# Team Aurevia at CLPsych 2026: Local Healthcare NLP for Schema-Constrained Self-State Modeling

**Nathan Roll**[1], **Irene Yi**[1], **Sufian Aldogom**[2], **Grace Brown**[1], **Eric Basile**[4], **Isaac Gutterman**[3], **Lakshika Tennakoon**[1], **Ammar Ahmed**[4]
[1]Stanford University [2]Massachusetts Institute of Technology
[3]Harvard College [4]Aurevia
nroll@stanford.edu

## Abstract

Team Aurevia introduces a local open-weight healthcare NLP system for the CLPsych 2026 Shared Task, predicting MIND-coded self-state elements, moments of change, summaries, and dynamic signatures from social media timelines. The task is difficult because coarse presence, fine-grained ABCD subelements, and timeline-level change require different longitudinal evidence over privacy-sensitive mental-health language. Our system combines TF-IDF retrieval, schema-constrained local Qwen2.5 prompting, ordinal calibration, and conservative post-processing. Among official runs, Aurevia ranked 3rd of 17 for Task 1.2 presence prediction, 5th of 13 overall for Task 3.1, 1st on Task 3.1 consistency, and 2nd of 9 for MIND-coded deterioration signatures, showing that constrained local LLM pipelines can remain competitive in sensitive healthcare NLP while reducing reliance on hosted proprietary inference.

## 1 Introduction

Healthcare NLP systems for mental-health language must support structured interpretation without turning social-media posts into diagnoses or risk scores. The CLPsych 2026 Shared Task asks systems to model longitudinal self-state dynamics through the MIND framework, which represents self-states using Affect, Behavior, Cognition, and Desire annotation categories rather than diagnostic constructs (Ali et al., 2026; Atzil-Slonim, 2025, 2026). The challenge is not only identifying evidence in one post, but determining how that evidence changes, persists, or reverses across a person's timeline.

Team Aurevia tests whether a privacy-conscious local pipeline can remain competitive without hosted proprietary inference over raw task data. Our contribution is a compact combination of TF-IDF retrieval, schema-constrained JSON prediction, local MLX inference with a 4-bit Qwen2.5-72B-Instruct checkpoint, ordinal calibration, and validators/post-processors that keep outputs inside official formats. We show that these controls were most useful for coarse ordinal presence and low-contradiction summaries, while fine-grained subelements, improvement signatures, and temporal labels still require stronger within-person grounding.

## 2 Task and Data

The shared task contains five evaluated components over longitudinal social media timelines (Ali et al., 2026). Task 1.1 identifies adaptive and maladaptive ABCD subelements in each post; Task 1.2 assigns adaptive and maladaptive presence ratings on a 1–5 ordinal scale; Task 2 detects sudden Switches and gradual Escalations; Task 3.1 generates summaries for organizer-provided post sequences; and Task 3.2 identifies recurrent dynamic signatures of MIND-coded deterioration and improvement for human evaluation. Together, these tasks test both local evidence extraction and higher-level reasoning over temporal patterns.

We do not reproduce user posts, prompts containing raw post text, or model outputs that could identify a user. The data are sensitive social-media timelines, and examples can remain identifying even when usernames are removed. All qualitative discussion below therefore uses labels, elements, aggregate behavior, and official scores only.

## 3 Related Work

This task extends a CLPsych line of work on longitudinal mental-health language. The 2022 shared task introduced temporally sensitive evaluation for Switches and Escalations in user timelines (Tsakalidis et al., 2022); the 2025 task broadened the setting to adaptive and maladaptive self-state evidence, well-being scores, and post- and timeline-level sum-



*Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026)*, pages 531–534

marization (Tseriotou et al., 2025). CLPsych 2026 keeps this longitudinal focus but grounds the labels in MIND-coded ABCD dynamics (Ali et al., 2026; Atzil-Slonim, 2025). Team Aurevia follows this trajectory with a local, schema-constrained LLM pipeline rather than a hosted generative workflow.

## 4 System

Across tasks, the system retrieves similar training examples, generates or scores task-specific outputs, enforces official schemas, calibrates predictions where needed, and applies conservative post-processing before submission.

**Task 1.** For each post and valence, the system predicts a `Presence` rating in $\{1, \ldots, 5\}$ and optional ABCD subelements. The selected run ensembled a non-neural predictor with LLM candidates. The non-neural path uses TF-IDF unigrams/bigrams with 5,000 features, nearest-neighbor votes, and keyword lists for valid valence-element-subelement triples. The LLM path uses local MLX inference with `mlx-community/Qwen2.5-72B-Instruct-4bit` (Yang et al., 2024), retrieves three similar posts as few-shot examples, uses greedy decoding, and requests JSON-only ABCD predictions. Deterministic repair normalizes keys, replaces invalid subelements with frequent training subelements for the predicted valence/element, clamps presence scores, filters mundane posts, and inserts missing states. Presence calibration blends model values with training/development element-count means using weights 0.30 and 0.70.

**Task 2.** Task 2 predicts Switch and Escalation separately using handcrafted temporal signals, Task 1 ABCD profiles, and LLM confidence outputs with a rule-based calibrator. Signals include polarity changes, sudden- and gradual-change keywords, distances between consecutive ABCD profiles, increases in active maladaptive elements, disappearance of previously adaptive evidence, and newly appearing maladaptive evidence. Confidence values are engineering scores, not calibrated probabilities of clinical worsening or risk. The selected run normalized outputs to official labels, forced Switch to `0` for first posts, and selected top-scoring posts after rate matching to training Switch/Escalation rates.

**Task 3.** For Task 3.1, the system retrieves two demonstrations by TF-IDF similarity over concatenated sequence text, attaches Task 1 profiles when available, uses greedy decoding, removes preambles, caps summaries at 350 words, and softens categorical absence, certainty, and exclusivity language. This favors cautious summaries over summaries that maximize reference overlap. For Task 3.2, the system records adjacent-post transitions in training timelines, prioritizes element-transition combinations appearing in at least two timelines where possible, and formats one deterioration and one improvement signature with supporting post-pair sequence identifiers but no raw post text.

## 5 Experimental Setup

Team Aurevia reports one official selected run for each evaluated component. Table 1 gives the official Codabench submission IDs for Tasks 1–3.1 and labels the Task 3.2 email submissions. Run T1 is submission 661781 for Tasks 1.1/1.2, T2 is submission 653243 for Task 2, T3 is submission 693454 for Task 3.1, and T4-D/T4-I are the Task 3.2 deterioration and improvement email submissions. Calibration and threshold choices used released training/development resources and local held-out validation outputs. The official organizer results workbook was used only to report ranks and identify Aurevia rows. Official submissions were deterministic after fixed retrieval, decoding, calibration, and post-processing choices.

## 6 Results and Analysis

Table 1 shows balanced performance across structured and generative tasks. Aurevia ranked 3rd of 17 for calibrated presence prediction, 5th of 17 for subelement classification, 7th of 18 for change detection, 5th of 13 for sequence summarization, and 2nd of 9 for deterioration signatures. The strongest results align with the system design: schema-constrained prediction and calibration were most useful for coarse ordinal presence or conservative aggregate signatures.

Task 1 shows the clearest contrast. Presence prediction achieved 0.981 average RMSE, with quadratic weighted kappa 0.645, Spearman correlation 0.674, and MAE 0.688. In Task 1.1, element-presence macro F1 was 0.546 for adaptive and 0.680 for maladaptive states, while subelement macro F1 was lower at 0.328 and 0.434. This matters because clinical-style interpretation often requires calibrated evidence strength even when fine-grained category boundaries remain ambigu-

| Task | Run | Primary metric (direction) | Score | Rank |
|---|---|---|---|---|
| 1.1 Subelements | 661781 | Avg. subelement macro F1 ↑ | 0.381 | 5 / 17 |
| 1.2 Presence | 661781 | Avg. adaptive/maladaptive RMSE ↓ | **0.981** | **3 / 17** |
| 2 Change detection | 653243 | Post/timeline macro F1 ↑ | 0.484 | 7 / 18 |
| 3.1 Summaries | 693454 | Score rank average ↓ | 6.25 | 5 / 13 |
| 3.2 Deterioration | T4-D | Human overall score ↑ | **0.676** | **2 / 9** |
| 3.2 Improvement | T4-I | Human overall score ↑ | 0.465 | 6 / 9 |

Table 1: Team Aurevia's selected official submissions. Lower is better for RMSE and Task 3.1 score-rank average; higher is better otherwise.

ous.

Task 2 exposes the temporal bottleneck. The selected submission reached 0.484 combined macro F1: post macro F1 was 0.556, while timeline macro F1 was 0.413. The system was therefore better at identifying local evidence than at deciding exactly where a person's trajectory changed.

Task 3.1 shows a summary tradeoff. Aurevia ranked 5th overall, with the best consistency score (0.866, rank 1) and second-best contradiction score (0.625, rank 2), but lower lexical and embedding recall. Contradiction-aware cleanup helped factual caution but left reference details uncovered. Task 3.2 similarly shows that deterioration signatures were easier than improvement signatures: deterioration ranked 2nd with overall score 0.676, while improvement ranked 6th with overall score 0.465. Improvement signatures may require richer context than transition counts alone provide.

## 7 Discussion

The main strength of the submission is not a single model choice, but the control stack around local generation. Retrieval narrows the evidence context, JSON prompting exposes format errors, repair keeps outputs legal under the official schema, and calibration tempers ordinal presence scores when the predicted evidence structure is sparse. These controls are especially useful in healthcare NLP because misleading certainty can matter even when a system is used only for shared-task research.

The main remaining weakness is temporal grounding. The system can often detect local self-state evidence, but Switches, Escalations, and improvement signatures require a stronger account of what changed relative to a user's earlier baseline. Future systems should therefore combine calibrated post-level prediction with timeline-level models that represent persistence, reversal, uncertainty, and missing context explicitly.

## 8 Conclusion

Team Aurevia's CLPsych 2026 submission demonstrates that a local open-weight, schema-constrained healthcare NLP pipeline can be competitive for coarse self-state presence prediction and conservative summarization without sending raw task text to hosted proprietary APIs. Its weaker performance on subelement selection, improvement signatures, and timeline-level change detection points to the central remaining challenge: modeling self-state dynamics as within-person trajectories rather than isolated post-level cues.

## Limitations

The system should not be used for diagnosis, treatment, crisis triage, surveillance, moderation, or intervention triggering. It predicts shared-task labels, not clinical truth. The labels depend on an annotation framework and limited training data; ambiguous posts, missing context, and annotator disagreement limit what the labels can mean. Social media posts are context-poor, performative, and culturally mediated, and may not map cleanly onto symptoms, impairment, risk, or need for care. Errors may also reflect cultural, demographic, linguistic, or platform-specific biases in both the data and the models.

Performance varied across domains. Aurevia performed well on presence detection and MIND-coded deterioration, but lower scores in secondary extraction and MIND-coded improvement detection (T4-I score 0.465; rank 6 of 9) suggest difficulty capturing positive clinical trajectories and fine-grained semantic distinctions. The pipeline also compounds errors across tasks: missed elements or miscalibrated presence scores can affect later predictions. We did not perform controlled ablations or clinical validation of the generated signatures, and we report only officially selected submissions.

## Ethics Statement

The data concern mental health and social media behavior, a setting that requires particular care around privacy, consent, and downstream use. Even de-identified posts may contain personal experiences that users did not write for clinical modeling. We therefore avoid quoting raw posts, avoid releasing derived examples that could identify users, avoid including raw prompts or raw post text, and describe errors only in aggregate. We relied on the organizer-provided shared-task access and data-use conditions, made no re-identification attempts, limited access to the author team's local shared-task workspace, and prepared outputs under the organizer rules. All processing was designed for the shared-task research setting. We did not build user-facing interventions, clinical alerts, automated monitoring, moderation tools, or individual risk profiles.

Team Aurevia did not obtain separate institutional review board approval for this secondary shared-task analysis. Our work used the organizer-provided shared-task materials under the competition's access and data-use conditions, and we did not collect new human-subject data or attempt to contact, identify, or profile individual users. Where language-model inference was required, we used local execution rather than sending raw task text to a hosted proprietary API. This reduces one third-party disclosure pathway but does not remove all risk: local files, logs, caches, generated artifacts, and open-weight models can still leak information, hallucinate, reproduce biases, and generate over-confident summaries.

## 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 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*, pages 157–186. Oxford University Press.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.