shared tasks provide a controlled setting for evaluating such capabilities under limited-data conditions.

The CLPsych 2026 shared task (Ali et al., 2026) builds on prior CLPsych work on mental health modeling from social media (Tsakalidis et al., 2022; Tseriotou et al., 2025). In particular, it extends the CLPsych 2025 task, which was grounded in the MIND framework (Atzil-Slonim, 2025, 2026). Following the ABCD view of personality states (Revelle, 2007), MIND represents self-states through four component types: Affect (A), Behavior toward Self and Others (B-S, B-O), Cognition about Self and Others (C-S, C-O), and Desire (D). These components define adaptive and maladaptive self-states, which reflect how individuals experience themselves and others at a given moment. While the broader task emphasizes longitudinal modeling of self-states for mental health analysis, Task 1 focuses specifically on post-level identification of these components and their relative psychological centrality.

In this work, we address Tasks 1.1 and 1.2, which require predicting ABCD subelement compositions (Atzil-Slonim, 2025) and estimating the presence of adaptive and maladaptive self-states for individual posts. The task is challenging because it combines sparse fine-grained labels, ordinal presence ratings, and limited training data annotated under a specialized psychological framework. We propose a retrieval-augmented in-context learning system that annotates each post independently using two open-weight LLMs and three targeted prompts per model. Our contributions are:

- A retrieval pipeline that combines semantic similarity, label-aware scoring, and diversity re-ranking to select in-context examples;
- Separate unified, adaptive-focused, and Affect-focused prompts for better coverage of adaptive content and Affect subelements;
- Subtask-specific cross-model merging that favors precision for subelement classification and recall for presence estimation.

Empirically, we show that retrieval contributes most to performance, that no single merge strategy optimizes both subtasks, and that the Affect-focused prompt accounts for most of the F1 gain from prompt decomposition.

*All authors contributed equally.



*Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026)*, pages 521–530

## 2 Background and Related Work

### 2.1 LLMs for Mental Health NLP

The application of LLMs to clinical and mental health NLP has grown rapidly, covering mental health prediction from online text, interpretable analysis of social media posts, and broader clinical applications such as screening, monitoring, and decision support (Uluslu et al., 2024; Xu et al., 2024; Yang et al., 2024; Guo et al., 2024; Malgaroli et al., 2025). Their ability to capture nuanced linguistic patterns makes them well-suited to clinically grounded tasks that require detailed interpretation beyond surface-level sentiment. Recent work has shown that prompt engineering and structured inference strategies are particularly important for adapting general-purpose LLMs to such tasks, especially when annotated data follows specialized clinical or psychological frameworks (He et al., 2023; Sivarajkumar et al., 2024).

### 2.2 In-Context Learning and Retrieval-Augmented Prompting

In-context learning has become a standard way to adapt LLMs to specialized tasks without fine-tuning (Brown et al., 2020; Liu et al., 2023), and the choice of in-context examples can substantially affect performance (Liu et al., 2022). Retrieval-augmented approaches select examples from a labeled pool using dense similarity to the target input (Rubin et al., 2022), and can be combined with diversity criteria such as Maximal Marginal Relevance (Carbonell and Goldstein, 1998) to avoid redundancy. In the CLPsych 2025 shared task, top systems used open-weight LLMs with in-context learning, including retrieval-augmented example selection (Antony and Schoene, 2025) and structured prompt design (Chan et al., 2025).

### 2.3 Task and Data

CLPsych 2026 Shared Task 1 focuses on post-level identification of adaptive and maladaptive psychological self-states. We address both subtasks.

**Task 1.1 (Subelement Classification).** Systems predict the composition of the adaptive and/or maladaptive self-state expressed in each post. A self-state is represented by selected subelements from six ABCD elements: Affect (A), Behavior toward Self (B-S), Behavior toward Others (B-O), Cognition toward Self (C-S), Cognition toward Others (C-O), and Desire (D). Each subelement has a predefined valence, either adaptive or maladaptive, and at most one subelement can be selected for each element and valence.

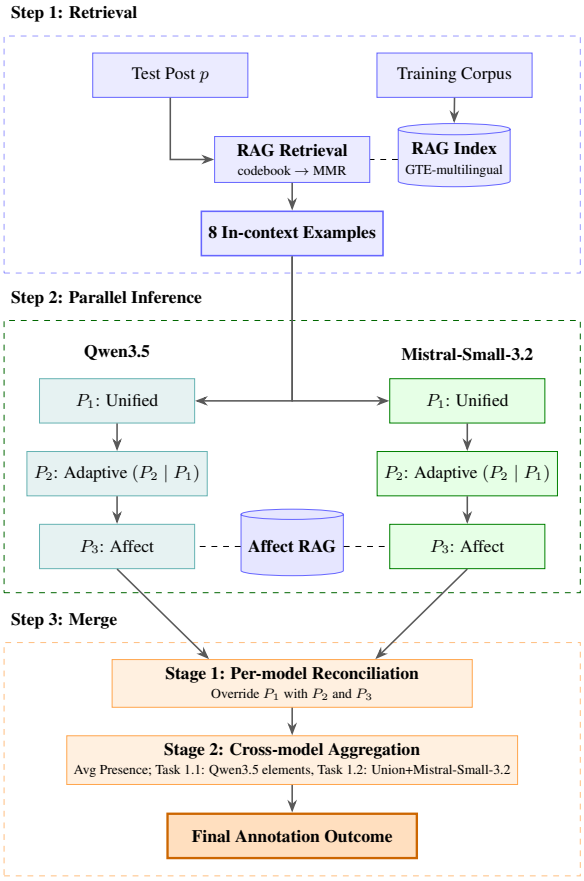


Figure 1: System architecture with the RAG and LLM ensemble.

**Task 1.2 (Presence Rating).** Systems assign separate presence scores to the adaptive and maladaptive self-states on a scale from 1 (not present) to 5 (highly present). The score reflects how central the self-state is to the post.

**Dataset.** The shared task dataset is an extended subset of the CLPsych 2022 "Reddit-New" corpus (Tsakalidis et al., 2022) and comprises 40 longitudinal social media timelines, partitioned into 30 for training (373 posts, 236 with evidence annotations) and 10 for testing (92 posts). The training set provides the post text and metadata (timeline ID, post ID, date, and chronological index), together with post-level MIND annotations: adaptive and maladaptive self-state classifications, ABCD subelements with textual evidence spans, and presence scores on the 1–5 scale. The Task 1.1 label space consists of 12 element–valence slots (6 elements $\times$ 2 valences) and 32 subelements in total; the full taxonomy is listed in Table 7. The test set provides

only the post text and the metadata required for evaluation.

## 3 Methods

Building on prior CLPsych systems that used retrieval-augmented in-context learning and structured prompting (Antony and Schoene, 2025; Chan et al., 2025), we adapt this setup to the fine-grained MIND label space by making example selection label-aware, separating prompts for error-prone parts of the taxonomy, and combining two open-weight LLMs with subtask-specific ensemble aggregation.

Figure 1 summarizes our pipeline. We use an ensemble of two open-weight LLMs: Qwen3.5-27B (Qwen Team, 2026) and Mistral-Small-3.2-24B-Instruct-2506 (Mistral AI, 2025), executed via vLLM (Kwon et al., 2023), with an in-context learning example corpus embedding using gte-multilingual-base (Zhang et al., 2024). For each post, we retrieve the 8 nearest in-context examples and run three sequential prompts per model; outputs are merged in two stages, with Task 1.1 keeping Qwen3.5's elements and Task 1.2 taking their union with Mistral-Small-3.2 arbitration on conflicts.

### 3.1 Retrieving In-Context Learning Examples

For every training post that carries an evidence annotation, we precompute a dense embedding using gte-multilingual-base (Zhang et al., 2024) and store these vectors in an in-memory index. We additionally build a small codebook of 12 embeddings, one per (element, valence) pair, by encoding the concatenated subelement descriptions for that pair. At prediction time, for a target post $p$ we (i) score every candidate $c$ as an equal mix of text cosine similarity to $p$ and a codebook label score (which sums the similarities between $p$ and the codebook entries matching $c$'s gold elements), (ii) retain the top $K_{\text{pool}} = 20$ candidates, and (iii) greedily re-rank to $K = 8$ using a similarity-and-diversity criterion inspired by MMR (Carbonell and Goldstein, 1998), where diversity is measured as coverage of distinct subelement labels among the already-selected examples. Candidates from the target post's own timeline are excluded throughout (leave-one-timeline-out). The Affect-focused call (Section 3.2, P3) uses a variant of this procedure: candidates are restricted to training posts that carry an Affect annotation, the codebook is disabled, and the diversity objective targets coverage of Affect subelements specifically.

### 3.2 Prompting Strategy

For each post and for each of the two LLMs, we issue three prompts that include: task requirements, definitions of the adaptive and maladaptive valences and of the element inventory, the eight retrieved in-context examples, and explicit response-format guidelines. The three prompts differ in scope: P1 asks for everything at once, P2 narrows attention to the adaptive valence by conditioning on the maladaptive predictions of P1, and P3 narrows attention to a single element. Full templates are in Appendix C.

**(P1) Unified prompt.** A single prompt that asks the model to return, in a structured form, both valences' elements (with subelements) and both presence scores in one shot. The prompt is conditioned on the eight RAG examples from Section 3.1. The unified call lets the model use inter-element context when assigning subelements.

**(P2) Adaptive-focused prompt.** The maladaptive elements that the model predicted in P1 are injected into the prompt header, and the model is asked to return only the adaptive elements; the same eight RAG examples from P1 are reused. By making the maladaptive prediction explicit, the call shifts attention toward adaptive content that the unified call tends to under-predict.

**(P3) Affect-focused prompt.** A prompt that asks the model to return only the Affect element (in either valence). P3 uses element-specific RAG so that all eight in-context examples illustrate Affect annotations. Affect is the element with the largest and most semantically diverse subelement space; isolating it lets the model concentrate retrieval and reasoning on a single category.

### 3.3 Two-Stage Ensemble Merge

The six structured outputs (three prompts $\times$ two models) are combined in two stages.

**Stage 1: per-model reconciliation.** The three calls within each model are merged deterministically: maladaptive elements and both presence scores are taken from P1; adaptive elements are taken from P2; the Affect (A) element in either valence, when present in P3, overrides whatever P1 or P2 produced; if after this step a valence has no elements, its presence is reset to 1. This rule is identical for Qwen3.5 and Mistral-Small-3.2.

| Submission | Task 1.1 (Sub F1) ↑ | | | Task 1.2 (RMSE) ↓ | | |
|---|---|---|---|---|---|---|
| | Adapt. | Malad. | Avg | Adapt. | Malad. | Avg |
| Sub. 1 (Solo Qwen3.5, early RAG impl.) | **0.371** | 0.510 | 0.440 | 1.253 | 0.935 | 1.094 |
| Sub. 2 (Ensemble, Union+Mistral-Small-3.2) | 0.357 | 0.508 | 0.433 | **1.130** | **0.858** | **0.994** |
| Sub. 3 (Ensemble, Qwen3.5 elements) | 0.346 | **0.537** | **0.441** | 1.184 | **0.858** | 1.021 |
| Organizer baseline | 0.156 | 0.338 | 0.247 | 1.409 | 1.439 | 1.424 |

Table 1: Test-set scores for our submissions. **Bold** marks the best score among our submissions per metric.

**Stage 2: cross-model merge.** Presence scores and element predictions are aggregated separately for each valence.

**Presence (both tasks).** The two presence scores from Qwen3.5 and Mistral-Small-3.2 are averaged and rounded. When the two models agree, the score is unchanged. In the critical 1/2 boundary case (one model votes "not present", the other "somewhat present"), the merged score is 2, favoring presence detection.

**Elements: Task 1.1.** Take Qwen3.5's elements (and their subelements). The Task 1.1 metric (Macro F1, see Section 3.4) is sensitive to false positives, so we use only one model's elements rather than the union, accepting some recall loss in exchange for precision.

**Elements: Task 1.2.** Take the union of both models' elements; on shared elements with disagreeing subelement labels, prefer Mistral-Small-3.2. This design choice was based on preliminary development experiments, where Mistral-Small-3.2 tended to yield lower presence-rating error, whereas Qwen3.5 produced more reliable subelement predictions. The Task 1.2 metric (RMSE, see Section 3.4) is sensitive to missing elements (a missing element forces presence to 1, incurring a large quadratic penalty when the gold value is 4 or 5), so the union acts as a safety net.

### 3.4 Evaluation Metrics

The performance of each submission was evaluated using task-specific metrics specified by the shared task organizers. For Task 1.1, element presence is evaluated as 12 binary classifications (6 elements $\times$ 2 valences) using Precision, Recall, and F1, while subelement classification is evaluated via multi-class F1 per element. The official ranking metric is the mean of the Adaptive subelement Macro F1 and the Maladaptive subelement Macro F1.

For Task 1.2, the 1–5 presence ratings for the adaptive and maladaptive self-states are evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Quadratic Weighted Kappa (QWK), and Spearman correlation. The official ranking metric for this subtask is the mean of the Adaptive RMSE and the Maladaptive RMSE.

## 4 Results and Discussion

Our system was evaluated across two subtasks: subelement classification (Task 1.1) and presence rating (Task 1.2).

### 4.1 System Configurations

We submitted three configurations:

- **Sub. 1**: Qwen3.5 with an alternative RAG setup: utilizing all-MiniLM-L6-v2 as the underlying embedding model (Wang et al., 2020) and augmented training examples for the Affect-focused prompt.
- **Sub. 2**: Qwen3.5 + Mistral-Small-3.2 ensemble; merged elements take the union of both models, with Mistral-Small-3.2 winning on subelement conflicts (see Section 3.3 – Task 1.2).
- **Sub. 3**: same ensemble as Sub. 2, but merged elements come solely from Qwen3.5 (see Section 3.3 – Task 1.1).

Presence merging is identical for Sub. 2 and Sub. 3 and explained in Section 3.3. Test-set scores are reported in Table 1.

### 4.2 Task 1.1: Subelement Classification

Our Task 1.1 entry, Sub. 3, achieves **0.441 Avg F1**, placing 2nd of 17 teams on the official leaderboard, only 0.0004 below the leader. The following ablations show how each component of the pipeline contributes to this score.

**Prompt decomposition.** We hold the ensemble (Qwen3.5 + Mistral-Small-3.2) and MMR-8 retrieval fixed and vary the prompts described in

Section 3.2 (Table 2). The adaptive-conditioned prompt P2 contributes only a small gain over P1 alone, while the Affect-focused prompt P3 contributes the bulk of the improvement. The full three-call pipeline gains 0.029 Avg F1 over the single-call baseline.

**Retrieval strategy.** We compare three retrieval regimes within the full ensemble pipeline, holding the prompts and merging strategy fixed (Table 3). In-context examples are essential: removing RAG collapses Avg F1 (0.206 vs 0.441 Avg F1). One-step retrieval and MMR-8 re-ranking perform comparably (0.444 vs 0.441 Avg F1), suggesting that diversity re-ranking offers no measurable advantage on this test set.

**Ensemble merge.** We isolate the contribution of each individual model and of the merge strategy, holding the three-prompt pipeline and MMR-8 re-ranking fixed (Table 4). Qwen3.5 outperforms Mistral-Small-3.2 on subelement classification (0.427 vs 0.373 Avg F1), motivating Qwen3.5's role as the Task 1.1 element-extraction anchor. Consistent with this, the precision-oriented merge using Qwen3.5's elements (Sub. 3) achieves higher Avg F1 than the union-based Sub. 2 merge (0.441 vs 0.433; Table 1).

### 4.3 Task 1.2: Presence Rating

Our Task 1.2 entry, Sub. 2, achieves **0.994 Avg RMSE**, placing 5th of 17 teams on the official leaderboard. We analyze presence-rating performance using both the component ablations above and an additional Task 1.2-specific comparison of merge rules (Table 5).

**Prompt decomposition.** P2 does not help presence rating: adding it to P1 increases Avg RMSE from 1.015 to 1.035. The full three-call pipeline incurs only a small RMSE cost relative to P1 alone (1.021 vs 1.015), so the F1 gains discussed in Section 4.2 do not substantially hurt presence rating.

**Retrieval strategy.** Retrieval remains important for Task 1.2: removing in-context examples increases Avg RMSE from 1.021 to 1.958. One-step retrieval and MMR-8 re-ranking are effectively tied (1.020 vs 1.021).

**Ensemble merge.** The ensemble results support the asymmetric merge. Mistral-Small-3.2 achieves lower RMSE than Qwen3.5 (1.083 vs 1.153; Table 4), supporting its use as the Task 1.2 conflict arbiter. In the Task 1.2-specific merge ablation, the Sub. 2 union ensemble obtains the best RMSE (0.994), providing additional support for our submitted merge strategy among the alternatives tested (Table 5).

### 4.4 Valence Asymmetry

Across all configurations and metrics, our system performs better on the maladaptive valence than on the adaptive one. In our top Task 1.1 run (Sub. 3), the Maladaptive Subelement Macro F1 reaches 0.537 while the Adaptive Subelement Macro F1 stays at 0.346. The same pattern holds for presence rating: our top Task 1.2 run (Sub. 2) attains a Maladaptive RMSE of 0.858 versus an Adaptive RMSE of 1.130. On the maladaptive valence specifically, our Subelement Macro F1 ranks 1st and our RMSE ranks 2nd among all 17 participating teams, indicating that our pipeline is particularly effective at detecting and rating maladaptive states.

## 5 Conclusions

The system presented in this paper demonstrates the potential of retrieval-augmented open-weight LLMs for identifying psychologically grounded self-states in social media posts. We addressed CLPsych 2026 Task 1 with an ensemble that combines semantic and label-aware retrieval of in-context examples, targeted prompts for adaptive content and Affect, and subtask-specific aggregation across Qwen3.5-27B and Mistral-Small-3.2-24B-Instruct. This design was effective in the shared task: our system ranked 2nd of 17 teams on Task 1.1, 5th on Task 1.2, and 1st on maladaptive-valence subelement F1.

Our ablations show that retrieval is the strongest driver of performance, while prompt decomposition and ensemble aggregation address different error modes in the fine-grained MIND label space. Overall, these results suggest that open-weight LLMs can support clinically motivated annotation schemes when their predictions are grounded in related examples and guided by task-specific structure. At the same time, the sensitivity of performance to label sparsity, valence asymmetry, and merge strategy highlights the need for careful evaluation before such systems are used beyond shared-task settings.

## 6 Limitations

Despite competitive performance, our system exhibits several key limitations.

**Adaptive state blind spots.** As discussed in Section 4.4, our system remains weaker on adaptive than maladaptive self-states. Subtle adaptive signals such as coping, self-care, help-seeking, or constructive social engagement are still easier to miss, especially when they co-occur with more salient maladaptive content.

**Single-run test evaluation.** All test-set submissions were evaluated as single inference runs at $T = 0.1$ in the blind shared-task setting, rather than as repeated stochastic runs of the same configuration. Consequently, small differences across configurations should be interpreted with caution, as they may reflect stochastic decoding or submission-level variance in addition to true design effects (Reimers and Gurevych, 2017; Card et al., 2020).

**Limited model comparison.** We explored several open-weight LLMs (including Llama and Gemma variants) during development, but the final pipeline and ablations are based on a single model pair (Qwen3.5-27B and Mistral-Small-3.2-24B-Instruct). These models offered the best balance of following structured JSON instructions and feasible GPU memory use in preliminary tests, but we did not perform a systematic comparison across model families and sizes.

**Computational overhead.** Our pipeline runs three targeted prompts across two LLMs, requiring six retrieval-augmented inference passes per post. Although this design targets adaptive underprediction, Affect diversity, and cross-subtask robustness, it increases inference cost and latency and may hinder real-time use in high-volume clinical monitoring environments.

## Ethics

The CLPsych 2026 dataset is an extended subset of the CLPsych 2022 "Reddit-New" corpus (Tsakalidis et al., 2022), provided in de-identified form. Although Reddit is an anonymous platform, posts on mental health topics remain sensitive; we accessed the data only after signing the Data Access Agreement and complied with its restrictions on storage, sharing, and processing throughout the project.

In compliance with the shared task requirements, we used only open-weight LLMs, run locally; no closed-source models or external API services were used for experimentation or analysis on the shared task data.

Computational systems trained on annotated social media data should be regarded as supportive research tools rather than substitutes for clinical judgment. Real-world deployment would require validation by mental-health professionals, attention to demographic and cultural variation in how mental states are expressed, and safeguards against misuse. We also recognize that LLM-based systems may reflect biases present in their pretraining corpora; our in-context learning setup grounded in the MIND framework can mitigate some of this drift, but does not eliminate it.

To protect the privacy of the social media posters whose data was used, this work should not be reproduced on the original dataset by parties without an active Data Access Agreement; the contributions reported here are the design and analysis of our system, not the data itself.

## Acknowledgments

We thank Simon Clematide and Andrianos Michail for their supervision, feedback, and guidance during the development of this system.

## References

Iqra Ali, Talia Tseriotou, Guy Dvir, Callum Chan, Yuxiang Zhou, Juan Antonio Lossio-Ventura, Ayal Klein, Aya Shamir, Dan Sayda, Anthony Hills, Aya Zirikly, Diana Inkpen, Dana Atzil-Slonim, and Maria Liakata. 2026. Overview of the clpsych 2026 shared task: Capturing and characterizing mental health changes through social media timeline dynamics. In *Proceedings of the 11th Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Anson Antony and Annika M. Schoene. 2025. Retrieval-enhanced mental health assessment: Capturing self-state dynamics from social media using in-context learning. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 268–278. Association for Computational Linguistics.

Dana Atzil-Slonim. 2025. Multimodal intrapersonal and interpersonal dynamics (MIND): A transtheoretical coding manual.

Dana Atzil-Slonim. 2026. Leveraging theoretical and technological innovations to study the mechanisms that underlie therapeutic change in psychotherapy. In Louis G. Castonguay, Dana Atzil-Slonim, Michael Barkham, and Wolfgang Lutz, editors, *Practice-Based Evidence in the Psychological Therapies: Toward Policy Implications for Research, Training, and*

*Clinical Guidelines*. Oxford University Press, New York.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274. Association for Computational Linguistics.

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

Zhijun Guo, Alvina Lai, Johan H. Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large language models for mental health applications: Systematic review. *JMIR Mental Health*, 11:e57400.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Matteo Malgaroli, Katharina Schultebraucks, Keris Jan Myrick, Alexandre Andrade Loch, Laura Ospina-Pinillos, Tanzeem Choudhury, Roman Kotov, Munmun De Choudhury, and John Torous. 2025. Large language models for the mental health community: framework for translating code to care. *The Lancet Digital Health*, 7(4).

Mistral AI. 2025. Mistral-Small-3.2-24B-Instruct-2506. `https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506`. Model card. Accessed: 2026-05-19.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348. Association for Computational Linguistics.

William Revelle. 2007. Experimental approaches to the study of personality. In Richard W. Robins, R. Chris Fraley, and Robert F. Krueger, editors, *Handbook of Research Methods in Personality Psychology*, pages 37–61. Guilford Press, New York.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Medical Informatics*, 12(1):e55318.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, Seattle, USA. Association for Computational Linguistics.

Talia Tseriotou, Jenny Chim, Ayal Klein, Aya Shamir, Guy Dvir, Iqra Ali, Cian Kennedy, Guneet Singh Kohli, Anthony Hills, Ayah Zirikly, Dana Atzil-Slonim, and Maria Liakata. 2025. Overview of the CLPsych 2025 shared task: Capturing mental health

dynamics from social media timelines. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 193–217, Albuquerque, New Mexico. Association for Computational Linguistics.

Ahmet Yavuz Uluslu, Andrianos Michail, and Simon Clematide. 2024. Utilizing large language models to identify evidence of suicidality risk through analysis of emotionally charged posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 264–269, St. Julians, Malta. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. MentaLLaMA: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*, New York, NY, USA. Association for Computing Machinery.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

# A Ablation Tables

| Configuration | Avg F1 ↑ | Avg RMSE ↓ |
|---|---|---|
| P1 only | 0.412 | **1.015** |
| P1 + P2 | <u>0.414</u> | 1.035 |
| P1 + P2 + P3 (Sub. 3) | **0.441** | <u>1.021</u> |

Table 2: Ablation 1: prompt decomposition, full ensemble (Qwen3.5 + Mistral-Small-3.2) with MMR-8. P1 is the unified prompt, P2 adds the adaptive-focused prompt, and P3 adds the Affect-focused prompt. **Bold** marks the best score in each column; <u>underline</u> marks the runner-up.

| Configuration | Avg F1 ↑ | Avg RMSE ↓ |
|---|---|---|
| Zero-shot | 0.206 | 1.958 |
| One-step retrieval | **0.444** | **1.020** |
| MMR-8 (Sub. 3) | <u>0.441</u> | <u>1.021</u> |

Table 3: Ablation 2: retrieval strategy, full ensemble with all three prompts. Zero-shot removes in-context examples, one-step retrieval directly selects the eight highest-scoring examples, and MMR-8 additionally reranks a candidate pool for label diversity.

| Configuration | Avg F1 ↑ | Avg RMSE ↓ |
|---|---|---|
| Mistral-Small-3.2 | 0.373 | 1.083 |
| Qwen3.5 | 0.427 | 1.153 |
| Ensemble (Sub. 2) | **0.433** | **0.994** |

Table 4: Ablation 3: ensemble strategy, full three-prompt pipeline with MMR-8. Single-model rows use only that model's outputs; the ensemble row combines Qwen3.5 and Mistral-Small-3.2 with Sub. 2 cross-model aggregation rules.

| Configuration | Avg F1 ↑ | Avg RMSE ↓ |
|---|---|---|
| Confidence-based merge | 0.386 | 1.467 |
| Aggressive-presence merge | **0.434** | <u>1.0</u> |
| Union ensemble (Sub. 2) | <u>0.433</u> | **0.994** |

Table 5: Ablation 4: ensemble merge approaches for presence score, full three-prompt pipeline with MMR-8. Confidence-based merge uses model confidence in cross-model selection, aggressive-presence merge favors non-1 presence predictions from either model, and the union ensemble is Sub. 2 averaging-based merge.

# B Hyperparameter Settings

| Parameter | Value |
|---|---|
| *Inference* | |
| Weight precision | FP16 |
| Max context length | 16,384 tokens |
| Sampling temperature | 0.1 |
| Sampling top-$p$ | 1.0 (vLLM default) |
| Max generation tokens | 100 |
| *Retrieval* | |
| Candidate pool size | $K_{\text{pool}} = 20$ |
| In-context examples | $K = 8$ |
| Diversity weight | $\alpha = 0.6$ |
| *Other* | |
| Post text truncation | 1,800 characters |

Table 6: Hyperparameter settings for our submitted system.

## C Prompt Templates

We reproduce the actual prompt templates used in our pipeline. The static *ABCD taxonomy block* (which lists element definitions, valence definitions, and all 32 subelement descriptions) is abridged here for space; it remains constant across the three prompts. The example post shown below is generated for illustration, and no actual dataset content is reproduced.

### *(P1) Unified prompt*

**System:**
You are a clinical psychology annotation assistant trained in the MIND framework. The MIND framework analyzes social media posts to identify psychological self-states. Each post may express adaptive and/or maladaptive psychological states across 6 dimensions: Affect (A), Behavior toward Others (B-O), Behavior toward Self (B-S), Cognition about Others (C-O), Cognition about Self (C-S), and Desire (D). Your job: identify which specific subelements are present and rate their overall presence.

[ABCD TAXONOMY BLOCK — see Appendix D]

PRESENCE SCALE (reflects psychological centrality, not mere frequency of words):
1 = Not present
2 = Somewhat present
3 = Moderately present
4 = Much present
5 = Highly present

RULES:
1. For each ABCD element (A, B-O, B-S, C-O, C-S, D), decide if it is present.
2. If present, choose the single most prominent subelement.
3. Assign ADAPTIVE_PRESENCE and MALADAPTIVE_PRESENCE scores (1–5).
4. Output ONLY these 4 lines: ADAPTIVE_ELEMENTS, ADAPTIVE_PRESENCE, MALADAPTIVE_ELEMENTS, MALADAPTIVE_PRESENCE. No evidence lines, no extra text.
5. If no adaptive elements: write ADAPTIVE_ELEMENTS: NONE and ADAPTIVE_PRESENCE: 1.
6. If no maladaptive elements: write MALADAPTIVE_ELEMENTS: NONE and MALADAPTIVE_PRESENCE: 1.
7. Only annotate elements you are confident about. It is better to miss an element than to predict the wrong subelement.

**User:**
Here are 8 annotated examples:
[examples]

Now annotate the following post:

POST: "I haven't been able to focus on anything lately. Even small tasks feel overwhelming. I keep thinking that I should be doing better by now, which only makes me feel worse. But yesterday I called my friend, and talking to her helped a bit. I still feel exhausted, but at least I managed to take a shower and eat something afterward."

### *(P2) Adaptive-focused prompt*

**System:**
You are a clinical psychology annotation assistant trained in the MIND framework. Your task: identify ADAPTIVE psychological patterns that co-exist alongside maladaptive states in this post. Adaptive states often co-exist with maladaptive states. A person describing depression may also show self-care, self-acceptance, healthy grieving, or reaching out for support.

[ABCD ADAPTIVE TAXONOMY BLOCK — see Appendix D, adaptive subelements only]

[PRESENCE SCALE — as above]

RULES:
1. Focus ONLY on ADAPTIVE elements. Maladaptive annotation is already done.
2. For each element, decide if an adaptive subelement is present.
3. Assign ADAPTIVE_PRESENCE (1-5).
4. Output ONLY these 2 lines: ADAPTIVE_ELEMENTS and ADAPTIVE_PRESENCE.
5. If no adaptive elements: ADAPTIVE_ELEMENTS: NONE and ADAPTIVE_PRESENCE: 1.
6. For B-S and C-S: predict whenever you see ANY signal.
7. For A, B-O, C-O, D: only annotate if confident in the specific subelement.

**User:**
Here are 8 annotated examples:
[examples]

The following maladaptive patterns were already identified:
MALADAPTIVE_ELEMENTS: A:8, B-S:2
MALADAPTIVE_PRESENCE: 4

Now identify ADAPTIVE elements in this post:

POST: "[same example post as P1]"

### *(P3) Affect-focused prompt*

**System:**
You are a clinical psychology annotation assistant trained in the MIND framework.

[ABCD TAXONOMY BLOCK — see Appendix D, Affect (A) row only]

RULES:
1. ONLY evaluate the Affect (A) element. Decide if it is present.
2. If present, choose the single most prominent subelement.
3. Output ONLY these 2 lines: ADAPTIVE_ELEMENTS and MALADAPTIVE_ELEMENTS. No presence scores.
4. If adaptive A is not present: write ADAPTIVE_ELEMENTS: NONE.
5. If maladaptive A is not present: write MALADAPTIVE_ELEMENTS: NONE.
6. Only annotate if you are confident.

**User:**
Here are 8 annotated examples (all illustrating Affect annotations):
[examples]

Now annotate the following post:

POST: "[same example post as P1]"

## D ABCD Taxonomy Block

| Element | Valence | ID | Subelement |
|---|---|---|---|
| A | Adaptive | 1 | Calm/laid back — feeling at ease, relaxed, or peacefully accepting of one's situation |
| | Adaptive | 3 | Sad/emotional pain/grieving — expressing sorrow or grief in a healthy, processing way |
| | Adaptive | 5 | Content/happy/joy/hopeful — feeling satisfied, cheerful, or optimistic |
| | Adaptive | 7 | Vigor/energetic — feeling motivated, lively, or full of energy |
| | Adaptive | 9 | Justifiable/assertive anger — expressing appropriate anger or standing up for oneself |
| | Adaptive | 11 | Proud — feeling a sense of achievement, self-worth, or pride |
| | Adaptive | 13 | Feel loved/belong — feeling connected, valued, or accepted by others |
| | Maladaptive | 2 | Anxious/fearful/tense — feeling worried, nervous, scared, or on edge |
| | Maladaptive | 4 | Depressed/despair/hopeless — feeling deeply sad, empty, or worthless |
| | Maladaptive | 6 | Mania — showing elevated mood, grandiosity, impulsivity beyond normal excitement |
| | Maladaptive | 8 | Apathetic/blunted — feeling emotionally flat, numb, or unable to care |
| | Maladaptive | 10 | Angry/aggression/disgust/contempt — feeling hostile, bitter, or resentful |
| | Maladaptive | 12 | Ashamed/guilty — feeling intense shame, self-blame, or disproportionate guilt |
| | Maladaptive | 14 | Feel lonely — feeling isolated, disconnected, or painfully alone |
| B-O | Adaptive | 1 | Relating behavior — actively reaching out, connecting, or cooperating with others |
| | Adaptive | 3 | Autonomous or adaptive control behavior — taking independent action or setting healthy boundaries |
| | Maladaptive | 2 | Fight or flight behavior — reacting with aggression, withdrawal, or avoidance driven by fear |
| | Maladaptive | 4 | Over-controlled or controlling behavior — excessively managing, manipulating, or dominating interactions |
| B-S | Adaptive | 1 | Self-care and improvement — taking steps to look after oneself physically, mentally, or emotionally |
| | Maladaptive | 2 | Self-harm, neglect, and avoidance — self-destructive behavior, ignoring needs, or avoiding self-care |
| C-O | Adaptive | 1 | Perception of other as related — viewing others as caring, supportive, or emotionally connected |
| | Adaptive | 3 | Perception of other as facilitating autonomy — seeing others as encouraging independence |
| | Maladaptive | 2 | Perception of other as detached or over-attached — seeing others as emotionally unavailable or intrusive |
| | Maladaptive | 4 | Perception of other as blocking autonomy — viewing others as controlling or restricting |
| C-S | Adaptive | 1 | Self-acceptance and compassion — treating oneself with kindness or accepting flaws without harsh judgment |
| | Maladaptive | 2 | Self-criticism — harshly judging, blaming, or devaluing oneself |
| D | Adaptive | 1 | Relatedness — wanting to connect, bond, or maintain meaningful relationships |
| | Adaptive | 3 | Autonomy and adaptive control — desiring independence, self-direction, or healthy control |
| | Adaptive | 5 | Competence/self-esteem/self-care — wanting to feel capable, worthy, or to take good care of oneself |
| | Maladaptive | 2 | Expectation that relatedness needs will not be met — believing one will remain unloved or rejected |
| | Maladaptive | 4 | Expectation that autonomy needs will not be met — believing one will remain trapped or powerless |
| | Maladaptive | 6 | Expectation that competence needs will not be met — believing one will remain incompetent or worthless |

**Table 7:** Full ABCD subelement taxonomy used in the prompt templates.
Abbreviations: A = Affect, B-O = Behavior toward Others, B-S = Behavior toward Self, C-O = Cognition of the Other, C-S = Cognition of the Self, D = Desire.