P2P - from Posts to Patterns: An LLM Ensemble Approach to Mental Health Dynamics Detection

Federico Ravenda[★], **Volodymyr Karpenko**[★]
Antonietta Mira[★,◇], **Andrea Raballo**[★,♣]
{name.lastname}@usi.ch,
[★] Euler Institute, Università della Svizzera italiana
[♣] Cantonal Socio-Psychiatric Organization , [◇] University of Insubria

Abstract

The automatic detection of mental health dynamics represents a challenging and socially relevant task, requiring models to capture subtle psychological changes over time. Large language models have demonstrated remarkable capabilities in clinical and diagnostic contexts, yet their outputs often lack consistency and may diverge significantly across different model families and prompting strategies. This paper presents the USAI team's submission to the CLPsych 2026 Shared Task, targeting Tasks 1.1, 1.2, 2, and 3.1. We propose an ensemble-based approach combining multiple open-source large language models, where the contribution of each model is weighted according to its alignment with clinically grounded human annotations on the training set. Our system achieves competitive results across the evaluated subtasks, with particularly strong performance on Tasks 1.2 and 2.

1 Introduction & Related Works

Recent advances in NLP have enabled remarkable progress in sensitive domains such as digital health and mental health, where the ability to extract meaningful signals from unstructured text is of critical clinical value. The growing availability of social media data has further opened new avenues for passive, longitudinal monitoring of psychological states at scale.

Large language models have shown increasing promise in mental health research. For instance, MentaLLaMA (Yang et al., 2024a) demonstrates how interpretability can be integrated into model outputs to provide clinically meaningful insights, while Varadarajan et al. (2024) combine theoretical psychological frameworks with computational modelling to advance suicide risk assessment. Together, these efforts illustrate the growing potential of LLM-based systems to support real-time mental health monitoring and intervention (Yang et al., 2024b). Ravenda et al. (2025a) further explored the use of LLMs to enhance symptom severity assessment across different mental health conditions.

Beyond clinical monitoring, LLMs have recently gained traction as diagnostic and educational tools. Studies such as Elyoseph et al. (2025); Levkovich et al. (2024); Raballo et al. have shown that LLMs can match or even surpass physicians in diagnostic reasoning based on standardised clinical vignettes, highlighting their potential as reliable aids in medical education and decision support.

The CLPsych 2026 Shared Task (Ali et al., 2026) extends the previous year shared task (Tseriotou et al., 2025) by introducing a fine-grained longitudinal benchmark (Tsakalidis et al., 2022) grounded in the MIND framework (Atzil-Slonim, 2025, 2026), which conceptualises self-states as structured combinations of Affect, Behaviour, Cognition, and Desire (ABCD) components. Participants are asked to model adaptive and maladaptive self-state dynamics over time from sequences of social media posts, with the goal of characterising psychological change processes at both the post and timeline level.

The main contributions of this work are:

- We propose a weighted ensemble of open-source LLMs for ABCD subelement classification, self-state presence rating, and moment-of-change detection, covering Tasks 1.1, 1.2, 2, and 3.1 of the shared task.
- We introduce a training-time weight optimisation procedure that aligns each model's contribution with clinically grounded human annotations, independently for each task and metric.

2 Methodology

2.1 Overview

Our system takes inspiration from the CLPsych 2025 submission by Ravenda et al. (2025b) and

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 498–503

follows a weighted ensemble paradigm designed to exploit the capabilities of large language models without any parameter fine-tuning. Fine-tuning is indeed impractical in this setting for two reasons: the limited number of available training timelines (30) makes it prone to overfitting, and the computational cost of adapting large-scale models is prohibitive. Instead, each LLM is prompted for the specific task, and its predictions are treated as fixed inputs to a lightweight optimisation stage. This is conceptually analogous to stacking in ensemble learning: rather than combining model outputs through a learned meta-classifier, we directly optimise a set of scalar weights over the simplex to maximise the target evaluation metric on the training set. The pipeline consists of three stages: (i) independent inference from each LLM using task-tailored prompting strategies, (ii) weight learning via constrained optimisation on training-set predictions, and (iii) weighted ensemble decoding at test time. In Figure 1, we illustrate the three stages of the proposed pipeline: independent LLM inference, weight optimisation on the training set, and weighted ensemble decoding at test time.

2.2 Models and Prompting Strategies

We employ six open-source LLMs: DeepSeek-R1-Distill-Qwen-32B, Kimi Moonlight-16B-A3B-Instruct, GPT-OSS-20B, Gemma-27B , Xiaomi MiMo-v2-Flash, and Llama-3.1-8B.

Each model is assigned the prompting strategy, either *zero-shot* or *chain-of-thought*, that achieves the highest individual performance on the training set.

All prompts include the complete ABCD subelement schema of the task guidelines.

2.3 Task 1.1: ABCD Subelement Classification

For each post, the model is asked to identify the single most dominant adaptive and maladaptive subelement within each ABCD element (Affect, Behavior-Self, Behavior-Other, Cognition-Self, Cognition-Other, Desire), following the definitions in Atzil-Slonim (2025). The prompt enumerates all valid subelement indices for each element and valence, and constrains the output to at most one subelement per element per valence.

At ensemble time, subelement predictions across the six models are combined via soft weighted voting: for each post and each element–valence pair, each model contributes its predicted subelement index with a weight w_m. The index receiving the highest total weight is selected as the final prediction, provided it exceeds the total weight assigned to abstention (i.e. no subelement predicted).

The per-element weights $\mathbf{w} \in \Delta^5$ (the probability simplex) are found by solving:

$$\mathbf{w}^* = \mathbf{w} \in \Delta^5 \arg\max\, F_1(\hat{y}(\mathbf{w}),\, y) \qquad (1)$$

where $\hat{y}(\mathbf{w})$ is the ensemble decision under weights $\mathbf{w}$ and y is the gold label. Since F_1 is non-differentiable with respect to $\mathbf{w}$ due to the hard voting threshold, we adopt a derivative-free search strategy. We evaluate the objective at multiple candidate weight vectors, iteratively refining the search region towards configurations that improve training-set F1. To mitigate the risk of converging to poor local solutions, the search is repeated from $N{=}10$ independent starting points sampled uniformly at random from the simplex, and the best solution found across all runs is retained.

2.4 Task 1.2: Self-State Presence Rating

Each model assigns a presence score in $\{1, \ldots, 5\}$ to the adaptive and maladaptive self-states independently. The ensemble prediction is the weighted average of model scores, clipped to $[1, 5]$:

$$\hat{p} = \mathrm{clip}\left(\sum_m w_m \cdot p_m,\ 1,\ 5\right) \qquad (2)$$

Separate weight vectors are learned for adaptive and maladaptive presence by minimising the following objective on the training set:

$$\mathbf{w}^* = \mathbf{w} \in \Delta^5 \arg\min\, \mathrm{RMSE}(\textstyle\sum_m w_m p_m,\ y) \qquad (3)$$

where p_m is the presence score predicted by model m and y is the gold annotation.

2.5 Task 2: Moments of Change Detection

For each post, each model independently predicts a Switch label (`S` or `0`) and an Escalation label (`E` or `0`), treating the two labels as independent binary classification problems. Each model is provided with the full chronologically ordered sequence of posts in the timeline, enabling it to assess wellbeing trajectory.

At ensemble time, the binary predictions are combined via weighted majority voting, using the same formulation as Task 1.1 (Equation 1). Separate weight vectors are optimised for Switch F1 and Escalation F1 independently.

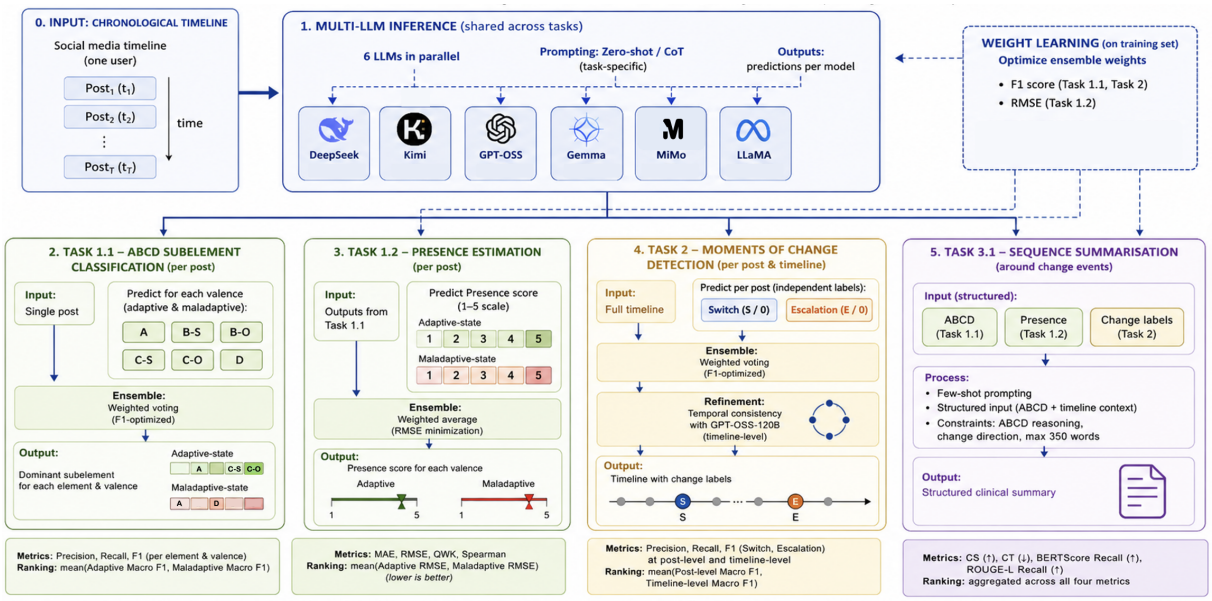

Figure 1: Overview of the proposed LLM ensemble pipeline for CLPsych 2026.

Finally, GPT-OSS-120B is employed in a refinement stage to enforce temporal consistency across the timeline: it handles missing wellbeing values by referring to prior posts within a defined temporal window, and revises predictions deemed clearly inconsistent with the surrounding context.

2.6 Weight Learning

All weight vectors are learned on the 30 training timelines provided for the shared task. A diversity check is performed prior to optimisation: if the standard deviation of per-model singleton F1 scores falls below a threshold $\epsilon{=}10^{-4}$, the weight vector defaults to uniform $\mathbf{w} = 1/M$ to avoid spurious solutions. Learned weights are fixed at test time; no further adaptation is performed.

2.7 Task 3.1: Sequence Summarisation

We address Task 3.1 through a few-shot prompting approach and `qwen-2.5-72b-instruct` model. For each test sequence, we construct a structured input by combining the ABCD self-state annotations from Task 1 (subelements and presence ratings for adaptive and maladaptive states) with the Switch and Escalation labels from Task 2, aligned to post indices. Up to five in-context demonstrations are selected from the training set to maximise diversity across change type (Switch, Escalation, or both) and direction of change (deterioration vs. improvement). One demonstration is always a fixed gold example from the task guidelines. The system prompt enforces ABCD element abbreviations, explicit identification of the change event and its direction, and coverage of within-state dynamics and cross-state interactions.

3 Results

Table 1 reports the performance of our system across all evaluated subtasks.

On Task 1.1, our system achieves competitive results, ranking second on Maladaptive Element Presence Macro F1, Average Element Presence Macro F1, and Adaptive Subelement Macro F1.

On Task 1.2, our system ranks first on Combined QWK (0.730) and Combined Spearman Correlation (0.752), and second on the majority of RMSE-based metrics, demonstrating strong alignment with human presence annotations.

Task 2 represents our strongest result overall, with our system ranking first on Combined Macro F1 (0.600), Post Macro F1 (0.639), and Post Switch Macro F1 (0.583), and second on Timeline Switch and Timeline Escalation Macro F1.

On Task 3.1, our system achieves the best performance on ROUGE-L Recall (0.333) and BERTScore Recall (0.365), indicating strong semantic coverage in the generated summaries. However, lower Consistency and higher Contradiction scores suggest that the system prioritises content coverage over internal coherence. Incorporating consistency as an explicit constraint in the prompt, or applying a post-hoc coherence filtering step, represents a possible direction for future work.

Task	Metric	USAI
1.1 - Subelement Classification	Average Subelement Macro F1	0.410
	Adaptive Element Presence Macro F1	0.549
	Maladaptive Element Presence Macro F1	0.736
	Average Element Presence Macro F1	0.642
	Adaptive Subelement Macro F1	0.333
	Maladaptive Subelement Macro F1	0.487
1.2 - Presence Rating	Avg RMSE ↓	0.942
	Adaptive RMSE ↓	0.979
	Maladaptive RMSE ↓	0.905
	Combined RMSE ↓	0.943
	Combined QWK	**0.730**
	Combined Spearman	**0.752**
	Combined MAE ↓	0.681
2 - Moment of Change	Combined Macro F1	**0.600**
	Post Macro F1	**0.639**
	Post Switch Macro F1	**0.583**
	Post Escalation Macro F1	0.694
	Timeline Macro F1	0.561
	Timeline Switch Macro F1	0.510
	Timeline Escalation Macro F1	0.611
3.1 - Sequence Summarisation	Consistency Score ↑	0.681
	Contradiction ↓	0.849
	ROUGE-L Recall ↑	**0.333**
	BERTScore Recall ↑	**0.365**

Table 1: USAI system results across all evaluated subtasks. ↑ = higher is better, ↓ = lower is better. **Bold** = best across all teams; underline = second best.

4 Conclusion

In this paper, we presented the USAI submission to the CLPsych 2026 Shared Task on mental health change detection in social media timelines. We proposed a weighted ensemble of open-source large language models.

Our system achieved strong results across multiple subtasks, ranking first on Task 2 and obtaining top-two performance on the majority of Task 1.2 metrics. These results suggest that optimisation-based ensemble learning is an effective strategy for psychologically grounded NLP annotation tasks, where individual model outputs may be noisy or inconsistent.

The main limitation of our approach lies in its reliance on large-scale models that entail significant computational costs, which may limit practical deployment in clinical settings. A promising direction for future work is the distillation of temporal dynamics into smaller, more efficient models, reducing the dependence on large language models while preserving the ability to capture longitudinal psychological patterns.

5 Limitations

Our approach presents several limitations that should be acknowledged. First, the ensemble weights are optimised on the training set with limited number of training timelines (30) available for the shared task. Second, the effectiveness of the weight optimisation procedure is contingent on sufficient diversity among model predictions: when models converge to similar outputs, weights default to uniform, providing no advantage over a simple average. Third, our system operates at the post level for Tasks 1.1 and 1.2, without explicitly modelling dependencies between consecutive posts within a timeline, which may limit its ability to capture gradual psychological transitions. Finally, while our system achieves competitive performance, it remains largely a black-box ensemble: the learned weights indicate the relative contribution of each model, but do not provide insight into which linguistic or psychological features drive individual predictions. Interpretability is particularly critical in mental health applications, where clinicians and practitioners need to understand and trust model outputs before acting on them.

6 Ethics Statement

This work involves the analysis of social media posts from individuals discussing mental health topics, which raises important ethical considerations. All data used in this study was provided by the CLPsych 2026 shared task organisers under a formal data access agreement, and handled strictly in accordance with its terms. No additional data collection was performed, and no attempt was made to re-identify any individual from the provided anonymised dataset.

The posts processed in this work may contain sensitive disclosures, including expressions of suicidal ideation, self-harm, and severe psychological distress. We recognise the potential harm of misclassifying such content and emphasise that our system is intended solely as a research prototype and is not suitable for deployment in clinical or real-world settings without thorough validation by qualified mental health professionals.

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.

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.

Zohar Elyoseph, Inbar Levkovitch, Yuval Haber, and Yossi Levi-Belz. 2025. Using genai to train mental health professionals in suicide risk assessment: Preliminary findings. *The Journal of clinical psychiatry*, 86(3):24m15525.

Inbar Levkovich, Eyal Rabin, Michal Brann, and Zohar Elyoseph. 2024. Large language models outperform general practitioners in identifying complex cases of childhood anxiety. *Digital Health*, 10:20552076241294182.

Andrea Raballo, Federico Ravenda, and Antonietta Mira. Diagnosing schizophrenia spectrum disorders: Large language models (llms) vs. leading international psychiatrists (lips). *Psychiatry and Clinical Neurosciences*.

Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, Antonietta Mira, and Noriko Kando. 2025a. Are LLMs effective psychological assessors? leveraging adaptive RAG for interpretable mental health screening through psychometric practice. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8975–8991, Vienna, Austria. Association for Computational Linguistics.

Federico Ravenda, Fawzia-Zehra Kara-Isitt, Stephen Swift, Antonietta Mira, and Andrea Raballo. 2025b. From evidence mining to meta-prediction: a gradient of methodologies for task-specific challenges in psychological assessment. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 242–248, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Vasudha Varadarajan, Allison Lahnala, Adithya V Ganesan, Gourab Dey, Siddharth Mangalik, Ana-Maria Bucur, Nikita Soni, Rajath Rao, Kevin Lanning, Isabella Vallejo, and 1 others. 2024. Archetypes and entropy: Theory-driven extraction of evidence for suicide risk. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 278–291.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024a. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.

Minqiang Yang, Yongfeng Tao, Hanshu Cai, and Bin Hu. 2024b. Behavioral information feedback with large

language models for mental disorders: Perspectives and insights. *IEEE Transactions on Computational Social Systems*, 11(3):3026–3044.