McMasters of Change: Predicting Wellbeing States and Transitions from Longitudinal Language

Hongyi Zhang[1*], Derron Li[1], Scarlett Cleary[1], Aadi Sanghani[1], Akshay Krishna Sirigana[1], Brian Miguel Pimentel[1], Kelsey Isman[2], Kian Omoomi[1], Vasudha Varadarajan[3], Charles Welch[1], Allison Lahnala[1*]

[1]McMaster University, Department of Computing and Software, Hamilton, ON, Canada
[2]Indiana University, Department of Psychological and Brain Sciences, Bloomington, IN, USA
[3]Carnegie Mellon University, Language Technologies Institute, Pittsburgh, PA, USA

[*]**Correspondence**: {zhanh279,lahnalaa}@mcmaster.ca

Abstract

Most existing work on mental health prediction from language focuses on isolated posts, overlooking temporal dynamics in longitudinal timelines. We present McMaster NLP's system for the CLPsych 2026 Shared Task, which centers on modeling mental health dynamics in social media timelines using the MIND framework (Atzil-Slonim, 2025). The task comprises: (1) identifying adaptive and maladaptive self-state components within posts, (2) detecting moments of change in wellbeing, and (3) generating structured summaries. For self-state prediction, we leverage LLM-generated archetypal representations of language use as semantic anchors within a dual-encoder architecture, enabling interpretable prediction of subelements and their intensities through alignment with prototypical expressions of psychological states. For temporal dynamics, we use BiLSTM-based sequence models to detect moments of change. For summarization, we employ a prompt-based LLM to generate grounded, structured summaries emphasizing causal interactions and temporal progression of self-states. Finally, we analyze model failure modes with respect to human evaluation and identify directions for reconciling the MIND framework with how state-assessment models encode meaning.

1 Introduction

Mental health disorders affect over one billion people worldwide, impacting reasoning, emotions, and social behaviour (Stein et al., 2020; World Health Organization, 2025). Early detection is critical to alleviate the severe consequences of these conditions, yet traditional methods for assessing mental health largely rely on face-to-face interviews, self-reporting, or questionnaires, which are limited by bias and barriers to care (Lin et al., 2017). Natural Language Processing (NLP) models of language on social media have the potential to overcome these limitations by leveraging behavioural and linguistic cues to model mental health at scale (Atzil-Slonim, 2026), but most prior work focuses on analyzing individual posts in isolation (Chancellor and De Choudhury, 2020). Longitudinal analysis of temporally ordered social media timelines remains comparatively understudied (Tsakalidis et al., 2022).

We present our submission to the CLPsych 2026 Shared Task (Ali et al., 2026), which addresses modeling mental health dynamics in social media timelines using the MIND framework (Atzil-Slonim, 2024, 2025). The task comprises: (1) identifying adaptive and maladaptive self-state components within posts, including dominant ABCD subelements and presence levels; (2) detecting moments of change in wellbeing, including both abrupt transitions (switches) and gradual progressions (escalations); and (3) generating structured summaries of self-state sequences surrounding change events. The task extends the prior year's shared task (Tseriotou et al., 2025) with finer-grained application of the MIND framework.

We propose a multi-stage framework combining representation learning, sequence modeling, and prompt-based summarization (Figure 1). For Task 1, we employ a dual-encoder pipeline that integrates contextual semantic embeddings with alignment features derived from prototypical ABCD self-state descriptions, followed by multi-output classification and regression models to predict subelements and presence scores. For Task 2, we model timelines as ordered sequences using BiLSTM architectures to capture temporal dependencies and identify both abrupt and gradual changes in wellbeing. For Task 3, we adopt an iterative prompt-engineering approach with an LLM to generate summaries that capture self-state dynamics and change processes.

Our analysis reveals systematic failure modes in how encoding models represent psychological con-

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 482–497

structs, including a tendency to overpredict adaptive states and to conflate stylistic cues such as verbosity with markers of wellbeing. We further examine the operationalization of wellbeing through the GAF scale used in the shared task, discussing its limitations with respect to construct and ecological validity, and suggest future work explore alternative operationalizations with updated wellbeing measures adapted to social media contexts.

2 Methodology

TASK 1: POST-LEVEL ABCD SUBELEMENT CLASSIFICATION AND PRESENCE REGRESSION. The objective is to identify adaptive and maladaptive self-states expressed in posts using the MIND framework, which is divided into identifying dominant ABCD elements and subelements (Sub-task 1.1) and estimating the degree of presence of each self-state within a post (Subtask 1.2). We approach Task 1 as a feature-engineering problem solved by per-target linear models, motivated by the small labeled corpus (30 timelines). Each post is represented by an 800-dimensional vector formed by concatenating two complementary encodings. The first is a 768-dimensional contextual embedding obtained by mean-pooling the final hidden states of `mental/mental-roberta-base`, a RoBERTa variant adapted to the mental-health domain (Ji et al., 2022). For the rest of the 32-dimensions, we used the archetype framework (Varadarajan et al., 2024; Mahwish et al., 2026). We derive a vector of cosine similarities between the post and 32 first-person sentences of the ABCD framework that were generated by ChatGPT 5.1 (see Appendix B). Posts and archetypes are independently encoded by `sentence-transformers/all-MiniLM-L6-v2` (Reimers and Gurevych, 2019; Reimers, 2021), an encoder optimized for semantic textual similarity. This dual-encoder design separates the two roles of the representation: MentalRoBERTa captures rich mental-health-related contextual semantics, while MiniLM provides linguistic alignment scores against each ABCD archetype.

Given these 800-dimensional features, we train two independent linear systems. For Subtask 1.1, we fit one multinomial logistic regression per element × valence target, wrapped in a `MultiOutputClassifier`. Each target is a single multi-class label where class 0 means "element absent" and classes $1 \ldots K$ index the subelement ($K \in \{2, 4, 6, 14\}$, depending on the element). For Sub-task 1.2, we fit a ridge regressor (α=1.0) for each of the two presence scores using a `MultiOutputRegressor`; predictions are bounded to $[1, 5]$ and rounded to integers.

Both systems are trained on an 80/20 split made at the timeline level, so no timeline appears in either half. We augment the training data using `ModernBERT-base` to fill ABCD-conditioned masked tokens and discard low-quality generations, growing the training set from 312 to 495 posts ($+58.7\%$).

TASK 2: IDENTIFYING MOMENTS OF CHANGE. The objective is to detect moments of change in wellbeing, which is broken into detecting abrupt transitions (*switches*, Sub-task 2.1) and gradual development from mild to more extreme states (*escalations*, Sub-task 2.2). We address the two sub-tasks with complementary models that share the same `mental/mental-roberta-base` post encoder, motivated by their different temporal characters: switches are sharp transitions, whereas escalations span over the timeline.

For Sub-task 2.1, we frame the problem as wellbeing regression followed by a difference-based decision rule. Each post is represented by a 780-dimensional feature vector that concatenates (i) ten cosine similarities between the post and ten GAF-aligned archetype sentences encoded using `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019; Reimers, 2021), (ii) the previous post's wellbeing score, (iii) a Language Style Matching (LSM) score (Ireland and Pennebaker, 2010) with the previous post, and (iv) the 768-dimensional mean-pooled MentalRoBERTa embedding of the current post. A Ridge regressor predicts the wellbeing score $\hat{y}_t$, and a post is labeled as a switch whenever $|\hat{y}_t - \hat{y}_{t-1}| \geq 2$, mirroring the two-point gap that the annotation guidelines associate with clinically salient transitions.

For Sub-task 2.2 (Escalation), posts in a timeline are arranged chronologically into a $(T, 768)$ MentalRoBERTa embedding sequence and fed to a single-layer bidirectional LSTM (Schuster and Paliwal, 1997). Although an ablation also explored concatenating the wellbeing score and per-timeline Δwellbeing (each projected to 32 dimensions inside the network), these scalar features did not yield consistent gains over the embedding-only configuration, and the trend information they carry is largely captured by the BiLSTM's temporal mod-

eling. Given that only ∼24 timelines are available for training, the network is optimized with `BCEWithLogitsLoss` whose `pos_weight` is set to the negative-to-positive ratio of the training labels to mitigate the heavy class imbalance.

TASK 3: SUMMARY OF CHANGE. The goal is to characterize and summarize psychological change dynamics surrounding moments of change in social media timelines, which is split into generating sequence-level summaries of self-state progression around change events (Sub-task 3.1) and identifying recurrent dynamic signatures of deterioration and improvement across multiple timelines (Subtask 3.2). We use Qwen3.5-9B (Team, 2026) in a two-stage, prompt-based pipeline *without* model fine-tuning, where both stages rely on carefully structured prompts to enforce consistency, grounding, and interpretability.

For Task 3.1, the prompt specifies explicit temporal segmentation, requiring the model to anchor descriptions in time and identify a single dominant change process, while also characterizing how psychological states evolve. It further enforces relational reasoning by directing the model to describe interactions among ABCD elements in terms of reinforcing or suppressive dynamics, rather than listing them independently. Additional constraints ensure that all inferences remain grounded in the provided data and that outputs emphasize causal explanatory structures over surface-level description.

For Task 3.2, the prompt is designed to extract structured change signatures by guiding the model to implicitly group and compare patterns before synthesizing them into concise representations. It enforces grounding in prior summaries, limits redundancy by constraining the number of supporting examples, and requires uniqueness and specificity in the extracted evidence. Our prompts can be found in Appendix D.

3 Results

Tables 2–6 summarize our official results for all tasks. Each task's results show the comparison of our system against the shared-task baseline. Our systems consistently outperformed the provided baselines on Task 1, showed mixed results on Task 2, and had competitive behaviour in selected cases for Task 3.

Our system ranked seventh for Task 1.1, outperforming the baseline across all reported metrics. On the primary ranking metric, Avg Sub Macro F1, our system achieved 0.351 compared with the baseline score of 0.247. The largest improvement was observed in predicting adaptive subelements, where our system achieved 0.317 compared with the baseline's 0.156. These results suggest that the MentalRoBERTa-based representation and ABCD archetype similarity features helped capture subelement-level distinctions, particularly for categories involving broader semantic information.

For Task 1.2, our system also ranked seventh and improved over the baseline on the primary metric. Improvements were observed for both maladaptive and adaptive presence ratings, with Mal RMSE decreasing from 1.439 to 0.957 and Adp RMSE decreasing from 1.409 to 1.112. However, the baseline obtained slightly higher QWK and Spearman correlation scores, indicating that although our model produced more accurate absolute predictions, its ordinal agreement and rank correlation with the gold ratings were not consistently better than the shared-task baseline.

For Task 2, our system ranked eleventh on the combined Macro F1 metric. It achieved a combined Macro F1 of 0.412, outperforming the Llama zero-shot baseline and the Llama zero-shot plus Task 1.1 baseline, but falling below the TempoFormer baseline. At the post level, our system obtained 0.375 escalation F1, 0.357 Macro F1, and 0.339 switch F1. At the timeline level, performance was stronger for escalation detection, where our system reached 0.660 escalation F1, but lower for switch detection, with a score of 0.274 switch F1. This suggests that the system was more effective at identifying broad escalation regions than detecting exact switch points. The gap between escalation and switch performance highlights the difficulty of localizing discrete moments of change in longitudinal sequences.

For Task 3.1, our system ranked ninth by average rank. It achieved strong consistency (CS) and contradiction (CT) scores, with CS = 0.770 and CT = 0.761, both slightly higher than the two baselines. However, it performed worse on content-overlap metrics, obtaining a ROUGE-L Recall of 0.208 and BERTScore Recall of 0.255. This pattern suggests that our LLM-based prompting strategy produced summaries that were coherent and avoided contradiction, but sometimes missed specific gold-reference content. In other words, the system favored internally consistent and non-contradictory summaries over maximal lexical or semantic coverage.

For Task 3.2, our system showed asymmetric

performance across improvement and deterioration directions. For improvement signatures, our system ranked fourth and outperformed the baseline on fit, specificity, and overall score, achieving an overall score of 0.594 compared with the baseline's 0.389. However, for deterioration signatures, our system ranked eighth and underperformed the baseline, with an overall score of 0.211 compared with 0.483. This indicates that the model was more effective at identifying patterns of improvement than deterioration. One possible explanation is that improvement signatures may contain more explicit recovery cues, whereas deterioration may involve subtler or more gradual negative changes that are harder to summarize reliably.

Taken together, the results show that our strongest performance was obtained on Tasks 1.1 and 1.2, where embedding and archetype-based features consistently improved over the baselines. For Task 2, the model captured escalation more effectively than switch localization, suggesting that future work should focus on temporal boundary modeling. For Tasks 3.1 and 3.2, the results suggest that prompt-based summarization can produce coherent outputs, but further work is needed to improve content recall and robustness for deterioration-oriented explanations.

4 Analysis

TASK 1. We perform error analysis on a held-out 20% split of the training data (46 posts, 6 timelines), examining: (i) element-level false positives and false negatives; (ii) subelement identity errors given correct presence detection; and (iii) stylistic correlates of error.

Asymmetry of adaptive/maladaptive model predictions. The model overpredicts adaptive elements: in 8 of 9 posts with no true adaptive labels, the model incorrectly assigns at least one adaptive label. Overall, this results in 74 false positive (FP) adaptive element instances, compared to 27 for maladaptive elements. Per-element adaptive FP rates range from 41.7% to 64.1%, versus 14.3%–38.7% for maladaptive, indicating a clear bias toward predicting adaptive self-states broadly.

Subelement confusion on affect and desire. We plot confusion matrices for each element–valence pair (Figure 4). Errors are concentrated in the adaptive elements of affect (A; 7 subelements) and desire (D; 3 subelements). Even when the model correctly detects the presence of an adaptive A element, it identifies the correct subelement only 20.0% of the time; for adaptive D, this improves to 47%. Since affect and desire are among the most abstract dimensions in the MIND framework, these patterns suggest the model captures general emotional tone but struggles to distinguish finer-grained self-state distinctions.

Errors associated with style and themes. We analyze how error rates vary with post style by manually annotating 46 posts for seven features (e.g., length, self-deprecation, suicidality), with consensus labels from two annotators. We then test associations between features and error counts using nonparametric tests with Bonferroni correction. The most striking result is that longer posts and self-deprecating language substantially increase adaptive FPs. For example, long posts yield over 4× more adaptive FPs than short ones. This suggests the model systematically misinterprets verbosity and self-critical tone as signals of adaptive coping, rather than vulnerability.

TASK 2. Since Task 2 is defined over longitudinal emotional development, we analyze model behavior at the timeline level rather than post-level error analysis. Individual posts may contain partial evidence of escalation, while the S/E process is better understood as a gradual development across multiple posts.

Task 2.1 ablation results (Figure 2) show that predicting the continuous wellbeing score with Ridge regression and then applying a threshold generally outperforms direct binary switch classification with Logistic Regression. This suggests that switches are more naturally modeled as changes in continuous wellbeing dynamics rather than as isolated binary post-level events.

For Task 2.2, Figure 5 shows the BiLSTM probability curves are relatively smooth and often increase around gold escalation spans, suggesting that the model captures broad temporal patterns rather than reacting only to isolated posts. However, the model sometimes starts increasing slightly before the annotated span or remains high after the span ends. These cases indicate that many errors are better understood as boundary-localization or probability-calibration errors rather than complete post-level misclassifications.

Finer-grained error analysis reveals a consistent pattern across both sub-tasks. For Task 2.1, the error-rate-by-P(Switch) bucket plot (Figure 6) shows errors concentrated in mid-confidence buckets, while high- and low-confidence predictions

have better performance; the error-rate-vs-post-length plot is roughly flat (Figure 7), with only very short posts showing slightly higher error rates due to limited textual signal. For Task 2.2, the BiLSTM diagnostics follow a similar pattern: errors are localized in the mid-P(E) range rather than at the extremes (Figure 8), and post length shows no strong monotonic effect on error rate (Figure 9). Together, these results indicate that Task 2 errors are driven mainly by borderline, ambiguous posts near trajectory boundaries rather than by surface-level post length.

Overall, Task 2 performance appears to be driven by the model's ability to learn gradual wellbeing dynamics from timeline-level context. Because S/E escalation is a developmental process with discrete annotated boundaries, small mismatches between predicted and gold spans are expected.

5 Discussion and Future Work

Finally, we raise a further discussion about the operationalization of "wellbeing" and avenues for future work. The Global Assessment of Functioning (GAF) is a measure of illness severity that was included in multiple previous iterations of the Diagnostic and Statistical Manual (DSM) (Piersma and Boes, 1997). However, due to concerns regarding reliability and clinical utility (Abuse and Administration, 2016), the APA replaced the GAF in the DSM-5 with the World Health Organization Disability Assessment Schedule (WHODAS-2). Previous work has found GAF to have weak inter-rater reliability (Grootenboer et al., 2012; Vatnaland et al., 2007), and requires clinicians to assign a value 0-100 to describe the patient's impairment, where lower scores indicate greater functional impairment and higher scores indicate better overall functioning. The WHODAS-2, on the other hand, is a semi-structured questionnaire measuring functioning across different domains, combined into a total disability score (Gspandl et al., 2018).

This background points to avenues for future work into operationalizations of wellbeing to address its *construct validity* and *ecological validity*. Regarding construct validity, while our model's performance in Sub-task 2.1 (detecting *switches*) reflect limitations in our model design, the current GAF-based wellbeing operationalization may be more limited in measuring the underlying latent construct than a measure like WHODAS-2, considering the aforementioned psychology literature. Weaker construct validity suggests a weakened ability to accurately capture the intended clinical construct across modeling approaches. Therefore, future research could investigate operationalizations based on the WHODAS-2 and/or a more direct measure of clinical wellbeing such as the WHO-5 wellbeing Index (Topp et al., 2015). However, applying a measure designed for clinician administration, which relies on clinical expertise and understanding of the patient, to social media posts is a significant departure from its intended use and domain; this implies weaker ecological validity. This contextual difference may introduce additional noise to the labels on top of the GAF scale's existing limitations. Therefore, updating the wellbeing measurement instrument with an existing validated clinical instrument alone may not be sufficient; rather, empirical investigations are needed into clinically-relevant measurement instruments adapted to the social media context.

6 Conclusion

We presented a multi-stage framework for modeling mental health dynamics in longitudinal social media data for the CLPsych 2026 Shared Task, combining LLM-generated archetypes for self-state prediction, sequence modeling for change detection, and prompt-based summarization. Our results showed that LLM-derived archetype representations, constructed without any clinical expertise, can still improve over purely encoder-based approaches while remaining interpretable. This offers a promising alternative to relying solely on extensively validated, time-consuming, manually constructed archetypes. Our analyses highlighted systematic mismatches in the MIND framework and the performance of encoding models: our models overpredict adaptive states, collapse distinctions between affect and desire, and misread stylistic cues such as verbosity or self-deprecation as signs of wellbeing. These patterns suggest that models rely on surface-level signals rather than capturing underlying psychological structure. This highlights the need for tighter alignment between theory-driven constructs and learned representations in longitudinal mental health modeling. Finally, we reflected on the operationalization of wellbeing in this shared task, and suggested future directions that explore alternative operationalizations with updated wellbeing measurement instruments adapted for social media posts.

Limitations

This study has several limitations. First, the data is limited to English social media posts and does not include author demographic information. Therefore, our findings may not generalize across languages, platforms, cultures, or specific demographic groups.

Second, the dataset is highly imbalanced. Some ABCD subelements have very few or even no training examples, and some classes are extremely sparse or missing. This limits the generalizability of standard classifiers or regressors, which may overfit frequent labels and perform poorly on rare classes.

Third, we did not explore a sequential or joint pipeline across subtasks. Although predictions from one task could potentially inform another, such a setup may also introduce noise. Future work should investigate whether joint modeling can improve consistency while reducing downstream error.

Fourth, our Task 3 system was mainly based on prompt engineering, and we did not conduct a broad comparison across different LLMs or summarization methods. Thus, the results reflect one prompting-based approach rather than a comprehensive model comparison.

Finally, our archetype similarity features were based on LLM-generated archetypes. While these provided richer semantic descriptions than label names alone, they may contain biases or incomplete interpretations. Future work should involve clinical experts in writing or validating archetypes, as well as in assessing the safety, practicality, and clinical relevance of the overall approach.

Ethics

Although this work may help inform clinical practice, it is critical that the implications of our findings are considered in combination with intervention by mental health professionals. Mental health conditions are highly prevalent and often go untreated, especially in marginalized communities (Cook et al., 2017). As such, machine learning models may be useful in early detection of mental health challenges through social media language that may otherwise go undetected. This technology should be utilized as a way to connect individuals to mental health care who might otherwise lack access rather than be used as a way to deliver or replace psychiatric intervention.

It is also critical to consider the stigma associated with mental health conditions when analyzing social media language related to maladaptive behaviors (Clement et al., 2015). Interventions or other mental health detection strategies informed by this work should be carefully designed with the input of clinicians to avoid further stigmatization of those at risk of mental illness.

Acknowledgments

This work is supported in part by the McMaster University's Faculty of Engineering, Department of Computing and Software via start-up funding awards issued to Dr. Charles Welch and Dr. Allison Lahnala. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding Dr. Welch's Discovery Grant. The conclusions and opinions reflect those solely of the authors and are not attributable to institutes at McMaster University, Indiana University, Carnegie Mellon University, or NSERC.

References

Substance Abuse and Mental Health Services Administration. 2016. Dsm-5 changes: Implications for child serious emotional disturbance [internet].

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. 2024. *Self-Other Dynamics (SOD): A Transtheoretical Coding Manual*. Bar-Ilan University. Transtheoretical coding manual for psychotherapy research.

Dana Atzil-Slonim. 2025. Multimodal intrapersonal and interpersonal dynamics (mind): A transtheoretical coding manual. Open Science Framework.

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.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3(1):43.

Sarah Clement, Oliver Schauman, Tanya Graham, Francesca Maggioni, Sara Evans-Lacko, Nikita Bezborodovs, Craig Morgan, Nicolas Rüsch, June SL Brown, and Graham Thornicroft. 2015. What is the impact of mental health-related stigma on help-seeking? a systematic review of quantitative and qualitative studies. *Psychological medicine*, 45(1):11–27.

Benjamin Lê Cook, Nhi-Ha Trinh, Zhihui Li, Sherry Shu-Yeu Hou, and Ana M Progovac. 2017. Trends in racial-ethnic disparities in access to mental health care, 2004–2012. *Psychiatric services*, 68(1):9–16.

Esther MV Grootenboer, Erik J Giltay, Rosalind van der Lem, Tineke van Veen, Nic JA van der Wee, and Frans G Zitman. 2012. Reliability and validity of the global assessment of functioning scale in clinical outpatients with depressive disorders. *Journal of evaluation in clinical practice*, 18(2):502–507.

Scott Gspandl, Ryan P Peirson, Ramzi W Nahhas, Tracey Goodman Skale, and Douglas S Lehrer. 2018. Comparing global assessment of functioning (gaf) and world health organization disability assessment schedule (whodas) 2.0 in schizophrenia. *Psychiatry Research*, 259:251–253.

Molly E Ireland and James W Pennebaker. 2010. Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, 99(3):549.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Huijie Lin, Jia Jia, Jiezhong Qiu, Yongfeng Zhang, Guangyao Shen, Lexing Xie, Jie Tang, Ling Feng, and Tat-Seng Chua. 2017. Detecting stress based on social interactions in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 29(9):1820–1833.

Syeda Mahwish, Ryan L Boyd, Vasudha Varadarajan, Roman Kotov, Benjamin J Luft, H Andrew Schwartz, and Sean AP Clouston. 2026. Measuring resilience using language modeling: A computational approach to observing resilience. *Journal of Traumatic Stress*.

Harry L Piersma and Janna L Boes. 1997. The gaf and psychiatric outcome: a descriptive report. *Community Mental Health Journal*, 33(1):35–41.

Nils Reimers. 2021. sentence-transformers/all-MiniLM-L6-v2. `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`. Accessed: March 2026.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Dan J. Stein, Peter Szatmari, Wolfgang Gaebel, Michael Berk, Eduard Vieta, Mario Maj, Ymkje Anna De Vries, Annelieke M. Roest, Peter De Jonge, Andreas Maercker, Chris R. Brewin, Kathleen M. Pike, Carlos M. Grilo, Naomi A. Fineberg, Peer Briken, Peggy T. Cohen-Kettenis, and Geoffrey M. Reed. 2020. Mental, behavioral and neurodevelopmental disorders in the icd-11: an international perspective on key changes and controversies. *BMC Medicine*, 18(1):21.

Qwen Team. 2026. Qwen3.5-omni technical report.

Christian Winther Topp, Søren Dinesen Østergaard, Susan Søndergaard, and Per Bech. 2015. The who-5 well-being index: a systematic review of the literature. *Psychotherapy and psychosomatics*, 84(3):167–176.

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, Lucie Flek, H. Andrew Schwartz, Charles Welch, and Ryan Boyd. 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, St. Julians, Malta. Association for Computational Linguistics.

Torbjørn Vatnaland, J Vatnaland, Svein Friis, and Stein Opjordsmoen. 2007. Are gaf scores reliable in routine clinical use? *Acta Psychiatrica Scandinavica*, 115(4):326–330.

World Health Organization. 2025. World mental health today: latest data. Technical report, World Health Organization, Geneva. License: CC BY-NC-SA 3.0 IGO.

A Code Availability

Our code is available at `https://github.com/McMasterNLP/CLPsych-2026-McMasterNLP`.

B Archetype Sentences

We used ChatGPT (accessed April 2026) to rephrase the subelements of the ABCD scheme. There are 32 elements across the adaptive and maladaptive sections. The elements are rephrased to be expressed in first person such that we can compare the embedding to the posts in our data. Table 1 shows each sentence we used. This is inspired by the archetype similarity approach used by Varadarajan et al. (2024).

C Style Annotations

1. **Length:** Short (roughly 1–30 tokens), medium (30–200 tokens), or long (over 200 tokens).
2. **Off-topic:** Posts whose primary content was unrelated to mental health, including jokes, community pleasantries, and questions on unrelated topics (e.g., video games, entertainment trivia, general lifestyle questions). On-topic posts concerned the author's mental, emotional, or behavioral state.
3. **Help-seeking:** Posts that explicitly referenced seeking assistance, such as asking about medication, requesting advice, or expressing a desire for connection.
4. **Humor or sarcasm:** Posts containing at least one passage with non-literal communicative intent, including ironic understatement (e.g., "well, this is fine"), self-mocking exaggeration (e.g., "classic me, ruining everything"), or absurdist framing of distress (e.g., "time to befriend the void").
5. **Self-deprecation:** Posts containing explicit negative self-evaluation, framed either seriously or comically (e.g., "I'm such a loser," "I always mess up," "nobody likes me anyway").
6. **Somatic language:** Posts containing first-person descriptions of bodily sensations linked to emotional state (e.g., "my chest feels heavy," "can't sleep").
7. **Themes of suicidality:** Posts containing explicit references to wanting to die, ending one's life, or self-harm. We use this broader term rather than *suicidal ideation* because our annotators do not hold clinical training; the clinically meaningful distinction between passive death wishes, active ideation with intent, and ideation with plan was not made at this stage.

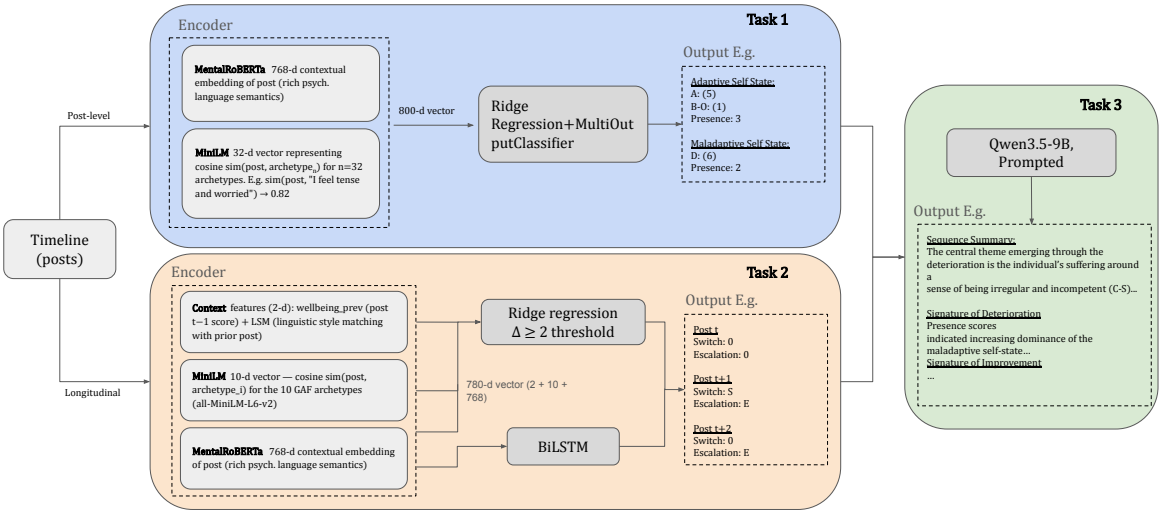


Figure 1: **System overview.** A user's temporally ordered posts (timeline) are the shared input. Task 1 operates post-level, encoding each post with MentalRoBERTa and MiniLM archetype similarity features to classify self-state components and rate their intensity. Task 2 operates longitudinally: switch detection uses regression with temporal context to flag abrupt wellbeing changes ($|\Delta| \geq 2$), while escalation uses a BiLSTM over the full timeline. Task 3 generates structured summaries of change sequences and extracts recurrent patterns of deterioration and improvement using a prompted LLM (Qwen3.5-9B).

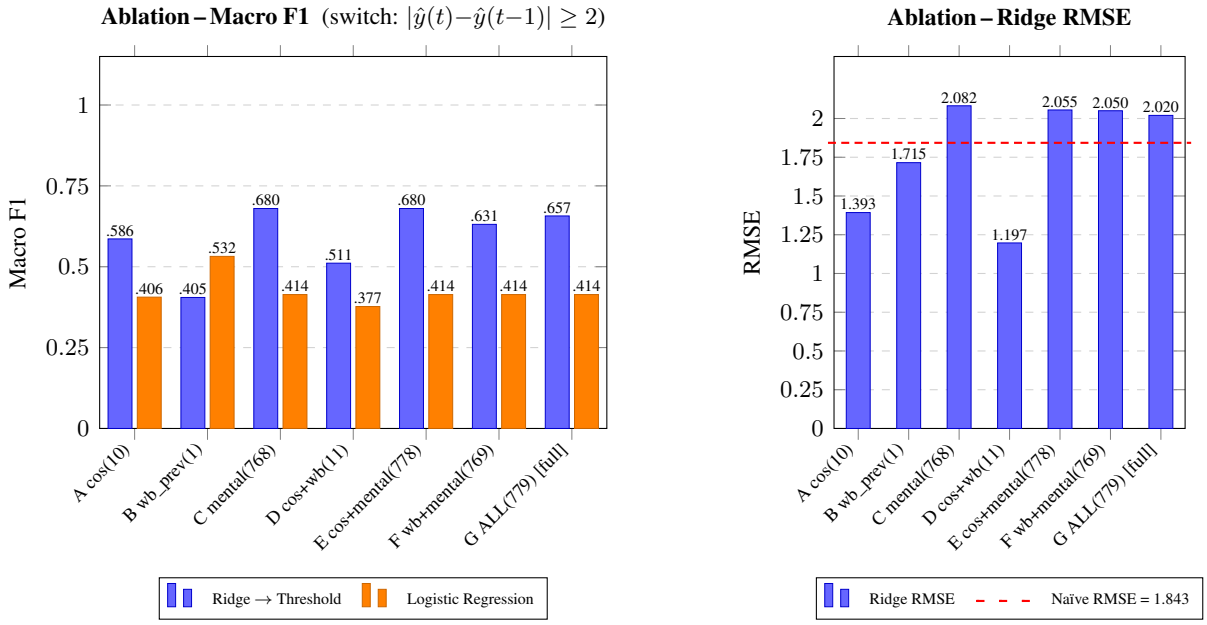


Figure 2: Feature ablation study for Task 2.1 (switch condition $|\hat{y}(t) - \hat{y}(t-1)| \geq 2$, predicted wellbeing score; cos = MiniLM archetype similarities, mental = MentalRoBERTa). **Left:** Macro F1 for switch detection, comparing Ridge → Threshold against direct Logistic Regression across seven feature sets. **Right:** Ridge RMSE for continuous wellbeing prediction; dashed line marks the naïve baseline (RMSE = 1.843).

Dimension	ID	Archetype Statement
Affect (A)		
A	1	I feel relaxed and steady, and things do not easily disturb my sense of balance.
A	2	I feel tense and worried, like something could go wrong at any moment.
A	3	I feel a deep sadness and emotional pain that is hard to shake.
A	4	I feel hopeless and weighed down, like nothing will really get better.
A	5	I feel genuinely happy and hopeful about my life and what lies ahead.
A	6	I feel intensely energized and unstoppable, like my mind is racing and everything is possible.
A	7	I feel energized and motivated, ready to take on challenges and move forward.
A	8	I feel emotionally numb and indifferent, like nothing really matters.
A	9	I feel angry about something unfair, and I want to stand up for what I believe is right.
A	10	I feel intense anger and hostility toward others, and it is hard not to lash out.
A	11	I feel proud of myself and what I have accomplished.
A	12	I feel ashamed of what I have done and keep blaming myself.
A	13	I feel valued and cared for, like I truly belong with others.
A	14	I feel alone and disconnected, like no one is really there for me.
Behavior toward Other (B-O)		
B-O	1	I try to connect with others and respond to them with openness and care.
B-O	2	When others threaten or pressure me, I either confront them or withdraw to protect myself.
B-O	3	I interact with others in a balanced way, setting boundaries while respecting theirs.
B-O	4	I feel the need to tightly control situations or others so things do not go wrong.
Behavior toward Self (B-S)		
B-S	1	I take care of myself and try to grow or improve my wellbeing.
B-S	2	I avoid taking care of myself and sometimes act in ways that harm or neglect my own needs.
Cognition of Other (C-O)		
C-O	1	I see others as supportive and connected to me.
C-O	2	I feel that others are either distant from me or cling too tightly to me.
C-O	3	I feel that others support my independence and help me grow.
C-O	4	I feel that others restrict or interfere with my ability to be independent.
Cognition of Self (C-S)		
C-S	1	I accept myself as I am and treat myself with understanding.
C-S	2	I constantly judge myself and feel like I am not good enough.
Desire (D)		
D	1	I want meaningful connection and closeness with other people.
D	2	I feel like my need for connection will probably be rejected or ignored.
D	3	I want the freedom to make my own decisions and control my life.
D	4	I feel like I will not be allowed the independence I need.
D	5	I want to feel capable, confident, and proud of what I can do.
D	6	I feel like I will never be good enough or capable enough.

Table 1: LLM-generated ABCD archetype statements used to construct category-level similarity features. Each statement represents a first-person semantic prototype for one ABCD category.

System	Avg Sub Macro F1↑	Adp Pres Macro F1↑	Mal Pres Macro F1↑	Avg Pres Macro F1↑	Adp Sub Macro F1↑	Mal Sub Macro F1↑
Ours (rank 7)	0.351	0.541	0.666	0.603	0.317	0.385
Baseline	0.247	0.392	0.605	0.498	0.156	0.338

Table 2: **Task 1.1: Subelement Classification results.** *Avg Sub Macro F1* is the primary ranking metric. Adp = Adaptive, Mal = Maladaptive, Pres = Element Presence, Sub = Subelement.

System	Avg RMSE↓	Mal RMSE↓	Adp RMSE↓	QWK↑	Comb RMSE↓	Spear↑	MAE↓
Ours (rank 7)	1.035	0.957	1.112	0.526	1.037	0.570	0.799
Baseline	1.424	1.439	1.409	0.555	1.424	0.603	1.083

Table 3: **Task 1.2: Presence Rating results.** *Avg RMSE* (Maladaptive + Adaptive) is the primary ranking metric (lower is better). Mal = Maladaptive, Adp = Adaptive, QWK = Quadratic Weighted Kappa, Spear = Spearman Correlation, MAE = Mean Absolute Error.

System	Comb	Post-level			Timeline-level		
	MF1↑	Esc	MF1	Sw	Esc	MF1	Sw
Ours (rank 11)	0.412	0.375	0.357	0.339	0.660	0.467	0.274
BL3 TempoFormer	0.572	0.667	0.561	0.456	0.736	0.583	0.430
BL2 Llama ZS + T1.1	0.365	0.407	0.383	0.359	0.343	0.346	0.350
BL1 Llama ZS	0.272	0.000	0.169	0.337	0.400	0.376	0.352

Table 4: **Task 2: Identifying Moments of Change**. *Combined (Post/Timeline) Macro F1* is the primary ranking metric. Comb = Combined, Esc = Escalation, Sw = Switch, MF1 = Macro F1, ZS = Zero-shot, T1.1 = Task 1.1 input. BL = Baseline.

System	CS↑	CT↑	R-L↑	BERTscore↑	Avg Score	Rank Avg
Ours (rank 9)	0.770	0.761	0.208	0.255	0.368	8.0
Baseline A	0.763	0.753	0.255	0.226	0.373	9.3
Baseline B	0.767	0.745	0.269	0.235	0.382	7.5

Table 5: **Task 3.1: Change Sequence Summary.** *Score Rank Average* is the primary ranking metric (lower is better). CS = Consistency, CT = Contradiction, R-L = ROUGE-L Recall, BERTscore = BERTscore Recall.

Direction	System	Fit↑	Recurrence↑	Specificity↑	Overall↑
Improvement	Ours (rank 4)	0.688	0.375	0.750	0.594
	Baseline (rank 7)	0.375	0.438	0.375	0.389
Deterioration	Ours (rank 8)	0.188	0.188	0.313	0.211
	Baseline (rank 7)	0.563	0.375	0.438	0.483

Table 6: **Task 3.2: Signatures for Improvement and Deterioration.** Human evaluation. Scores are proportion correct.

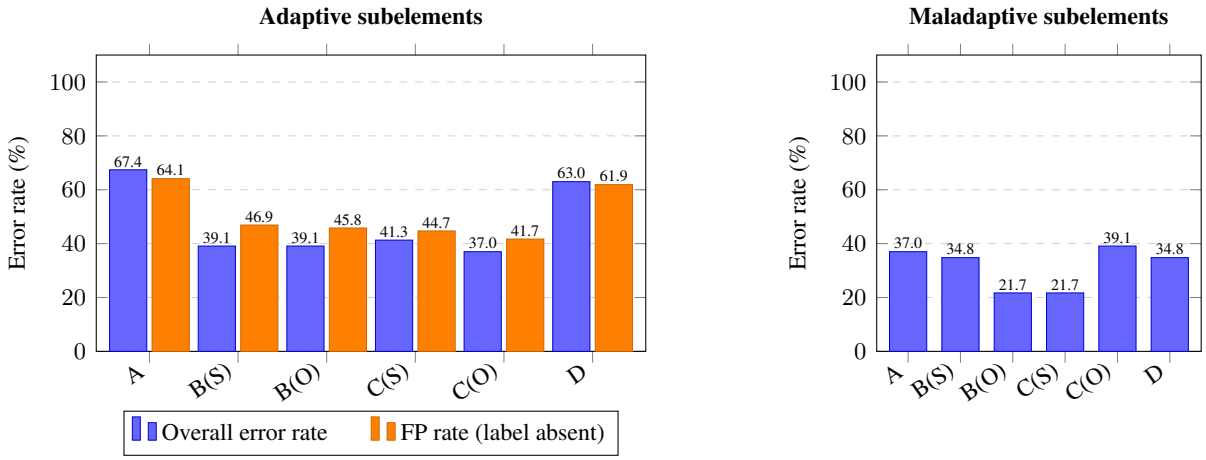


Figure 3: Per-target error rates across 46 held-out posts. **Left**: Adaptive subelements, showing overall error rate (blue) and false positive rate when the label is absent in the gold annotation (orange). **Right**: Maladaptive subelements (overall error rate only). Adap-A and Adap-D exhibit the highest error and FP rates among all targets.

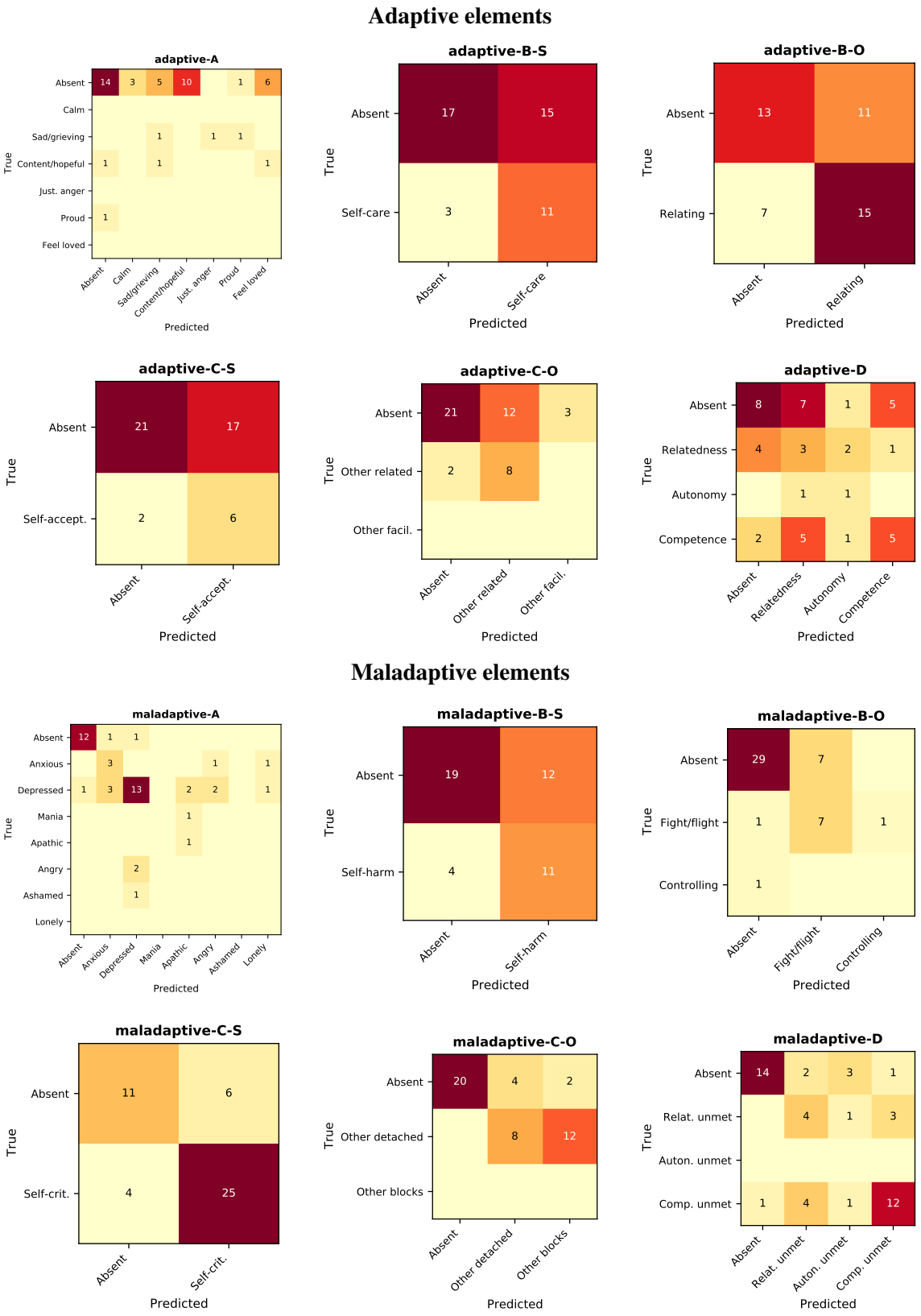


Figure 4: Confusion matrices for all 12 element-valence targets across the 46 held-out posts. Rows = true subelement, columns = predicted subelement. Value 0 = element absent. Adaptive-A and adaptive-D show the most subelement confusion; the remaining adaptive elements have only one possible non-zero value.

D Task 3 Prompt Template

For Task 3, we used the following prompt template to generate structured change-sequence summaries. The template was filled with the corresponding timeline identifier, sequence identifier, change type, ordered post indices, ordered post IDs, and textual evidence for each sequence.

Task 3 Change-Sequence Summary Prompt

You are an expert annotator for psychological change-sequence summaries.
Your task is to generate a structured narrative summary for a sequence of chronologically ordered posts surrounding a change event in wellbeing.
You must follow these rules exactly:

TASK
Write a sequence summary describing:

1. The central recurring psychological theme across the sequence.
2. The dynamics within the adaptive and maladaptive self-states.
3. The relationship between the adaptive and maladaptive self-states over time.
4. When the change event occurs.
5. Whether the change is a Switch or an Escalation.
6. Whether the direction is improvement or deterioration.

DEFINITIONS

- A Switch occurs within a single post.
- An Escalation unfolds across multiple posts.
- The sequence includes the pre-change phase and the change itself.
- The exact post where the change occurs is not explicitly marked, so infer it from the sequence.
- Presence scores indicate how strongly adaptive and maladaptive self-states are expressed, but do **not** print raw numeric scores.

ABCD FRAMEWORK
Use these abbreviations whenever referring to elements:

- (A) = Affect
- (B-S) = Behavior toward self
- (B-O) = Behavior toward others
- (C-S) = Cognition toward self
- (C-O) = Cognition toward others
- (D) = Desire

STRICT REQUIREMENTS

- You **must** explicitly describe the pre-change phase.
- You **must** explicitly state when the change occurs in the sequence.
- You **must** explicitly state the change type, Switch or Escalation.
- You **must** explicitly state the direction, improvement or deterioration.
- You **must** describe how dynamics culminate in the change, for Switch, or unfold through it, for Escalation.

SELF-STATE DYNAMICS

- When describing a self-state, you **must** include relational dynamics between at least two ABCD elements, such as mutual reinforcement, amplification, or suppression.
- You **must** describe how the presence of adaptive and maladaptive self-states changes over time.
- You **must** describe relationships between self-states using concepts such as dominance, suppression, coexistence, or reflective dialogue.

GROUNDING

- Only include ABCD elements and dynamics supported by the provided data.

- Do **not** hallucinate unsupported elements.

STYLE

- Write as one cohesive paragraph.
- Maximum 350 words.
- Be specific, analytic, and sequence-focused.
- No bullet points.

OUTPUT FORMAT
Return only:

```
Sequence Summary: <your summary paragraph>
```

CRITICAL OUTPUT REQUIREMENTS

- You **must** explicitly label the pre-change phase using the phrase: "In the pre-change phase,"
- You **must** explicitly specify when the change occurs using temporal grounding, such as early, middle, late in the sequence, or between posts.
- You **must** explicitly state both the change type and direction together, such as "a Switch towards deterioration occurs".
- Your summary **must** clearly follow this structure:
  1. Pre-change phase
  2. Change event or escalation process
  3. Post-change phase
- You **must** describe how the relationship between adaptive and maladaptive states evolves over time, such as coexistence to dominance to suppression.
- When describing a self-state, include at least one multi-element interaction involving two or more ABCD elements, such as C-S reinforcing A, which drives B-S.
- Your output **must** begin exactly with: `Sequence Summary:`

INPUT
You will receive:

- `timeline_id`
- `sequence_id`
- `change_type`
- ordered posts with structured annotations

IMPORTANT
Use only the provided structured fields and textual evidence. Do **not** assume access to any gold summary.

Now go ahead and write the summary for the following input.

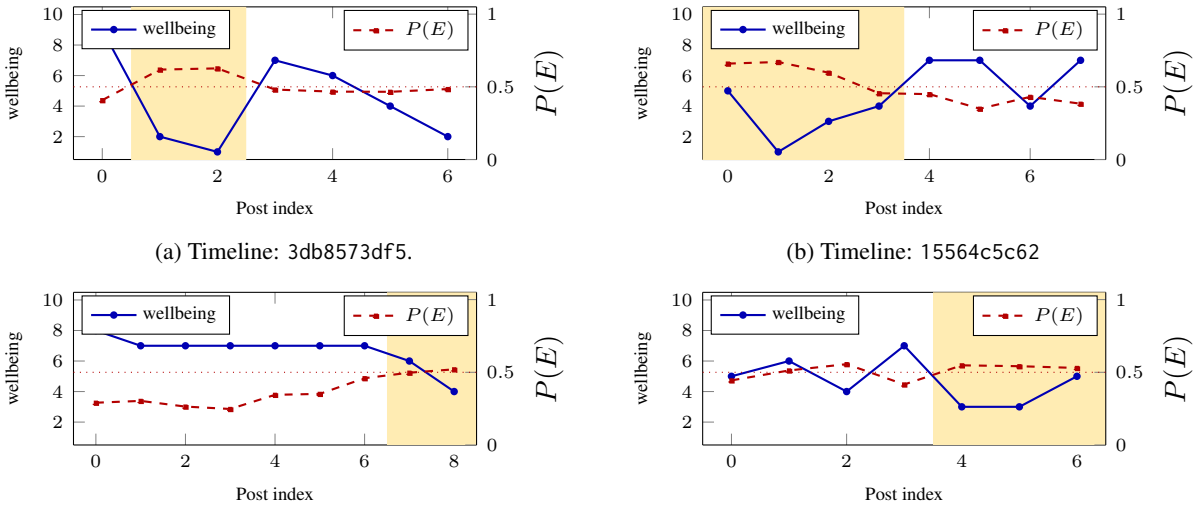
–

Structured fields for this sequence

- `timeline_id: {timeline_id}`
- `sequence_id: {sequence_id}`
- `change_type: {change_type}`
- `ordered post indices (chronological): {postindices}`
- `ordered post ids: {postids}`

Textual evidence for the sequence

```
{summary}
```



(a) Timeline: `3db8573df5`.

(b) Timeline: `15564c5c62`

(c) Timeline: `d0fb4b962e`

(d) Timeline: `62defd8df2`

Figure 5: Escalation trajectories on four held-out test timelines. Blue solid line: wellbeing score (left axis). Red dashed line: Bi-LSTM $P(E)$ (right axis, dotted line at 0.5). Gold shading: ground-truth escalation spans.

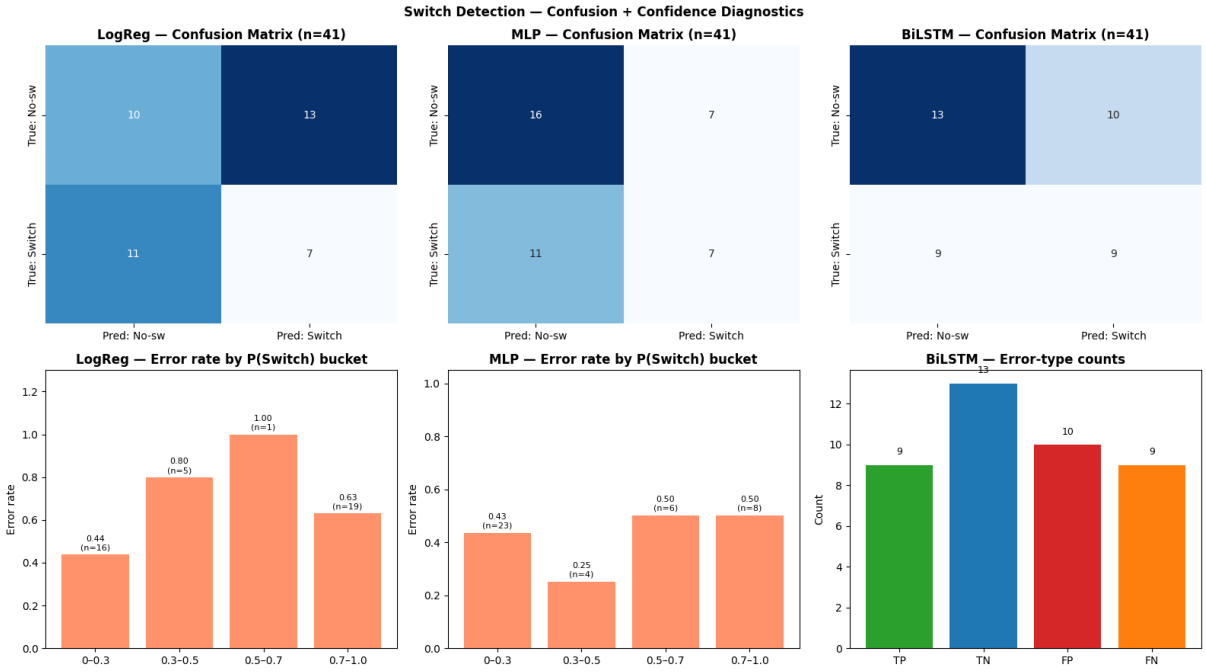


Figure 6: Confusion matrices and confidence diagnostics for switch detection.

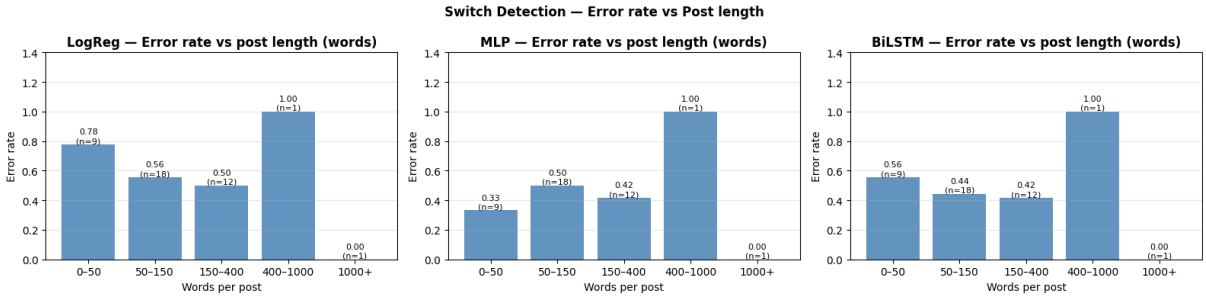


Figure 7: Error rate by post-length bucket for switch detection.

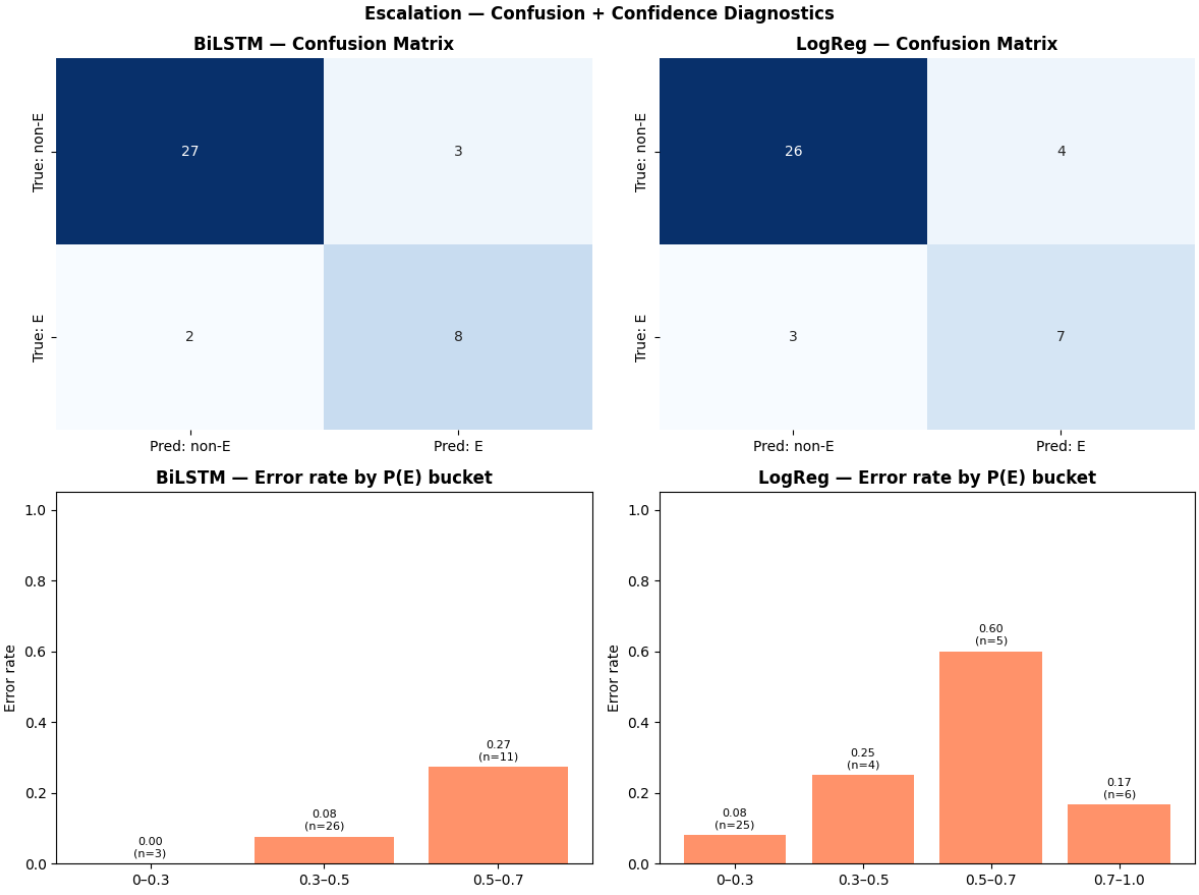


Figure 8: Confusion matrices and confidence diagnostics for escalation detection.

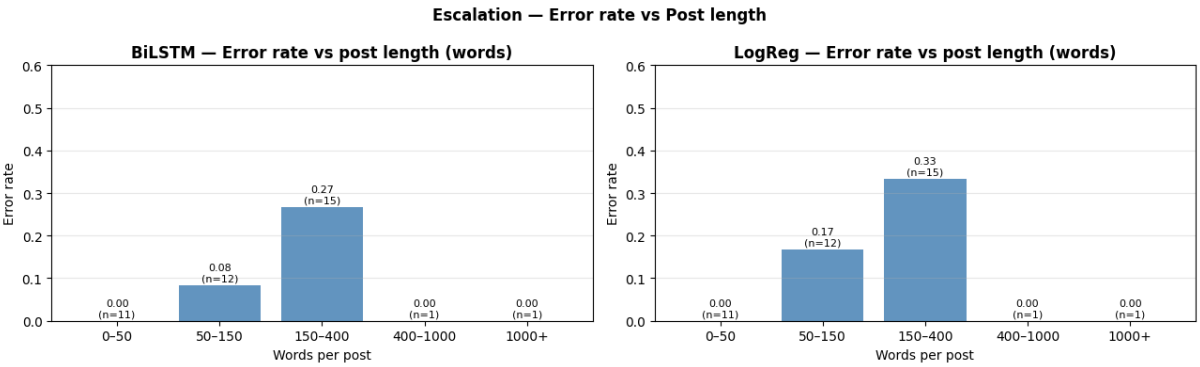


Figure 9: Error rate by post-length bucket for escalation detection.