Overview of the CLPsych 2025 Shared Task: Capturing Mental Health Dynamics from Social Media Timelines

Talia Tseriotou[1*], **Jenny Chim**[1*], **Ayal Klein**[3], **Aya Shamir**[3], **Guy Dvir**[3], **Iqra Ali**[1], **Cian Kennedy**[1], **Guneet Singh Kohli**[1], **Anthony Hills**[1], **Ayah Zirikly**[4], **Dana Atzil-Slonim**[3], **Maria Liakata**[1,2]
[1]Queen Mary University of London (UK), [2]The Alan Turing Institute (UK),
[3]Bar Ilan University (Israel), [4]Johns Hopkins University (US)
{t.tseriotou; c.chim; m.liakata}@qmul.ac.uk

Abstract

We provide an overview of the CLPsych 2025 Shared Task, which focuses on capturing mental health dynamics from social media timelines. Building on CLPsych 2022's longitudinal modeling approach, this work combines monitoring mental states with evidence and summary generation through four subtasks: (A.1) Evidence Extraction, highlighting text spans reflecting adaptive or maladaptive self-states; (A.2) Well-Being Score Prediction, assigning posts a 1 to 10 score based on social, occupational, and psychological functioning; (B) Post-level Summarization of the interplay between adaptive and maladaptive states within individual posts; and (C) Timeline-level Summarization capturing temporal dynamics of self-states over posts in a timeline. We describe key findings and future directions.

1 Introduction

Mental health concerns is a pressing global issue (WHO, 2022), necessitating solutions that both expand access to care and continuously monitor individuals over time, thereby reflecting the multifaceted and dynamic nature of mental health.

Over the past decade, social media platforms have emerged as major venues where people openly discuss mental health, sharing experiences and emotional states that can span years (Coppersmith et al., 2014; Shing et al., 2018; Zirikly et al., 2019; Tsakalidis et al., 2022b). This abundance of user-generated data offers an unprecedented opportunity to monitor individuals longitudinally, providing early detection, prevention, and "just-in-time" interventions well before difficulties escalate.

While traditional NLP approaches to mental health centered on static classification tasks (e.g. depression detection in De Choudhury et al. (2013)), recent work has recognized the complexity of mental health trajectories as fluctuating dynamic states influenced by evolving contexts, interactions and psychological processes, emphasizing the need for longitudinal, context-rich models that capture how mood, behavior, and cognition fluctuate over time (Tsakalidis et al., 2022b; Tseriotou et al., 2023). Moreover, accounting for both maladaptive and adaptive states delivers a more nuanced picture of well-being while also uncovering factors that can lead to personalized interventions (Slonim, 2024).

CLPsych shared tasks have followed this trend, shifting from user-level classification (Coppersmith et al., 2014; Shing et al., 2018; Zirikly et al., 2019) to longitudinal tasks such as detecting "Moments of Change" (MoC) (Tsakalidis et al., 2022a) and evidence generation (Chim et al., 2024).

The CLPsych 2025 shared task combines longitudinal modeling in social media timelines with evidence generation, promoting humanly understandable rationales that support recognizing mental states as they dynamically change over time. Adopting the MIND transtheoretical framework (Slonim, 2024), we seek to identify both adaptive and maladaptive self-states in a users longitudinal data via the following tasks: (A.1) Evidence Extraction, highlighting text spans within posts that reflect adaptive or maladaptive states; (A.2) Well-Being Score Prediction, assigning a 1–10 rating indicative of individuals' social, occupational, and psychological functioning, informed by maladaptive and adaptive states; and (B-C) Summarization, capturing individuals' mental health progression at the post level (B) and across the entire timeline (C) on the basis of adaptive and maladaptive states.

Our dataset comprises Reddit-based user timelines from mental health related subreddits (MHS), with posts annotated by clinical experts following the MIND scheme, which captures how an individuals self-state evolves in response to personal challenges, life events, or social interactions. From a clinical perspective, this means not only detecting

[*]Denotes equal contribution.

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 193–217

risk and symptoms but also identifying and tracking a person's strengths and coping abilities as they emerge and evolve. Summaries at various time resolutions further enhance explainability – critical for mental health professionals seeking clear, evidence-based insights. Computational challenges involve working with models that can process longitudinal data, incorporating and synthesizing appropriate evidence to generate rationales for the progression of an individual.

After providing a quick review of the landscape in NLP for mental health, focusing on temporality and explainability (§2), we describe the shared task (§3) and data annotation (§4.2). We discuss evaluation metrics (§5), methods by participating teams and results (§6.1), and conclude with an overview of key findings (§6.4), limitations, clinical implications and directions for future research (§7).

2 Related Work

2.1 NLP for Mental Health Applications

Explainability in mental health: Early work primarily focused on classification tasks, either at the document-level (Sawhney et al., 2022a) or user-level, with the latter addressing both static assessment of mental health conditions (Coppersmith et al., 2015; Shing et al., 2018; Zirikly et al., 2019; Sawhney et al., 2022b) and longitudinal monitoring of psychological states over time (Tsakalidis et al., 2022a,b; Hills et al., 2023).

Recent developments have shifted toward more fine-grained analysis and explainable approaches to mental health assessment. The 2023 eRisk Task focused on ranking sentences based on their relevance to depressive symptoms (Parapar et al., 2023), while Nguyen et al. (2022) developed BERT-based methods that incorporate PHQ-9 symptoms for improved interpretability in depression detection. Similarly, Nemesure et al. (2021) employed SHAP values (Lundberg and Lee, 2017) to explain predictions for anxiety and depression models, and Zirikly and Dredze (2022) leveraged PHQ-9 questions as auxiliary tasks to provide explanations for depression detection, evaluating performance on manually annotated text spans.

In the context of fostering interpretability, Garg (2024) annotated a dataset with highlighted text spans across various 'wellness' dimensions, while the CLPsych 2024 shared task explored Large Language Models (LLMs) to identify evidence supporting suicide risk assessments (Chim et al., 2024). This reflects the field's increasing emphasis on providing clinically meaningful explanations alongside predictions.

LLMs have been leveraged for mental health classification (Amin et al., 2023), data augmentation (Liyanage et al., 2023), and reasoning (Xu et al., 2023), demonstrating promise in detecting psychological indicators (Yang et al., 2023), extracting relevant evidence from text (Xu et al., 2024a), and generating clinically informed summaries (Song et al., 2024b). LLMs using instruction fine-tuning and Chain-of-Thought (CoT) prompting (Yang et al., 2023) have also been employed, though such approaches can pose risks of incorrect predictions and flawed reasoning, especially in complex conversations (Li et al., 2023).

Evidence extraction: Accurate span extraction is a crucial task in mental health assessment, enabling clinicians to identify and summarize the most relevant patient data for clinical evaluation. Prior work at the intersection of NLP and mental health have utilized LLMs to predict critical mental states and provide reasoning for predictions (Yang et al., 2024b; Xu et al., 2023, 2024b). Yet these approaches lack transparency and complex reasoning processes can lead to hallucination (Li et al., 2023).

LLMs and summarization: LLMs have been used to generate clinically meaningful summaries from social media posts (Song et al., 2024b, 2025; Sotudeh et al., 2022), summarize counseling sessions (Srivastava et al., 2022), generate structured medical reports from patient-doctor conversations(Adhikary et al., 2024; Michalopoulos et al., 2022), and summarize Mental State Examinations (MSE) (Mumtaz et al., 2024). However, uncertainties remain regarding the effectiveness of LLMs in generating contextually appropriate summaries, particularly in domains such as mental health (Klein et al., 2024; Asgari et al., 2024).

Temporal modeling: Most models have relied on recurrent neural networks without explicitly accounting for time intervals between posts (Tsakalidis et al., 2022a), or struggle to capture complex linguistic patterns over time (Bayram and Benhiba, 2022), despite the role of longitudinal linguistic features in mental health applications (Homan et al., 2022; Chim et al., 2025). Recent work has developed time-aware modeling approaches. Hills et al. (2023) introduced a Hawkes process-inspired approach capturing both temporal dynamics and linguistic context in user timelines, which was further

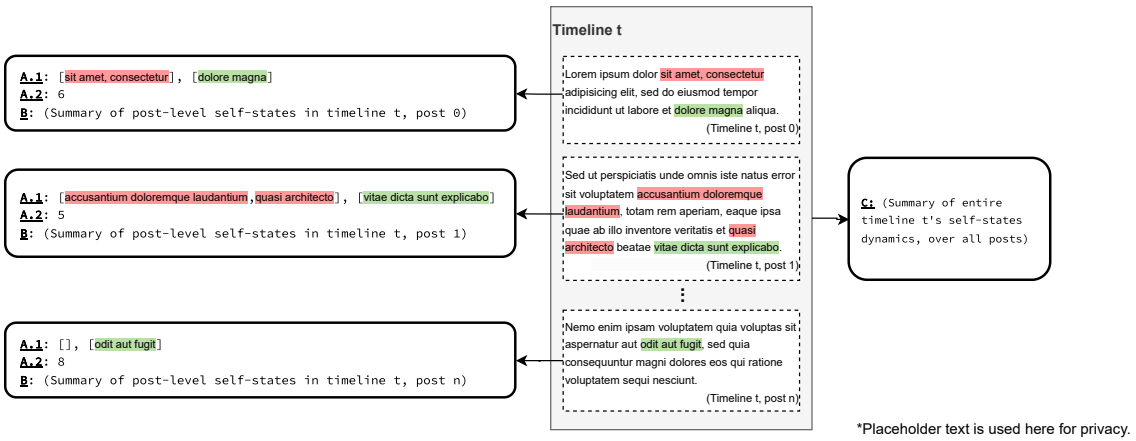

Figure 1: Participants are asked to tackle the tasks described in §3: identifying adaptive and maladaptive evidence (A.1), predicting well-being (A.2), and summarizing mental health dynamics at the post (B) and timeline level (C).

integrated into a hierarchical transformer architecture (Hills et al., 2024). Tseriotou et al. (2023) proposed sequential path signature networks to capture the temporal and linguistic progression in user posts, while Tseriotou et al. (2024b) introduced TempoFormer, which modifies the transformer architecture to account for relative temporal aspects between sequential data points, using time-sensitive rotary positional encodings. Such approaches have demonstrated superior performance in detecting subtle changes in mood and mental states by effectively modeling both linguistic and temporal context in social media posts.

2.2 Mental Health Self-State Dynamics

The MIND approach (Slonim, 2024) proposes a paradigm shift from categorical classification of trait-like psychopathology to modifiable intrapersonal dynamics. MIND provides a transtheoretical scheme that breaks down individuals experiences into core building blocks central to major therapeutic approaches, including cognitive-behavioral therapy (CBT), psychodynamic, interpersonal, relational, and experiential models. This highlights that human experience consists of multiple self-states that fluctuate and change over time (Beck et al., 2021; Bromberg, 2014; Stiles, 2001). Each self-state comprises identifiable elements characterized by specific combinations of Affect, Behaviour, Cognition, and Desire (ABCD) (Revelle, 2007) coactivated in a meaningful manner and directed either toward the self or others (Lazarus and Rafaeli, 2023). At a specific moment, one self-state may be dominant, while others, often complementary, may be subtler or remain in the background. Focusing on self-states in theory, research, and practice can reveal within-person adaptive and maladaptive states, as well as between-person differences in typical self-states.

Recent developments in NLP, specifically the emergence of LLMs, have demonstrated the capability to identify individuals' emotions (Mayer et al., 2024) and cognitions (Singh et al., 2024) from longitudinal mental health data. Nonetheless, no prior research has yet leveraged LLMs to systematically capture the complex interplay among ABCD elements especially as these manifest dynamically in adaptive and maladaptive self-states, a gap which this shared task directly seeks to address.

3 Task Definition and Instructions

We describe in detail the tasks introduced in §1 and provide an overview in Figure 1.

Task A consists of two sub-tasks: **Task A.1** involves identifying adaptive and maladaptive self-state evidence from an individual's post as a set of continuous spans. Each post can include either: (1) a single self-state (adaptive or maladaptive); (2) two complementary self-states (adaptive and maladaptive) or (3) evidence of neither. **Task A.2** requires rating the overall well-being of an individual on a scale from 1–10 based on GAF (Association et al., 2000). This score reflects the well-being of an individual based on three aspects: social, occupational and overall individual psychological functioning. While well-being scores are assigned at the post level, participants were encouraged to consider the sequence of previous posts and the extracted evidence from Task A.1 in this task.

Task B involves post-level summarization of self-states. Such summaries should capture the interplay between adaptive and maladaptive states manifesting in the post through identification of the central organizing aspects (ABCD) that drive the state and should provide the anchors for the summary. The expectation is to first identify the dominant self-state and then describe how core aspects influence the rest, emphasizing their evolution.

Task C involves summarizing self-states at the individuals' timeline-level. The focus should be on the temporal interplay between adaptive and maladaptive self-states, with emphasis on concepts such as flexibility, rigidity, improvement, and deterioration. When applicable, the temporal dynamics should capture changes in the dominant self-states and specifically how underlying changes in ABCD aspects contribute to potential transitions.

Ground truth data for all the above tasks were provided to the participants during training but not at test time. For Task A.1, additional information regarding each gold self-state evidence was made available in the training set only. Specifically, the types of evidence provided were under: Affect (A), Behavior towards the other (B-O), Behavior towards the self (B-S), Cognition of the other (C-O), Cognition of the self (C-S), and Desire/Expectation (D), along with further sub-categories stemming from each of the six categories. The full list of categories can be found in the Appendix A.2.

4 Data and Annotation

4.1 Data

We utilized the Reddit-New dataset originally introduced in the CLPsych 2022 shared task (Tsakalidis et al., 2022a). This dataset comprises user timelines extracted from various MHS from 2015 to 2021. Given the extensive nature of the MIND annotation scheme, annotating entire timelines proved to be prohibitively resource-intensive. To address this, we implemented a selective sampling strategy. Specifically, we reduced excessively long timelines by extracting subsets containing between 10 and 12 representative posts. Additionally, timelines of moderate length were preferentially sampled to balance feasibility with sufficient contextual richness. Beyond this length-based selection, timelines were chosen randomly, subject to two constraints:

- **User Uniqueness:** No user was represented by more than one timeline within the test set, and users appearing in the training set were explicitly excluded from the test set to ensure independence between training and evaluation data and to prevent potential data leakage that could inflate performance metrics.
- **Density Diversity:** Using the CLPsych 2022 annotations for mood switches and escalations (i.e. MoC), we define timeline 'density' as the proportion of posts labeled with MoC tags and use it for stratified sampling. This helps us to capture a diverse range of emotional fluctuation patterns and related mental health dynamics.

	Train	Test
# Timelines	30	10
# Posts	343	94
Avg. Tokens per Post	134.4	142.9
# Adaptive Evidences	399	145
# Maladaptive Evidences	526	171

Table 1: Dataset descriptive statistics.

The final dataset (see Table 1) contains timelines selected for length, content relevance, user uniqueness, and density distribution. This strategy maintains the longitudinal nature of the data while providing sufficient context for identifying adaptive and maladaptive self-states, as well as capturing the dynamics of psychological states over time. Furthermore, this enabled thorough annotations of detailed ABCD aspects in each post.

4.2 Annotation

Two Master's students in clinical psychology, both fluent in English, annotated the selected timelines using the MIND scheme (§2.2). Annotators received comprehensive training conducted by a clinical expert and ongoing supervision from a senior MA student with prior experience in annotation using the MIND scheme. Annotators underwent a preliminary training phase, during which they received iterative feedback and conducted reconciliation meetings to enhance consistency and inter-rater reliability.

Annotators followed a structured workflow. For each post, they first identified adaptive and maladaptive self-states. Within each identified self-state they annotated the present ABCD elements, selecting the most salient span as evidence for each element. Next, they assessed the individual's overall well-being based on GAF guidelines, considering both the specific post and the context of previous posts. They then composed a detailed sum-

mary for each post, specifying which self-state was dominant, the primary psychological dimension underpinning that self-state from the ABCD elements, and a description of the interplay between different elements constituting the self-state. This description considered temporality and causality to capture the evolving psychological dynamics within each post.

Beyond individual posts, annotators synthesized their insights at the timeline level, producing a comprehensive summary that mapped the interplay between adaptive and maladaptive self-states across the timeline. This included a description of how self-states dynamically changed (or remained stable) over time. Details about the annotation platform are specified in Appendix A.1.

Inter-annotator agreement was assessed using standard reliability metrics over 23 posts annotated by both annotators. For Task A.2 (Well-being Score), which involves numerical ratings, annotators demonstrated high agreement achieving a Pearson correlation coefficient (r) of 0.793 and an Intraclass Correlation Coefficient (ICC2) of 0.791, indicating high agreement. Additional evaluations of inter-annotator agreement using task-specific metrics detailed in Section 5 (BERTScore-based measures for Task A.1 and mean squared error for Task A.2), are provided in Appendix A.3.

5 Evaluation Metrics

5.1 Task A

Evidence Extraction. The main metric we consider is the recall of evidence spans. Recall is prioritized given that the costs of overlooking important evidence outweigh those of supplying excess evidence for the task of capturing mental health dynamics over time. Moreover, in our gold data, annotators selected the single most salient evidence span per self-state annotation (§4.2). As such, precision metrics could unfairly penalize valid predictions that simply differ from what the annotator considers as the most salient, whereas recall more accurately reflects performance.

Following Chim et al. (2024), for a given user, given predicted evidence spans H and gold evidence spans E, we average the maximum recall-oriented BERTScore (Zhang et al., 2020):

$$\text{Recall} = \frac{1}{|E|}\sum_{e\in E}\max_{h\in H} BERTScore(e,h)$$

We use `deberta-xlarge-mnli` to compute embeddings and apply rescaling as recommended by Zhang et al. (2020). In addition, we report a weighted version of recall, which is sensitive to predicted evidence lengths relative to gold evidence lengths. For a given user with gold evidence spans of cumulative token count n_{gold} and predicted spans with cumulative token count n_{pred}, if the predicted evidence spans are longer than the gold-standard ones, we apply weight w to the timeline-level recall:

$$w = \begin{cases} \frac{n_{\text{gold}}}{n_{\text{pred}}} & \text{if } n_{\text{pred}} > n_{\text{gold}} \\ 1 & \text{otherwise} \end{cases}$$

Well-being Score Prediction We evaluate well-being score predictions over all annotated posts using mean squared error (MSE), which appropriately penalizes larger errors and accommodates ordinal and continuous data:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

As supplementary metrics, we report MSE stratified by risk categories: serious impairment to functioning (1 to 4), impaired functioning (5 to 6), minimal impairment to functioning (7 to 10). Finally, based on these categories, we cast the task into a classification problem (serious, impaired, minimal) and report macro F1. This reflects a system's ability to identify well-being categories rather than exact scores, regardless of category prevalence.

5.2 Task B

Following prior work in general domain (Maynez et al., 2020) and mental health summarization (Song et al., 2024a), we leverage predictions from a natural language inference (NLI) model (Laurer et al., 2024) for summary evaluation.[1] We consider consistency to be the absence of contradiction. For each sentence in a submitted summary $s \in S$, we use the NLI model to compute its mean probability of contradicting each sentence in the corresponding gold-standard evidence summary $g \in G$, taking the gold sentence as premise and the submitted sentence as hypothesis:

$$\text{CS} = \frac{1}{|S|\cdot|G|}\sum_{s\in S}\sum_{g\in G}(1 - \text{NLI}(\text{Contradict}|g,s))$$

To complement consistency, we also evaluate summaries by their contradiction to expert summaries. We expect there to be some natural contradictory information in most summaries, since

[1] https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

summarized evidence can include both risk factors and protective factors. We compute the contradiction score by averaging the *maximum* contradiction probability of a predicted sentence against gold evidence summary sentences:

$$\text{CT} = \frac{1}{|S|} \sum_{s \in S} \max_{g \in G} \text{NLI}(\text{Contradict}|g, s)$$

To support post-hoc analysis, we assess whether generated summaries are supported by their corresponding supporting evidence spans. This is only informative if the assessed system actually uses predicted spans for post-level summarization.

$$\text{EA} = \frac{1}{|H|} \sum_{h \in H} \max_{s \in S} \text{NLI}(\text{Entail}|h, s).$$

5.3 Task C

Following Task B, we evaluate timeline-level summaries primarily with mean consistency (CS), supplementing with contradiction (CT).

6 Teams & Results

6.1 Participating Teams

A total of 26 teams (69 members) completed the registration process (see Appendix B.1), with members of 3 teams having participated in a past CLPsych shared task. Out of these 26 teams, 14 (49 members) submitted output files for one or more tasks and 11 teams submitted a paper (Table 9). Teams who submitted solutions averaged 3.5 members while those who did not averaged 1.6, suggesting that having more members increased the chance of completion.

6.2 Baselines

A range of LLMs and smaller model baselines were provided along with the official team submissions' results. This allowed for a direct comparison of teams' solutions, given strong setups for each task. Baselines are presented below (with prompts in Appendix C). All LLM baselines used `Llama-3.1-8b-Instruct` (Grattafiori et al., 2024).[2]

Task A.1: For evidence extraction, two zero-shot prompting baselines and two smaller BART-based models were used, representing both single-post and window-based approaches with the latter taking into account the context of recent posts. BART-based models allowed showcasing the effect of fine-tuning for generation.

[2]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

	All		Adapt.		Maladapt.	
Teams	**R**	**WR**	**R**	**WR**	**R**	**WR**
Aquarius	.51	.46	.50	.47	.52	.45
BLUE	.56	.39	.47	.40	.64	.38
BULUSI	.43	.37	.34	.34	.53	.40
CIOL	.25	.17	.23	.15	.26	.20
CSIRO-LT	.46	.43	.38	.38	.54	**.48**
EAIonFlux	.52	.47	.52	.48	.52	.46
ISM	.56	.45	.49	.46	.63	.44
MMKA	.60	.34	.52	.37	.68	.31
NoviceTrio	-.03	-.03	-.10	-.10	.05	.05
PsyMetric	.17	.17	.15	.15	.18	.18
ResBin	.47	.30	.26	.26	.68	.36
Seq2Psych	.28	.24	.25	.24	.31	.24
uOttawa	**.64**	**.50**	**.59**	**.54**	.68	.46
Zissou	.58	.32	.45	.31	**.71**	.34
Llama ZS Single-Post	.36	.34	.31	.29	.38	.41
Llama ZS 5-Post	.50	.26	.37	.25	.63	.27
BART Single-Post	.40	.38	.47	.46	.34	.30
BART 5-Post	.26	.26	.28	.28	.24	.24

Table 2: Results on Task A.1 (evidence extraction). We consider recall and weighted recall over all spans, adaptive spans only, and maladaptive spans only.

- Llama ZS Single-Post: Zero-shot prompting on each post by providing definitions of adaptive and maladaptive self-states and asking the LLM to generate an adaptive and a maladaptive evidence list.
- Llama ZS 5-Post: Same as above, but operating on each post along with the recent posting history from an individual's timeline (5 posts).
- BART Single-Post: BART (Lewis et al., 2020) fine-tuned separately for adaptive and maladaptive span generation on each post.
- BART 5-Post: BART fine-tuned as above but operating on a window of 5 posts separated by a [SEP] token (current post + 4 recent posts). Spans are based solely on the last post.

Task A.2: For well-being score prediction, two zero-shot prompting baselines and two smaller models were used. While one version of the models is single-post based, another version considered the context of recent posts, as is the case of the BiLSTM modeling the sequential aspect.

- Mode: Mode of training data scores (7).
- Llama ZS Single-Post: Zero-shot prompting on each post by providing definitions of each score for well-being prediction.
- Llama ZS 5-Post: Same as above, but operating on each post along with the recent posting history of the individual (5 posts).

- BERT Post-level: BERT model (Devlin et al., 2019) with a regression layer fine-tuned on the post-level for well-being score prediction, averaged over 5 seeds.
- BiLSTM 5-Post (BERT): BiLSTM operating on a window of 5 posts (current + 4 recent posts) for well-being score prediction, averaged over 5 seeds. For each post the [CLS] token of BERT representations is used.

Tasks B and C: Two zero-shot LLMs, with one version including an intermediate LLM generated summary, were used for the post and timeline summaries using prompts with clinical directions.

- Llama ZS Summary: Zero-shot prompting on single posts (Task B) and timelines (Task C). The model is instructed to identify the dominant and secondary self-states and highlight the central organizing ABCD aspects that drive the state along with their interplay, including guidance through definitions.
- Llama ZS w/ Intermediate Summary: A two-layer LLM approach following Song et al. (2024b), where first a post-level (Task B) or a timeline-level (Task C) concise summary is produced with zero-shot prompting, and then this summary is used as in *Llama ZS Summary* to generate the self-states summary.

6.3 Results

This section presents results and an overview of system submissions, focusing on the best run.[3]

System Characteristics The majority of submissions took a pipeline approach, using predictions from an earlier subtask to inform the next (e.g. use predicted evidences and scores to guide summarization). More than a third of teams used Retrieval Augmented Generation (RAG) through dense retrieving examples from the training set for in-context learning. Most used clinical information provided in the shared task description document in their prompts, and a few explored incorporating additional domain knowledge during feature extraction, prompt design, and data augmentation (Seq2Psych, CIOL, BLUE).

Model Characteristics A few teams employed Pretrained Language Models (PLMs), mostly for evidence extraction (MMKA, Seq2Psych). About a third used traditional approaches, such as KDE for sampling (ISM) and random forests for span classification (CIOL) and well-being score prediction (ResBin). All teams used LLMs in at least one task. LLMs have mostly been used to directly generate predictions, but also for feature extraction (Seq2Psych). Participants developed systems on private and self-hosted instances, without using Cloud APIs. All employed LLMs were generalist models, generally 9B or smaller in size (42%), and the majority can model long contexts of over 100k tokens (58%).

[3]For details of each submission and information about model families, sizes, and context lengths, see Appendix B.

Teams	MSE (↓) All	Min.	Imp.	Ser.	F1 (↑)
Aquarius	2.01	1.25	3.11	2.16	0.37
BLUE	2.26	2.06	3.69	**1.41**	**0.39**
BULUSI	**1.92**	**0.65**	1.19	3.04	0.35
CIOL	3.99	2.89	**0.49**	7.31	0.12
CSIRO-LT	2.04	1.08	3.68	1.82	0.34
EAIonFlux	2.08	2.11	3.71	1.77	0.32
ISM	2.76	2.74	5.00	1.93	0.32
MMKA	6.61	4.95	11.76	4.22	0.26
NoviceTrio	13.83	18.62	11.59	3.16	0.14
PsyMetric	3.23	3.28	6.63	2.52	0.30
ResBin	8.02	1.89	3.71	20.26	0.19
Seq2Psych	3.27	2.63	1.38	4.98	0.19
uOttawa	2.62	2.91	4.03	2.28	0.30
Zissou	3.14	3.09	4.32	2.91	0.34
Mode	7.30	0.47	1.31	19.20	0.13
Llama ZS Single-Post	4.22	3.20	3.66	4.67	0.26
Llama ZS 5-Post	4.46	7.06	3.20	1.67	0.27
BERT Post-level	2.90	2.81	2.32	3.39	0.14
BiLSTM 5-Post (BERT)	4.56	5.34	1.01	5.68	0.13

Table 3: Results on Task A.2 (well-being score prediction). In addition to overall MSE, performance on posts in different well-being score ranges are reflected by MSE computed over posts in the minimal impairment to functioning, impaired functioning, and serious impairment to functioning ranges, and macro F1.

Task A.1: Results for evidence identification are in Table 2. Instruction prompting with demonstrations proved effective, as shown in the system that achieved top recall and length-weighed recall (uOttawa). Most submissions followed this approach, although finetuned PLMs continue to be performant (MMKA). Systems that achieve high recall on adaptive tend to also perform well on maladaptive spans. Across the board, systems were better at identifying evidence for maladaptive self-states than adaptive ones, with the exception of EAIonFlux, which targets retrieval and achieves the same performance level on both self-state categories.

Task A.2: Results for well-being score prediction are in Table 3. The best-scoring system used an optimized weighted LLM ensemble (BULUSI). Sys-

	Task B			Task C	
	CS	**EA**	**CT (↓)**	**CS**	**CT (↓)**
Aquarius	.88	.69	.78	.92	.88
BLUE	**.91**	.59	**.53**	**.95**	.54
BULUSI	.87	.81	.81	.94	.71
CIOL	.61	**.81**	.97	.61	1.0
CSIRO-LT	-	-	-	-	-
EAIonFlux	.89	.76	.78	.91	.76
ISM	.86	.76	.78	.85	.83
MMKA	-	-	-	-	-
NoviceTrio	.69	.17	.89	.86	.60
PsyMetric	.70	.47	.56	.93	**.35**
ResBin	.76	.67	.84	.90	.82
Seq2Psych	-	-	-	-	-
uOttawa	.86	.70	.83	.94	.71
Zissou	.85	.74	.77	-	-
Llama ZS Summary	.88	-	.85	.88	.80
Llama ZS w/ Inter. Summary	.89	-	.84	.94	.58

Table 4: Results on Task B (post-level) and Task C (timeline-level). Summaries are assessed primarily on mean consistency to gold summaries (CS). We additionally report entailment by extracted evidences (EA) on post-level, and contradiction to gold summaries (CT).

tems that incorporate extracted evidence or jointly predict evidence and score tend to achieve better MSE (Aquarius, EAIonFlux), but tackling more than two subtasks in the same prompt remains challenging. Given this task's sequential nature, teams also explored using timeline-based features (CIOL) and person-contextualized modeling (Seq2Psych). Overall, systems that excel at posts in the impaired functioning range (5 to 6) also tend to excel at those indicating minimal impairment (7 to 10).

Tasks B & C: Table 4 shows results for post-level and timeline-level summarization. Almost all teams used LLMs, except team ResBin who fine-tuned long-context PLMs. Over half of the teams incorporated predictions from post-level tasks as additional well-being signals for summarization.

6.4 Performance Analysis and Discussion

Maladaptive vs Adaptive states: Figure 2 summarizes the post-level self-state summary performance across the best runs per team with respect to the number of labeled evidence spans in the test set. When the number of *adaptive* evidence spans in a post changes, the average team performance remains largely the same. By contrast, when the number of *maladaptive* evidence increases, performance increases. While this trend holds when the total number of spans increases, the mean consistency of the summaries clearly benefits from more maladaptive evidence spans. This uncovers model limitations; they more easily synthesize negative aspects compared to positive ones, potentially due to the latter being more subtle.

A closer look into the submissions in terms of *adaptive* self-state spans identification reveals that the top-performing teams either leverage large 70B LLMs with carefully selected demonstrations, through RAG or otherwise, or leverage the fine-tuning of much smaller models such as RoBERTa with data augmentation. By contrast for *maladaptive* self-states, while few-shot learning with 70B models and PLMs continued to work well, smaller LLM prompting in the range of 7-9B parameters achieved top performance. Furthermore, top systems perform better at capturing maladaptive (.71) compared to adaptive (.59) evidence. These results demonstrate that consistent with psychology literature (Baumeister et al., 2001), current LLMs, especially smaller ones, remain challenged by the task of identifying nuanced and subtle positive experiences, compared to negative experiences which are generally more salient and attention-grabbing.

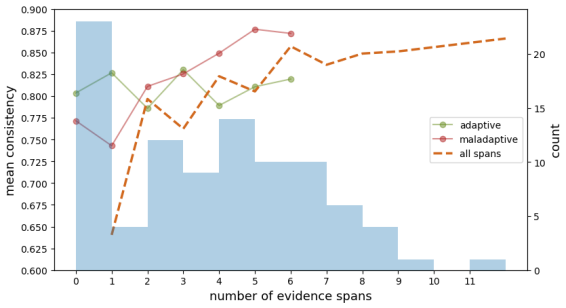

Figure 2: Post-level summarization performance in terms of average mean consistency across all teams with respect to the number of adaptive, maladaptive and total evidence spans (left) and histogram of posts per total evidence span (right).

Well-being Scores: Figure 3 provides different views to performance for well-being. As shown in the top boxplot, average MSE is the lowest (i.e. *best*) in the minimal impairment to functioning (scores 7-10) group. However, the middle line chart shows that the aggregate performance in terms of post-level self-state summarization for this group is *worse* compared to the group with serious impairment to functioning (1-4; left) and impaired functioning (5-6; middle). The bottom boxplot shows that posts in this group have the lowest median number of evidence spans. These results suggest that score prediction is differently impacted by the absence of self-state evidence compared to precise span extraction and summarization tasks;

posts with fewer evidence (and especially adaptive rather than maladaptive evidence) may be harder to summarize (Figure 2), but not necessarily harder to score on a well-being scale.

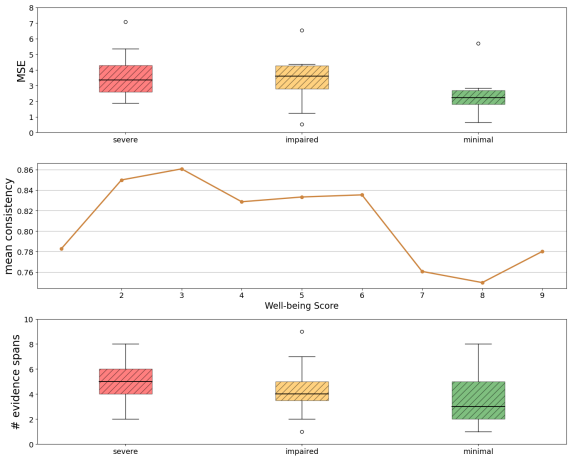

Figure 3: Distribution of average MSE per well-being score functioning bin per team (top), average mean consistency of post-level summary across well-being scores per team (middle), and number of labeled evidence spans per well-being bin (bottom).

Temporality and Well-being Score: In order to assess the temporal nature of well-being score's evolution in an individual's timeline, we compare this year's CLPsych annotations with CLPsych 2022's (Tsakalidis et al., 2022a) MoC annotations which are by definition longitudinal. For posts that do not present enough relevant information about an individual's mental state we assigned the latest recorded well-being score. We calculate the Spearman's correlation between the clinically annotated well-being score and the MoC seeing a statistically significant weak negative correlation (-.38). We further obtain a more longitudinal well-being score version by calculating the absolute well-being fluctuation between consecutive posts. The correlation of this variable with MoC is a stronger (significant) moderate correlation (.44). These results presented in Table 5, suggest that the currently provided well-being scores are of weak longitudinal nature, also manifesting through the lower performance of the BiLSTM baseline compared to the BERT Post-level baseline in Table 3.

Additionally, we append together statistics from the: current CLPsych 2025 Task A.2, CLPsych 2022 MoC and TalkLife MoC Dataset from Tseriotou et al. (2024a) in Table 6. Since the mean absolute well-being fluctuation is .91 and the standard deviation is .75, we define change in terms of fluctuations to be larger or equal to 2. As shown in the table, Reddit dataset changes are less frequent than TalkLife ones while their timelines span a longer period of time. Comparing this year's labels with other datasets', well-being score changes are considerably sparser than MoC (i.e. Switches and Escalations combined) and each change on average spans a longer time period, potentially limiting the degree of longitudinality of the well-being score prediction task. These findings may account for the lower performance of teams that attempted to tackle well-being score prediction in a temporal way (CIOL, Seq2Psych).

Variable	Spearman's Corr.	p-value
Well-being Score	-0.375	$4.8e^{-16}$
Well-being Fluctuation	0.440	$7.9e^{-12}$

Table 5: Spearman's correlation of well-being scores and fluctuation between consecutive posts with respect to corresponding *Moments of Change* labels.

Dataset	Reddit (current) Well-being	Reddit MoC		TalkLife MoC	
	Change	Switch	Escalation	Switch	Escalation
Mean Point Time Diff.	4d 14hr 27min	2d 6hr 58min		6hr 51min 11sec	
Median Point Time Diff.	2d 19hr 11min	22hr 42min 55sec		59min 38sec	
Mean consecutive events	1.39	1.19	2.83	1.58	4.12
Median consecutive events	1	1	2	1	3
Mean events in timeline	2.30	1.60	3.85	1.77	4.03
Median events in timeline	2	1	2	1	1

Table 6: Well-being and MoC statistics of datasets on time and event length.

7 Conclusion

Expanding on previous shared tasks, we introduced a novel multi-task framework grounded in the transtheoretical MIND approach. Participants were asked to identify adaptive and maladaptive self-states (Task A.1), predict post-level well-being scores (Task A.2), and generate post- and timeline-level summaries that reflect psychological progression (Tasks B and C). Systems using LLMs were able to identify both adaptive and maladaptive states although an asymmetry was observed in favor of maladaptive states.

Future directions could address the more longitudinal nature of well-being by reformulating the task towards a more temporal one and exploring temporal models that focus on capturing sparser and more subtle changes over longer time periods as well as amplify the signal of adaptive behavior which is important in achieving and monitoring better therapeutic outcomes.

Limitations

As in the vast majority of prior work leveraging social media for individual-level mental health assessments, this year's shared task involves individuals who generated content in self-selected online communities. The present tasks were conducted using social media posts made on various mental health-related subreddits in the English language, by users who willingly self-disclosed their thoughts and feelings. Generalization of the approaches presented in this work to other contexts and in other languages remains an open area of research.

Annotation was performed over 40 relatively short timelines due to the annotation load for clinical experts. This potentially hinders the performance of smaller supervised models, still leaving open questions around their true potential. Additionally, although the well-being score was annotated on the post-level with full timeline content visibility, the longitudinal manifestation of individuals' well-being remains underexplored. Since the annotation process involved selection of the most salient available adaptive and maladaptive spans for each ABCD element, this task does not yet explore the more nuanced selection of additional evidence spans and their connection to one another.

Although the dynamic evolution of self-states was to some extent addressed in this work with respect to summarization, there is still need to explore such dynamic progression through the lenses of other tasks such as monitoring and dialogue tracking. Finally, multimodality, which provides important cues especially in the clinical setting in terms of the manifestation of self-states, remains for now a future direction.

Ethics

This year's tasks explored the prediction of well-being scores from online posts of users over time, as well as the extraction of adaptive and maladaptive evidence spans and further summarization of self-state information at the post and timeline levels. This multi-task framework is grounded in the MIND scheme (Slonim, 2024) that views human experience as consisting of self-states fluctuating over time. Each self-state constitutes of identifiable units characterized by specific combinations of Affect, Behavior, Cognition, and Desire (ABCD).

While the evidence extraction and summaries provide some guidance with respect to ABCD elements and maladaptive and adaptive states, this cannot be used for diagnostic purposes, especially without the involvement of human experts. Adaptive and maladaptive evidence extracted by such models should be reviewed by clinical experts or used in the loop to augment their capacity by efficiently presenting information to them.

Additionally, the task cannot make any claims about the potential evidence providing explanations for well-being scores. Rather, it forms a research direction towards making causal links between the two, paving the way towards language models that can better reason along their decision making process.

In terms of data, even though we are using publicly available content from Reddit, we prohibited its redistribution and the use of any third-party LLMs that would require sending (part of) the information to the provider's servers, to ensure protection of the sensitive content.

Acknowledgements

This work was supported by a UKRI/EPSRC Turing AI Fellowship to Maria Liakata (grant ref EP/V030302/1) and the Alan Turing Institute (grant ref EP/N510129/1). This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible Ai UK (KP0016) as a Keystone project lead by Maria Liakata. The shared task organizers would like to express their gratitude to the anonymous users of Reddit whose data feature in this year's shared task dataset; to the clinical experts from Bar-Ilan University who annotated the data for all tasks; to all team members for their participation; and to NAACL for its support for CLPsych.

References

Prottay Kumar Adhikary, Aseem Srivastava, Shivani Kumar, Salam Michael Singh, Puneet Manuja, Jini K Gopinath, Vijay Krishnan, Swati Kedia Gupta, Koushik Sinha Deb, and Tanmoy Chakraborty. 2024. Exploring the efficacy of large language models in summarizing mental health counseling sessions: benchmark study. *JMIR Mental Health*, 11:e57306.

Mostafa M Amin, Erik Cambria, and Björn W Schuller. 2023. Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt. *IEEE Intelligent Systems*, 38(2):15–23.

Anson Antony and Annika Marie Schoene. 2025. Retrieval-enhanced mental health assessment: Capturing self-state dynamics from social media using in-

context learning. In *Proceedings of the Tenth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Elham Asgari, Nina Montana-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, and Dominic Pimenta. 2024. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *medRxiv*, pages 2024–09.

American Psychiatric Association et al. 2000. Diagnostic and statistical manual of mental disorders iv-tr washington. *DC: American Psychiatric Association*.

Abhin B and Renukasakshi V Patil. 2025. Transformer-based analysis of adaptive and maladaptive self-states in longitudinal social media data. In *Proceedings of the Tenth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Roy F Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs. 2001. Bad is stronger than good. *Review of general psychology*, 5(4):323–370.

Ulya Bayram and Lamia Benhiba. 2022. Emotionally-informed models for detecting moments of change and suicide risk levels in longitudinal social media data. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 219–225, Seattle, USA. Association for Computational Linguistics.

Aaron T Beck, Molly R Finkel, and Judith S Beck. 2021. The theory of modes: Applications to schizophrenia and other psychological conditions. *Cognitive Therapy and Research*, 45:391–400.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Philip M Bromberg. 2014. *Standing in the spaces: Essays on clinical process trauma and dissociation*. Routledge.

Suchandra Chakraborty, Sudeshna Jana, Manjira Sinha, and Tirthankar Dasgupta. 2025. Self-state evidence extraction and well-being prediction from social media timelines. In *Proceedings of the Tenth Workshop on Computational Linguistics and Clinical Psychology*. 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 Tenth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Jenny Chim, Julia Ive, and Maria Liakata. 2025. Evaluating synthetic data generation from user generated text. *Computational Linguistics*, 51(1):191–233.

Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the CLPsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 177–190, St. Julians, Malta. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, Imre Solti, et al. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Muskan Garg. 2024. Wellxplain: Wellness concept extraction and classification in reddit posts for mental health analysis. *Knowledge-Based Systems*, 284:111228.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Anthony Hills, Adam Tsakalidis, and Maria Liakata. 2023. Time-aware predictions of moments of change in longitudinal user posts on social media. In *International Workshop on Advanced Analytics and Learning on Temporal Data*, pages 293–305. Springer.

Anthony Hills, Talia Tseriotou, Xenia Miscouridou, Adam Tsakalidis, and Maria Liakata. 2024. Exciting mood changes: A time-aware hierarchical transformer for change detection modelling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12526–12537, Bangkok, Thailand. Association for Computational Linguistics.

Stephanie Homan, Marion Gabi, Nina Klee, Sandro Bachmann, Ann-Marie Moser, Martina Duri', Sofia Michel, Anna-Marie Bertram, Anke Maatz, Guido Seiler, Elisabeth Stark, and Birgit Kleim. 2022. Linguistic features of suicidal thoughts and behaviors: A systematic review. *Clinical Psychology Review*, 95:102161.

Md. Iqramul Hoque, Mahfuz Ahmed Anik, and Azmine Toushik Wasi. 2025. Ciol at clpsych 2025: Using large lanuage models for understanding and summarizing clinical texts. In *Proceedings of the Tenth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jihoon Kwon Sangmo Gu Yejin Kim Minkyung Cho Jy-yong Sohn Chanyeol Choi Junseong Kim, Seolhwa Lee. 2024. Linq-embed-mistral:elevating text retrieval with improved gpt data through task-specific control and quality refinement. Linq AI Research Blog.

Laerdon Kim. 2025. A baseline for self-state identification and classification in mental health data: Clpsych 2025 task. In *Proceedings of the Tenth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Ayal Klein, Jiayu Song, Jenny Chim, Liran Keren, Andreas Triantafyllopoulos, Björn W Schuller, Maria Liakata, and Dana Atzil-Slonim. 2024. Clinical insights from social media: Assessing summaries of large language models and humans.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Gal Lazarus and Eshkol Rafaeli. 2023. Modes: Cohesive personality states and their interrelationships as organizing concepts in psychopathology. *Journal of Psychopathology and Clinical Science*, 132(3):238.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chandreen Liyanage, Muskan Garg, Vijay Mago, and Sunghwan Sohn. 2023. Augmenting Reddit posts to determine wellness dimensions impacting mental health. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 306–312, Toronto, Canada. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 47684777, Red Hook, NY, USA. Curran Associates Inc.

Yu A Malkov. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Tobias Mayer, Neha Warikoo, Amir Eliassaf, Dana Atzil-Slonim, and Iryna Gurevych. 2024. Predicting client emotions and therapist interventions in psychotherapy dialogues. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1463–1477.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ummara Mumtaz, Awais Ahmed, and Summaya Mumtaz. 2024. Llms-healthcare : Current applications and challenges of large language models in various medical specialties.

Matthew D Nemesure, Michael V Heinz, Raphael Huang, and Nicholas C Jacobson. 2021. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific reports*, 11(1):1980.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8446–8459, Dublin, Ireland. Association for Computational Linguistics.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2023. Overview of erisk 2023: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 294–315. Springer.

Jason Phang, Yao Zhao, and Peter Liu. 2023. Investigating efficiently extending transformers for long input summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3946–3961, Singapore. Association for Computational Linguistics.

Federico Ravenda, Fawzia-Zehra Kara-Isitt, Stephen Swift, Antonietta Mira, and Andrea Raballo. 2025. From evidence mining to meta-prediction: a gradient of methodologies for task-specific challenges in psychological assessment. In *Proceedings of the Tenth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

William Revelle. 2007. Experimental approaches to the study of personality. *Handbook of research methods in personality psychology*, pages 37–61.

Anastasia Sandu, Teodor Mihailescu, Ana Sabina Uban, and Ana-Maria Bucur. 2025. Capturing the dynamics of mental well-being: Adaptive and maladaptive states in social media. In *Proceedings of the Tenth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Ramit Sawhney, Shivam Agarwal, Atula Tejaswi Neerkaje, Nikolaos Aletras, Preslav Nakov, and Lucie Flek. 2022a. Towards suicide ideation detection through online conversational context. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1716–1727.

Ramit Sawhney, Atula Tejaswi Neerkaje, and Manas Gaur. 2022b. A risk-averse mechanism for suicidality assessment on social media. *Association for Computational Linguistics 2022 (ACL 2022)*.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36.

Gopendra Singh, Sai Vemulapalli, Mauajama Firdaus, and Asif Ekbal. 2024. Deciphering cognitive distortions in patient-doctor mental health conversations: A multimodal llm-based detection and reasoning framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22546–22570.

Dana Atzil Slonim. 2024. Self-other dynamics (sod): A transtheoretical coding manual.

Jiayu Song, Mahmud Akhter, Dana Atzil Slonim, and Maria Liakata. 2025. Temporal reasoning for timeline summarisation in social media.

Jiayu Song, Jenny Chim, Adam Tsakalidis, Julia Ive, Dana Atzil-Slonim, and Maria Liakata. 2024a. Clinically meaningful timeline summarisation in social media for mental health monitoring.

Jiayu Song, Jenny Chim, Adam Tsakalidis, Julia Ive, Dana Atzil-Slonim, and Maria Liakata. 2024b. Combining hierachical VAEs with LLMs for clinically meaningful timeline summarisation in social media. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14651–14672, Bangkok, Thailand. Association for Computational Linguistics.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H Andrew Schwartz. 2022. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 622–636.

Nikita Soni, August Håkan Nilsson, Syeda Mahwish, Vasudha Varadarajan, H. Andrew Schwartz, and Ryan L. Boyd. 2025. Who we are, where we are: Mental health at the intersection of person, situation, and large language models. In *Proceedings of the Tenth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Sajad Sotudeh, Nazli Goharian, and Zachary Young. 2022. Mentsum: A resource for exploring summarization of mental health online posts.

Aseem Srivastava, Tharun Suresh, Sarah Peregrine, Lord, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling summarization using mental health knowledge guided utterance filtering.

William B Stiles. 2001. Assimilation of problematic experiences. *Psychotherapy: Theory, Research, Practice, Training*, 38(4):462.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Vu Tran and Tomoko Matsui. 2025. Team ism at clpsych 2025: Capturing mental health dynamics from social media timelines using a pretrained large language model with in-context learning. In *Proceedings of the Tenth Workshop on Computational Linguistics and Clinical Psychology*. 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, et al. 2022a. 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.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022b. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660.

Talia Tseriotou, Ryan Chan, Adam Tsakalidis, Iman Munire Bilal, Elena Kochkina, Terry Lyons, and Maria Liakata. 2024a. Sig-networks toolkit: Signature networks for longitudinal language modelling. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 223–237.

Talia Tseriotou, Adam Tsakalidis, Peter Foster, Terence Lyons, and Maria Liakata. 2023. Sequential path signature networks for personalised longitudinal language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5016–5031.

Talia Tseriotou, Adam Tsakalidis, and Maria Liakata. 2024b. Tempoformer: A transformer for temporally-aware representations in change detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19635–19653.

Zimu Wang, Hongbin Na, Rena Gao, Jiayuan Ma, Yining Hua, Ling Chen, and Wei Wang. 2025. From posts to timelines: Modeling mental health dynamics from social media timelines with hybrid llms. In *Proceedings of the Tenth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

WHO. 2022. *World mental health report: Transforming mental health for all*. World Health Organization.

Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Hong Yu, James Hendler, Anind K Dey, and Dakuo Wang. 2023. Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385*.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024a. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1).

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024b. 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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ayah Zirikly and Mark Dredze. 2022. Explaining models of mental health via clinically grounded auxiliary tasks. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 30–39.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology.

Appendix A Annotation

A.1 Annotation Interface

Figure 4 shows a screenshot of the annotation platform INCEpTION (Klie et al., 2018), which we adapted for our task. INCEpTION provides a user-friendly interface that enables annotators to efficiently assign labels and categories directly onto text segments. By customizing the annotation schema and label sets, we streamlined the annotation process, enhancing precision and consistency aligned with our research objectives.

A.2 MIND Framework

Table 10 shows categories and sub-categories within the framework.

A.3 Further IAA Measures

We complement the standard inter-annotator agreement measures reported in Section 4.2 with additional consistency metrics.

We first calculate relaxed pairwise F1 scores over spans identified in Task A.1 (State Evidence), following the previous CLPsych shared task (Chim et al., 2024) and established practices (Hripcsak and Rothschild, 2005; Deleger et al., 2012). In this relaxed metric, a minimal overlap of one token between spans is considered a match. Results are summarized in Table 7. These values indicate lower agreement compared to CLPsych 2024, likely due to the broader, more comprehensive nature of our task. Furthermore, we calculate the agreement between the annotators using the same metrics employed for system evaluation, recognizing that inter-annotator agreement serves as an essential reference point or upper bound for assessing system performance.

Span Type	F1
Adaptive spans	.51
Maladaptive spans	.58
Overall (micro-average)	**.56**

Table 7: Relaxed pairwise F1 agreement for Task A.1.

Our findings, detailed in Table 8, emphasize the complexity of the annotation tasks and underscore that achieving high agreement is challenging even for clinically informed annotators. Thus, achieving performance close to these inter-annotator agreement values can be considered as approaching the maximum attainable performance for these tasks.

Metric	Task	Value
BERTScore recall	A.1	0.469
BERTScore weighted recall	A.1	0.387
Mean squared error (MSE)	A.2	2.913
Macro F1	A.2	0.403

Table 8: Inter-annotator agreement over 23 posts using the system evaluation metrics. Note that recall measures were averaged across both comparison directions to reflect the symmetric nature of inter-annotator agreement.

Appendix B Participant Submissions

This section presents an overview of the registration process (§B.1), individual systems (§B.2) from each participating team and provides an overview of methods (§B.3).

B.1 Registration Process

The registration process consisted of three stages: a) completing an individual and a team registration through an online form, b) reading and signing a data sharing agreement, and c) receiving access instructions for training data stored in a password-protected compressed file. During stage a) the organizing team assisted participants looking for collaborators in the team-forming process. For b), the data sharing agreement asked the teams to determine the password-protected private storage of the data, while restricting explicit or implicit data distribution through third party LLM platforms.

Team	#Members	Task A	Task B	Task C	Paper submitted
Aquarius	5	2	2	2	✓
BLUE	3	3	3	3	✓
BULUSI	2	3	3	3	✓
CIOL	3	1	2	2	✓
CSIRO-LT	4	3	-	-	
EAIonFlux	2	2	2	2	✓
ISM	2	2	2	2	✓
MMKA	4	2	-	-	✓
NoviceTrio	3	1	1	1	
PsyMetric	2	1	1	1	
ResBin	2	1	1	1	✓
Seq2Psych	6	3	-	-	✓
uOttawa	4	3	3	3	✓
Zissou	7	1	1	-	✓
Total	**49**	**28**	**21**	**20**	**11**

Table 9: Team information and submissions for the CLPsych 2025 shared task.

Each team was allowed up to three submissions for the official team results. Additional submissions were allowed in order to facilitate ablation and further analysis by the teams. Upon receiving the submissions, results were returned within 24 hours based on our evaluation metrics (§5) on a test set of 10 timelines (§4). A summary of the team

specifics including the number of submissions is provided in Table 9.

B.2 Individual Team Submissions

Aquarius Wang et al. (2025) integrated extracted evidence to guide well-being score prediction and summarization. For evidence span identification, they used fine-tuned Qwen2.5-7b (Yang et al., 2024a) to explore a sentence classification and a span generation approach. Then, they combined the content of each post with extracted evidences for well-being score prediction. For post-level and timeline-level summarization, the team employed Qwen2.5-32B, using as input post content, extracted evidences, predicted score(s), and a retrieved annotated example from the train set that had the highest embedding similarity.

BLUE Sandu et al. (2025) utilized a range of LLMs, prompting strategies, and machining learning approaches to the tasks. For evidence extraction they achieved the highest recall using Gemma 2 9B (Team et al., 2024) coupled with a default prompt, providing instruction for the task without including definitions of concepts or additional context, while the same model performs the best for well-being scoring using an expert prompt based on emotional, cognitive, and behavioral indicators. For post-level and timeline-level summarization, LLaMA 3.2 3B (Grattafiori et al., 2024) utilizing the default prompt performed best.

BULUSI For evidence extraction, Ravenda et al. (2025) formed candidate segments and then extracted the most relevant ones using retrieval based on the training evidence which were fed in three (22-72B) LLMs for consensus self-state classification with in-context learning. For well-being score prediction, the team explored three strategies for aggregating LLM predictions: an average ensemble, an Oracle-style meta-model, and an optimized weighted ensemble minimizing mean squared error while accounting for missing values. The optimized ensemble yielded the best result. Finally, for post and timeline-level summarization the team used post(s) content, predicted self-state segments, and retrieved top five relevant example posts to prompt the LLM.

CIOL Hoque et al. (2025) extracted evidence spans of adaptive and maladaptive self-states using Random Forest classifiers on thousands of TF-IDF features. For well-being score prediction they formulated a supervised approach through Gradient Boosting regression on sentiment and ratio-based features reflecting the relationship between adaptive and maladaptive evidence. For post-level summaries they DPO-finetuned Qwen2.5-7B-Instruct-1M (Yang et al., 2024a). This was followed by a few-shot prompting strategy guiding the model to identify the dominant self-state determining the ABCD elements based on evidence spans and well-being scores. This approach was extended on the timeline-level by fine tuning the post-level model above on timeline-level examples. Then they used this model to generate summaries based on a narrative arc analysis framework that treats each timeline as a psychological development trajectory. For an extension of the timeline-level summary their prompt directs the model to identify temporal self-state patterns and changes, highlighting key transitions between states.

EAIonFlux Antony and Schoene (2025) proposed systems based on vector similarity retrieval of relevant in-context demonstrations for LLM prompting. They used LLaMA 3.3 70B (Grattafiori et al., 2024) and experimented with different numbers of retrieved examples. They built one post-level and one timeline-level vector database (capturing temporal patterns) out of the training data embedding them through Linq-Embed-Mistral (Junseong Kim, 2024) to capture emotional content. Retrieval is based on cosine similarity with HNSW (Malkov, 2018) for fast nearest neighbor search. For each task, a task-specific module generates prompts and processes outputs tailored to the different objectives, and predictions from previous task(s) are integrated into the next ones.

ISM Tran and Matsui (2025) explored in-context learning with Llama-3-8B, using random sampling followed by Gaussian kernel density estimation to select training data instances as demonstration. The team jointly modeled post-level tasks in the same prompt, and focused on summarization only in the prompt for timeline-level generation.

MMKA Chakraborty et al. (2025) focused on Tasks A.1 and A.2. They fine-tuned a RoBERTa classification model (Liu et al., 2019) to extract adaptive and maladaptive self-states at the token level, augmenting the training data using `nlpaug` and implementing post-processing to obtain the most frequent label per sentence. For well-being score generation (with a justification generation),

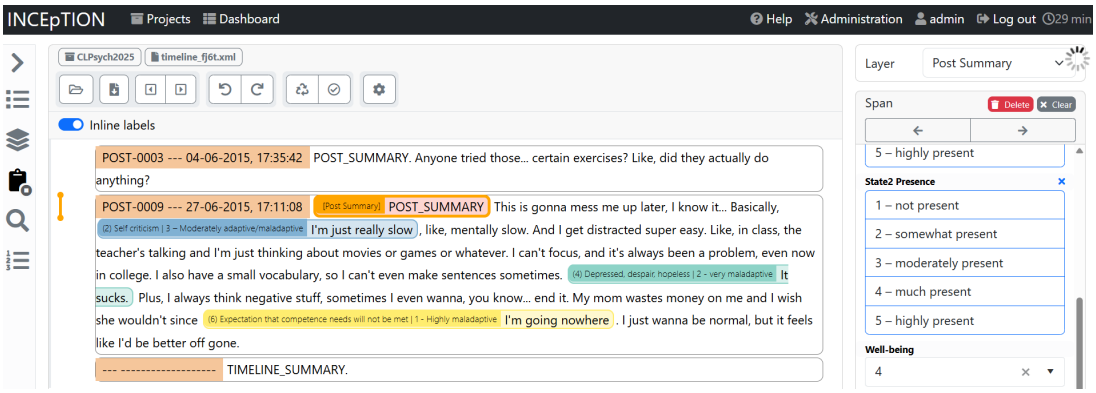

Figure 4: A screenshot from our annotation interface, leveraging the INCEpTION platform. Example timeline is reduced and paraphrased due to the sensitive nature of the data.

Category		Sub-Categories	
		Adaptive Example	Maladaptive Example
Affect	Type of emotion expressed by a person.	Calm/Laid back, Emotional Pain/Grieving, Content/Happy, Vigor/Energetic, Justifiable Anger/Assertive Anger, Proud	Anxious/Tense/Fearful, Depressed/Desperate/ Hopeless, Mania, Apathetic/Don't care/Blunted, Angry (Aggressive, Disgust, Contempt), Ashamed/Guilty
Behavior	**Behavior of the self with the Other (BO)** The person's main behavior(s) toward the other.	Relating behavior, Autonomous behavior	Fight or flight behavior, Overcontrolled/controlling behavior
	Behavior toward the Self (BS) The person's main behavior(s) toward the self.	Self-care behavior	Self-harm/Neglect/ Avoidance behavior
Cognition	**Cognition of the Other (CO)** The person's main perceptions of the other.	Perception of the other as related, Perception of the other as facilitating autonomy/ competence needs	Perception of the other as detached or over attached, Perception of the other as blocking autonomy needs
	Cognition of the Self (CS) The person's main self-perceptions.	Self-acceptance and self-compassion	Self-criticism
Desire	The person's main desire, need, intention, fear or expectation.	Relatedness, Autonomy and adaptive control, Competence, Self-esteem, Self-care	Expectation that relatedness need will not be met, Expectation that autonomy needs will not be met, Expectation that competence needs will not be met

Table 10: ABCD elements (Categories) with explanations, and their sub-categories.

they used RAG to select the top-k most similar posts to the current one using all-MiniLM-L6-v2 for embeddings. These were included for in-context learning with DeepSeek-7B (Bi et al., 2024). In post-hoc analysis, they found random forest regression yielding better results than the LLM approach.

ResBin B and Patil (2025) explored diverse approaches: they used Mixtral-8x7b (Jiang et al., 2024) for evidence extraction, obtained embeddings from PLMs such as MentalBERT to train random forest classifiers for well-being score prediction, fine-tuned Longformer (Beltagy et al., 2020) for post-level summarization with predicted evidences and post content as input, and fine-tuned Pegasus-X-Large (Phang et al., 2023) for timeline-level summarization with timeline content as input.

Seq2Psych Soni et al. (2025) focused on Task A, leveraging principled baseline features, such as Situational 8 DIAMONDS (S8D) and Person-Level Traits (PLT) including resilience quantification utilizing the Resilience through Language Modeling (ReLM) framework. For S8D, the team used the Deepseek-R1 (Guo et al., 2025) model with few-shot prompting to infer eight situational dimensions at the post-level. Using different feature combinations they fine-tuned the HaRT model (Soni et al., 2022) which processes temporal user language, to generate person-contextualized embeddings towards well-being score prediction and sentence-level binary adaptive/maladaptive classification.

uOttawa Chan et al. (2025) applied prompt engineering techniques with Llama-3.3-70B-Instruct to address all four subtasks. They compared variedly structured zero-shot, one-shot, and few-shot prompts, finding one-shot to be most performant for evidence extraction and post-level summarization, and few-shot to be best for well-being score prediction and timeline-level summarization.

Zissou Kim (2025) prompted a 4-bit quantized Gemma-2 9b (Team et al., 2024) with few-shot learning and presented an approach that explored the impact of preprocessing on span extraction. Each post was divided into sentences, identifying only the important sentences, and then classified

as adaptive or maladaptive through prompting with self-state definitions. Providing the previous sentence context improved performance. For the other tasks, they generated post summaries and well-being scores based on the list of classifications and post text also with few-shot prompting.

B.3 Overview

We outline methods used in the *best run per team* in Table 11. For a complete picture of each team's approaches, including ablations, please refer to their respective paper. We consider whether a system:

- **LLM**: Uses a large language model.
- **PLM**: Uses a pretrained language model.
- **ML**: Uses traditional machine learning, focusing on algorithms (e.g. random forest) and excluding techniques (e.g. feature engineering).
- **RAG**: Uses retrieval augmented generation, focusing on automatic retrieval and excluding manually selected examples.
- **Pipeline**: Involves an approach where predictions from at least one task are used in predictions for another. Excludes joint modeling that happens in the same step.
- **Domain**: Involves explicit domain knowledge *beyond* what was provided in the shared task documents provided to all participants.
- **Temporal**: Involves explicit modeling of temporality/relationship between posts within a single timeline in Tasks A and B. Excludes cases of contextual modeling within an individual post (e.g. between sentences in one post).

Team	LLM	PLM	ML	RAG	Pipeline	Domain	Temporal
Aquarius	✓				✓		
BLUE	✓						
BULUSI	✓			✓	✓		
CIOL	✓		✓		✓	✓	✓
EAIonFlux	✓			✓	✓		
ISM	✓		✓	✓			
MMKA	✓	✓		✓			
ResBin	✓	✓	✓		✓		
Seq2Psych	✓	✓				✓	✓
uOttawa	✓						
Zissou	✓				✓		
Total	11	3	3	4	6	2	2

Table 11: Methods in each team's top submission.

LLMs Every team in this year's shared task used an LLM to tackle at least one subtask. Focusing on the *best run* per team, we categorize the types of models used, counting each model once per submission per team. We summarize the model type (Figure 5), context length (Figure 6), and size in terms of parameter count (Figure 7).

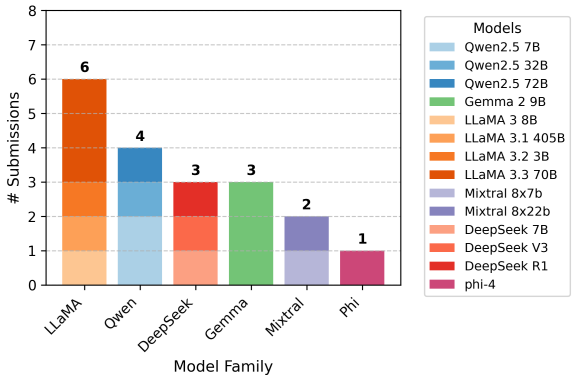

Figure 5: LLMs used in best runs of official submissions, grouped by model family and lineage.

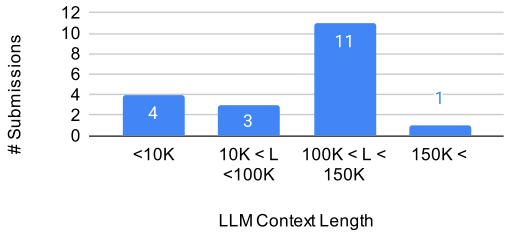

Figure 6: Maximum number of tokens that can be fed into the employed models.

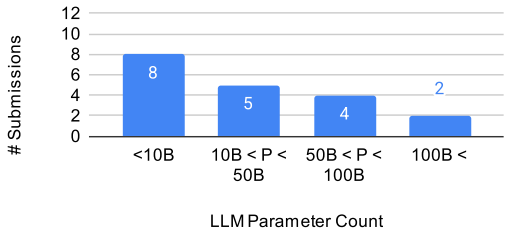

Figure 7: Size of employed models. We use the active parameter count for mixture-of-expert models.

Compared to the previous shared task which explored how well open LLMs can be leveraged to identify and synthesize textual evidences (Chim et al., 2024), we see noticeable increases in (1) model family diversity, (2) context length, and (3) use of retrieval augmented generation.

While LLaMA remains the most popular model family, many submissions leveraged alternatives such as Qwen and DeepSeek. The maximum theoretical context length that can be handled has increased from 32K to 1M tokens, and the majority

of models used in the top submission this year have a context length in the 100K to 150K token range. These changes highlight the rapid technological progress in open LLMs and long context models, as well as opportunities to advance mental health modeling over longer horizons.

Appendix C Baselines

This section outlines implementation details of our baseline models (§6.2). In LLM-based methods, we employ `Llama 3.1-8B-Instruct`.

C.1 Task A.1

For the *Llama ZS Single-Post* and *Llama ZS 5-Post* baselines we used the prompts presented in Listings 1 and 2 respectively. The LLM generated outputs in JSON. We used top-p sampling with temperature (p=.9, t=.01), permitting decoding up to 550 new tokens.

After hyperparameter tuning the following parameters were used for *BART Single-Post* and *BART 5-Post*: learning rate $= 2e{-}5$, epochs $= 10$. BERTScore recall was used for best model selection. We used `BART-base`.[4]

C.2 Task A.2

For the *Llama ZS Single-Post* and *Llama ZS 5-Post* baselines we used the prompts presented in Listings 3 and 4 respectively. The LLM generated outputs in JSON. We use top-p sampling with temperature (p=.9, t=.01), permitting decoding up to 100 new tokens.

BERT Post-level was fine-tuned end-to-end with a regression head to predict well-being scores as a regression task. After hyperparameter tuning the following parameters were used for this model: learning rate $= 2e{-}5$, epochs $= 15$. We used `BERT-base-uncased` [5]. MSE was used for best model selection.

For the *BiLSTM 5-Post (BERT)* model, the following parameters were selected after hyperparameter tuning: num_layers$= 1$, dropout$= .25$, hidden_size$= 100$, learning rate $= 1e{-}4$, epochs$= 100$. The `[CLS]` BERT representation token for each post was produced using off-the-shelf `BERT-base-uncased`. While the 5 latest posts (4 recent + current) were used for the BiLSTM sequence, only the score for the current post was predicted. In the absence of 4 posts in recent history padding was used. MSE was used for best model selection.

C.3 Task B & C

For Task B the *Llama ZS Summary* and *Llama ZS w/ Intermediate Summary* baselines we used the prompts presented in Listings 5 and 6 respectively. The corresponding prompts for Task C are presented in Listings 7 and 8 respectively. We use top-p sampling with temperature (p=.9, t=.1), permitting decoding up to 300 new tokens.

[4] https://huggingface.co/facebook/bart-base

[5] https://huggingface.co/google-bert/bert-base-uncased

```
Your goal is to identify and extract any sentence from the following social media
post which demonstrates an adaptive or maladaptive self-state of the user.

Definitions:
- **Adaptive Self-State** pertains to aspects of Affect, Behaviour, and Cognition
    towards the self or others, which is conducive to the fulfillment of basic
    desires/needs, such as relatedness, autonomy, and competence.
- **Maladaptive Self-State** pertains to aspects of Affect, Behaviour, and Cognition
     towards the self or others that hinder the fulfillment of basic desires/needs.

Instructions:
- Extract only the specific sentence spans from the post.
- Do not return these instructions or the entire post in your output.

Post Content: {post}
Output:

{
      "Adaptive": ["list of adaptive spans"],
      "Maladaptive": ["list of maladaptive spans"]
}
```

Listing 1: Prompt for Task A.1's Llama ZS Single-Post baseline.

```
Your goal is to identify and extract any sentence from the following social media
posts which demonstrates an adaptive or maladaptive self-state of the user.

Definitions:
- **Adaptive Self-State** pertains to aspects of Affect, Behaviour, and Cognition
    towards the self or others, which is conducive to the fulfillment of basic
    desires/needs, such as relatedness, autonomy, and competence.
- **Maladaptive Self-State** pertains to aspects of Affect, Behaviour, and Cognition
     towards the self or others that hinder the fulfillment of basic desires/needs.

Instructions:
- Extract only the specific sentence spans from the last post.
- Do not return these instructions or the entire post in your output.

Post Content: {posts}
Output:
{
      "Adaptive": ["list of adaptive spans"],
      "Maladaptive": ["list of maladaptive spans"]
}
```

Listing 2: Prompt for Task A.1's Llama ZS 5-Post baseline.

```
Your goal is to analyse and score the following social media post according to the
wellbeing scale below.

# Wellbeing Scale
- **10** No symptoms and superior functioning in a wide range of activities.
- **9** Absent or minimal symptoms (eg., mild anxiety before an exam), good
    functioning in all areas, interested and involved in a wide range of activities.
- **8** If symptoms are present, they are temporary and expected reactions to
    psychosocial stressors (eg., difficulty concentrating after family argument).
    Slight impairment in social, occupational or school functioning.
- **7** Mild symptoms (eg., depressed mood and mild insomnia) or some difficulty in
    social, occupational, or school functioning, but generally functioning well, has
     some meaningful interpersonal relationships.
- **6** Moderate symptoms (eg., panic attacks) or moderate difficulty in social,
    occupational or school functioning.
- **5** Serious symptoms (e.g., suicidal thoughts, severe compulsions) or serious
    impairment in social, occupational, or school functioning (eg., no friends,
    inability to keep a job).
- **4** Some impairment in reality testing or communication, or major impairment in
    multiple areas (withdrawal from social ties, inability to work, neglecting
    family, severe mood/thought impairment).
- **3** A person experiences delusions or hallucinations or serious impairment in
    communication or judgment or is unable to function in almost all areas (eg., no
    job, home, or friends).
- **2** In danger of hurting self or others (eg., suicide attempts; frequently
    violent; manic excitement) or may fail to maintain minimal personal hygiene or
    significant impairment in communication (e.g., incoherent or mute).
- **1** The person is in persistent danger of severely hurting self or others or
    persistent inability to maintain minimal personal hygiene or has attempted a
    serious suicidal act with a clear expectation of death.

Instructions:
- Only return the score for the entire post.
- Do not return these instructions or the entire post in your output.

Post Content: {post}
Output:
{
  "wellbeing scale": "score"
}
```

Listing 3: Prompt for Task A.2's Llama ZS Single-Post baseline.

```
Your goal is to analyse and score the following social media posts according to the
wellbeing scale below.

# Wellbeing Scale
- **10** No symptoms and superior functioning in a wide range of activities.
- **9** Absent or minimal symptoms (eg., mild anxiety before an exam), good
    functioning in all areas, interested and involved in a wide range of activities.
- **8** If symptoms are present, they are temporary and expected reactions to
    psychosocial stressors (eg., difficulty concentrating after family argument).
    Slight impairment in social, occupational or school functioning.
- **7** Mild symptoms (eg., depressed mood and mild insomnia) or some difficulty in
    social, occupational, or school functioning, but generally functioning well, has
     some meaningful interpersonal relationships.
- **6** Moderate symptoms (eg., panic attacks) or moderate difficulty in social,
    occupational or school functioning.
- **5** Serious symptoms (e.g., suicidal thoughts, severe compulsions) or serious
    impairment in social, occupational, or school functioning (eg., no friends,
    inability to keep a job).
- **4** Some impairment in reality testing or communication, or major impairment in
    multiple areas (withdrawal from social ties, inability to work, neglecting
    family, severe mood/thought impairment).
- **3** A person experiences delusions or hallucinations or serious impairment in
    communication or judgment or is unable to function in almost all areas (eg., no
    job, home, or friends).
- **2** In danger of hurting self or others (eg., suicide attempts; frequently
    violent; manic excitement) or may fail to maintain minimal personal hygiene or
    significant impairment in communication (e.g., incoherent or mute).
- **1** The person is in persistent danger of severely hurting self or others or
    persistent inability to maintain minimal personal hygiene or has attempted a
    serious suicidal act with a clear expectation of death.

Instructions:
- Only return the score for the entire last post.
- Do not return these instructions or the entire post in your output.

Post Content: {post}
Output:
{
  "wellbeing scale": "score"
}
```

Listing 4: Prompt for Task A.2's Llama 5-Post baseline.

```
Analyze the following social media post and identify the dominant self-state
(adaptive or maladaptive). Begin by determining which self-state is more dominant
and describe it first. For each self-state, highlight the central organizing aspect-
A (Affect), B (Behavior), C (Cognition), or D (Desire/Need)that drives the state.
Describe how this central aspect influences the other aspects, focusing on the
potential causal relationships between them. If the self-state is maladaptive,
explain how negative emotions, behaviors, or thoughts hinder psychological needs,
and if adaptive, explain how positive aspects support psychological needs. If both
adaptive and maladaptive states are present, describe each in turn. If only one
self-state is evident, focus solely on that. You must not make anything up. Keep the
description concise and only describe observations if they are fully supported by
the text.

Post Content: {post}
Summary:
```

Listing 5: Prompt for Task B's Llama ZS Summary baseline.

```
# Prompt 1 (General Summary)

Analyze the following social media post and generate the summary based on post
content. You must not make anything up. Keep the description concise and only
describe observations if they are fully supported by the text.

Post Content: {post}
Summary:


---------


# Prompt 2 (Self-State Analysis)

Analyze the following social media post summary and identify the dominant self-state
(adaptive or maladaptive). Begin by determining which self-state is more dominant
and describe it first. For each self-state, highlight the central organizing aspect-
A (Affect), B (Behavior), C (Cognition), or D (Desire/Need)that drives the state.
Describe how this central aspect influences the other aspects, focusing on the
potential causal relationships between them. If the self-state is maladaptive,
explain how negative emotions, behaviors, or thoughts hinder psychological needs,
and if adaptive, explain how positive aspects support psychological needs. If both
adaptive and maladaptive states are present, describe each in turn. If only one
self-state is evident, focus solely on that. You must not make anything up. Keep the
description concise and only describe observations if they are fully supported by
the text.

Post Summary: {post}
Final Summary:
```

Listing 6: Prompts for Task B's Llama ZS with Intermediate Summary baseline.

```
Generate a timeline-based summary analyzing the evolution of self-states across all
posts in chronological order. Emphasize the interplay between adaptive and
maladaptive self-states, focusing on temporal dynamics such as flexibility,
rigidity, improvement, and deterioration. Describe how the dominance of self-states
shifts over time, highlighting key emotional, cognitive, and behavioral changes that
contribute to these transitions. You must not make anything up. Keep the description
concise and only describe observations if they are fully supported by the text.

All Posts Content: {all_posts_concatenated}

Timeline Summary:
```

Listing 7: Prompt for Task C's Llama ZS Summary baseline.

```
# Prompt 1 (General Timeline Summary)

Generate a timeline-based summary analyzing all the posts in chronological order.
You must not make anything up. Keep the description concise and only describe
observations if they are fully supported by
the text.

All Posts Content: {all_posts_concatenated}

Timeline Summary:


---------


# Prompt 2 (Self-State Analysis over Timeline)

Generate a timeline-based summary analyzing the evolution of self-states across all
posts in chronological order. Emphasize the interplay between adaptive and
maladaptive self-states, focusing on temporal dynamics such as flexibility,
rigidity, improvement, and deterioration. Describe how the dominance of self-states
shifts over time, highlighting key emotional, cognitive, and behavioral changes that
contribute to these transitions. You must not make anything up. Keep the description
concise and only describe observations if they are fully supported by the text.

Post Summary: {timeline_summary}

Final Summary:
```

Listing 8: Prompts for Task C's Llama ZS with Intermediate Summary baseline.