Retrieval-Enhanced Mental Health Assessment: Capturing Self-State Dynamics from Social Media Using In-Context Learning

Anson Antony and Annika M. Schoene
The Institute for Experiential AI, Boston, MA
{a.antony, a.schoene}@northeastern.edu

Abstract

This paper presents our approach to the CLPsych 2025 (Tseriotou et al., 2025) shared task, where our proposed system implements a comprehensive solution using In-Context Learning (ICL) with vector similarity to retrieve relevant examples that guide Large Language Models (LLMs) without specific fine-tuning. We leverage ICL to analyze self-states and mental health indicators across three tasks. We developed a pipeline architecture using Ollama, where we are running Llama 3.3 70B locally and specialized vector databases for post- and timeline-level examples. We experimented with different numbers of retrieved examples (k=5 and k=10) to optimize performance. Our results demonstrate the effectiveness of ICL for clinical assessment tasks, particularly when dealing with limited training data in sensitive domains. The system shows strong performance across all tasks, with particular strength in capturing self-state dynamics.

1 Introduction

Mental health disorders affect approximately 970 million people worldwide, with depression and anxiety among the leading causes of disability globally.(World Health Organization, 2022) Social media is one of the many spaces where individuals often share aspects of their psychological well-being, seek support, and sometimes express distress. Given the widespread use of these platforms, they have been studied as potential sources of insight into mental health trends at scale.

CLPsych 2025 focuses on capturing mental health dynamics from social media timelines, viewing human experience as consisting of self-states that fluctuate over time. In this paper, we propose ICL to detect self-states and make the following contributions:

- A cascading framework that models mental health assessment across three progressive levels: evidence identification, post dynamics, and timeline patterns.
- A dual-granularity retrieval system (post-level and timeline-level) showing how optimal retrieval parameters (k=5, k=10) vary by assessment task complexity.

This approach allows us to leverage domain expertise without specific fine-tuning, which is particularly valuable when dealing with limited training data in sensitive domains like mental health.

2 Related Work

Mental health assessments on social media have gained significant attention in recent years. Previous CLPsych shared tasks have explored various aspects of mental health analysis, including longitudinal modeling of mood changes (Tsakalidis et al., 2022) and evidence generation for suicidality risk (Zirikly et al., 2019; Shing et al., 2018). *ICL* has emerged as a powerful technique for leveraging large language models without task-specific fine-tuning (Brown et al., 2020). By providing relevant examples within the prompt, ICL enables models to learn from demonstrations rather than parameter updates. Recent work by Uluslu et al. (2024) has shown the effectiveness of integrating emotional information retrieval with ICL for detecting suicidality risk, achieving top performance in the CLPsych 2024 shared task. *Retrieval-Augmented Generation* approaches and vector databases enhance LLM performance on specialized tasks by retrieving information beyond parametric knowledge (Lewis et al., 2020). This is particularly valuable in clinical domains, where accuracy and evidence-based reasoning are crucial. Similar cascading architectures have been effective in legal judgment prediction (Chalkidis et al., 2022) and medical diagnosis (Wang et al., 2023), refining insights through sequential processing. Framework-guided retrieval

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 268–278
May 3, 2025 ©2025 Association for Computational Linguistics

has also improved educational applications (Liu et al., 2023), where pedagogical principles inform example selection.

3 Task Description

This iteration of CLPsych 2025 analyzes social media timelines to capture mental health dynamics. Each social media timeline consists of chronologically ordered posts by the same individual, with each post potentially containing evidence of adaptive or maladaptive self-states.

Task A: Post-level Judgments Task A consists of two subtasks, where the detailed prompts can be found in Appendix A.1:

*Task A.1:*Identifying evidence of adaptive and maladaptive self-states in posts, which requires extracting spans of text that provide evidence for different types of self-states and is evaluated using recall-oriented BERTScore metrics.

Task A.2: Rating overall well-being on a scale from 1 (low well-being) to 10 (high well-being), which is evaluated using Mean Squared Error (MSE) and F1 Macro score.

Task B: Post-level Summaries Task B involves generating a summary of the interplay between adaptive and maladaptive self-states identified in each post. This requires determining which self-state is dominant and identifying the central organizing aspect (A, B, C, or D) that drives the state. Summaries are evaluated using Mean Consistency and Max Contradiction metrics based on Natural Language Inference models (see prompt in Appendix A.2).

Task C: Timeline-level Summaries Task C requires generating a summary focusing on the interplay between adaptive and maladaptive self-states along the entire timeline (see full prompt in Appendix A.3). This involves emphasizing temporal dynamics, such as flexibility, rigidity, improvement, and deterioration. Evaluation uses the same consistency-based metrics as Task B.

4 System Description

Our system implements a comprehensive approach to all three tasks using ICL with local LLM inference via Ollama. We utilized Llama 3.3 70B as our primary language model, running it locally through Ollama to maintain data privacy and control over the inference process. The system consists of three main components: vector database creation, task-specific processing, and result integration.

System Architecture Figure 1 presents the overall architecture of our system. The architecture consists of five main layers, where the system consists of multiple layers, each serving a distinct function in processing social media data. The *Data Layer* provides the training dataset, comprising social media posts annotated by experts. These posts are then transformed into vector representations through the *Embedding Layer*, which employs Linq-Embed-Mistral (Junseong Kim, 2024) to capture emotional content. The *Vector and ICL Processing Layer* integrates specialized vector databases for posts and timelines, facilitating interactions with Llama 3.3 70B. At the *Tasks Layer*, task-specific modules (A, B, C) generate prompts and process outputs tailored to different analytical objectives. Finally, the *Output Layer* structures and organizes the final outputs for each task, ensuring clarity and usability. The architecture enables experimentation with different k values for ICL, as shown by the parametrized connection between the vector databases and task modules.

Vector Database Foundation We built two specialized vector databases for efficient data organization and retrieval. The post-level database stores individual posts with annotations, evidence spans, well-being scores, and summaries, enabling detailed analysis. The timeline-level database captures broader temporal patterns by storing timeline representations, providing a comprehensive view of psychological trends.

Both databases utilize the Linq-Embed-Mistral embedding model, chosen for its strong performance on the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023). This model effectively captures semantic relationships between posts with similar psychological states. We measured similarity using cosine similarity between normalized embeddings and indexed vectors with HNSW (Malkov and Yashunin, 2018) for fast approximate nearest neighbor search. Instead of a distance threshold, we retrieved a fixed top-k (k=5 or k=10) nearest neighbors per query via ChromaDB, optimizing retrieval speed and quality for real-time ICL operations.

ICL Framework: Our approach follows a structured process across all tasks. First, the input, whether a post or a timeline, is embedded using

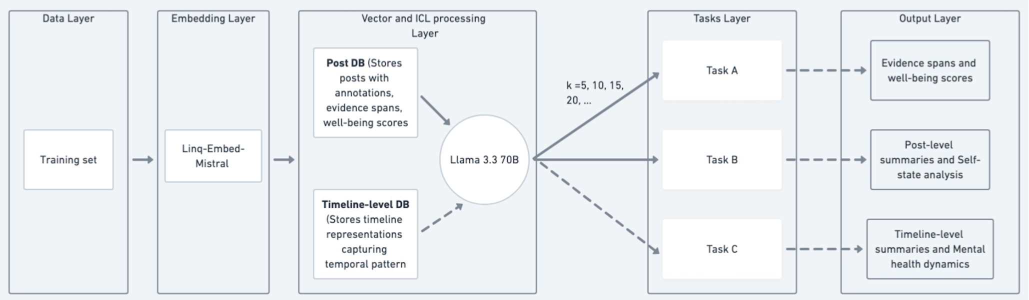
Data Layer
Training set
Embedding Layer
Linq-Embed-Mistral
Vector and ICL processing Layer
Post DB (Stores posts with annotations, evidence spans, well-being scores
Timeline-level DB (Stores timeline representations capturing temporal pattern
Llama 3.3 70B
k =5, 10, 15, 20, ...
Tasks Layer
Task A
Task B
Task C
Output Layer
Evidence spans and well-being scores
Post-level summaries and Self-state analysis
Timeline-level summaries and Mental health dynamics

Figure 1: Overview of System architecture.

Linq-Embed-Mistral. The system then queries the vector database to identify the most similar examples, which are subsequently formatted into a demonstration section. A detailed prompt is then constructed, incorporating task definitions, the ABCD framework, relevant example demonstrations, and the target input. This prompt is sent to Ollama, which runs Llama 3.3 70B for inference. Finally, the model's response is processed to extract the required outputs, ensuring task-specific insights are effectively derived.

The complete prompt templates used for each task are provided in Appendix A.

We experimented with different values of k (the number of examples retrieved for in-context learning), specifically k=5 and k=10, resulting in two separate submissions to the shared task. This allowed us to evaluate the impact of example quantity on model performance.

Task A Implementation: Our system retrieves similar posts from the vector database to serve as examples, guiding the LLM in identifying evidence spans and assessing well-being. The prompt is designed to include detailed definitions of self-states and the ABCD framework (see Appendix A.1). The TaskAWithICL class processes each post by first locating k similar posts in the vector database (k=5 or k=10, depending on the configuration). These examples are then formatted into the prompt before querying Llama 3.3 70B with a structured input. Finally, the system extracts the evidence spans and well-being scores from the LLM's response, ensuring accurate assessment of self-state indicators.

Task B Implementation: Our system builds on Task A's outputs and retrieves posts with high-quality summaries to serve as examples. While Task A retrieval is based solely on post content similarity, Task B employs a more selective approach. It queries the same vector database but applies additional filtering to prioritize examples that have both evidence annotations and existing summaries, ensuring higher quality demonstrations for the summarization task. As detailed in Appendix A.2, the prompt instructs the LLM to determine the dominant self-state, identify the central organizing aspect, and generate a cohesive paragraph summary.

Task C Implementation For Task C, our system generates a structured timeline representation to capture temporal mental health dynamics. We implement the following components:

- **Timeline metrics**: We calculate duration (days between first and last post) and posting frequency (posts per week) using date parsing functions that handle multiple formats
- **Well-being statistics**: We compute average scores, range (min/max), and trend analysis (improving, declining, fluctuating, or stable) using post-level well-being assessments from Task A
- **Self-state pattern analysis**: We identify predominant psychological patterns by counting adaptive, maladaptive, and mixed states across posts, classifying timelines as "Predominantly Adaptive," "Predominantly Maladaptive," "Mixed," or "Balanced"
- **Chronological mapping**: We create a sequence of posts with associated self-states, ordered by date when available, to track psychological evolution

Our system retrieves similar timeline patterns from our timeline-level vector database using cosine similarity between embeddings generated by Linq-Embed-Mistral. We specifically prioritize retrieving examples with high-quality summaries to serve as effective demonstrations. The system then constructs a prompt that instructs the LLM to focus on temporal dynamics such as flexibility, rigidity, improvement, and deterioration (see Appendix A.3). This structured approach helps the model generate coherent summaries that capture the evolution of self-states over time.

Example Output: Task B and C To illustrate the clinical relevance of our system, we provide example outputs in Appendix B. The Task B example demonstrates how our system identifies and summarizes the interplay between adaptive and maladaptive self-states within a single post, organizing the analysis around the dominant affect-driven maladaptive state while acknowledging a secondary adaptive state.

The Task C example effectively captures the temporal dynamics of self-states across a three-month period, highlighting the transition from predominantly maladaptive to increasingly adaptive states. It identifies key transition points (therapy engagement) and describes the specific ABCD elements that change over time (affect, cognition, and behavior). This structured analysis demonstrates the system's ability to synthesize complex psychological patterns across multiple posts, providing clinically relevant insights about a user's mental health trajectory.

Integrated Workflow: Our system implements an integrated workflow in which each task builds upon the previous one. Task A identifies evidence and well-being with ICL guidance. Task B then generates post summaries using Task A's outputs and ICL examples. Finally, Task C synthesizes a timeline analysis by integrating all previous outputs and applying timeline-level ICL. This cascading approach enables the system to conduct increasingly complex psychological assessments without requiring task-specific fine-tuning. Additionally, the prompts for each stage (see Appendix A) progressively increase in complexity and scope, ensuring a structured and scalable assessment process.

Error Mitigation Mechanisms To address the risk of cascading errors in our pipeline, we implemented several safeguards:

- **Quality filtering**: Our implementation ensures high-quality evidence identification through prompt instructions that require "exact text spans from the post, without modifications" and structured JSON output validation
- **Consistency checking**: Our system compares post summaries with identified evidence, ensuring Task B outputs align with Task A findings before generating timeline-level summaries
- **Similarity-based retrieval**: Our vector database retrieves the most semantically relevant examples using the specialized Linq-Embed-Mistral model, enhancing the quality and relevance of in-context examples
- **Format verification**: We implement regex pattern matching to validate structured timeline representations before processing, ensuring consistent input formatting across tasks
- **Exception handling**: Robust try-except blocks throughout the implementation prevent crashes when encountering unexpected data formats, providing graceful degradation

These mechanisms help reduce error amplification through the pipeline, though a human-in-the-loop validation would further enhance reliability in clinical applications. Future implementations could incorporate clinician feedback at key decision points.

Computational Considerations Running Llama 3.3 70B locally via Ollama requires substantial computational resources. Our implementation includes several practical considerations to balance performance and accessibility:

- **Model flexibility**: Our system architecture allows specifying different Ollama models through command-line arguments (as seen in our '–model' parameter), enabling users to select models based on their available hardware
- **Controlled batch processing**: We implement timed delays between processing files and posts (using 'time.sleep()') to prevent system overload, with configurable pause durations
- **Progressive task structure**: Our cascading pipeline allows running different components independently (Tasks A, B, and C), enabling incremental processing on systems with limited resources
- **Efficient vector retrieval**: We leverage ChromaDB's HNSW indexing for similarity search operations, making example retrieval faster and more resource-efficient

These design choices allow for deployment across various hardware configurations, though users should consider the performance trade-offs when using smaller models for complex timeline analysis tasks.

5 Results

Table 1 presents our submitted team results (EAIonFlux) compared to other participating systems. We submitted two different configurations—one with k=5 and another with k=10 for the number of examples retrieved during in-context learning—to evaluate the impact of example quantity on performance. Interestingly, our results showed that the configuration with k=10 generally outperformed k=5 for Tasks A.1 and B, suggesting that more examples provide better guidance for these complex tasks. However, for Tasks A.2 and C, the difference was less pronounced, indicating that well-being assessment and timeline-level summarization may be less sensitive to the number of examples.

Our system performed particularly well on Tasks B and C, demonstrating the effectiveness of our ICL approach with Llama 3.3 70B in generating coherent and clinically meaningful summaries. The relatively small difference in mean consistency metrics compared to other systems suggests that our approach effectively captures mental health dynamics at both the post and timeline levels.

6 Discussion

Our results demonstrate ICL's effectiveness for clinical assessment tasks with limited training data. The proposed system leveraged vector similarity retrieval using Linq-Embed-Mistral embeddings to identify semantically similar examples that reflected psychological patterns, which was crucial for nuanced mental health assessment. Model capabilities were enhanced through Llama 3.3 70B, which demonstrated strong reasoning abilities for complex psychological concepts, enabling the generation of clinically meaningful outputs. Example optimization experiments with k=5 and k=10 showed that incorporating more examples improved performance in Tasks A.1 and B, auiding the model's comprehension of intricate self-state patterns. To enhance clinical knowledge integration, prompts were structured using the ABCD framework (detailed in Appendix A.1), guiding the model toward more accurate psychological assessments. The system followed a cascading architecture, mirroring clinical workflows by allowing tasks to build upon previous insights without requiring task-specific fine-tuning. Lastly, privacy-preserving inference was ensured through local deployment via Ollama, maintaining data privacy while upholding performance quality.

7 Conclusion

In this paper, we presented our approach to the CLPsych 2025 shared task, which focuses on capturing mental health dynamics from social media timelines. Our system implements In-Context Learning with vector similarity to retrieve relevant examples that guide Llama 3.3 70B without specific fine-tuning. The results demonstrate the effectiveness of our approach for clinical assessment tasks, particularly when dealing with limited training data in sensitive domains. Our system shows strong performance across all tasks, with particular strength in capturing self-state dynamics at both the post and timeline levels.

Future work could explore several promising directions: developing specialized psychological embeddings to improve on our current Linq-Embed-Mistral implementation; implementing diversity-aware example selection strategies beyond simple vector similarity; integrating explainability features that highlight influential text spans; incorporating human-in-the-loop validation for error prevention; conducting comprehensive fairness evaluations across demographic groups; and extending to multimodal analysis for more holistic assessment. These enhancements would improve both technical performance and clinical utility, moving toward more equitable, transparent tools for mental healthcare support.

Ethical Considerations and Limitations

This work raises a number of important ethical considerations. All data used in this study was provided as part of the CLPsych 2025 shared task and has been properly de-identified to protect user privacy. No additional data collection was performed. While our approach prioritizes privacy and security by running models locally through Ollama rather than sending sensitive data to external API services, we acknowledge that automated mental health assessment tools should only be used as supportive aids and not as replacements for professional clinical judgment. Additionally, we want to emphasize that any practical deployment would require ex-

Task	Metric	EAIonFlux_1 (k=5)	Delta	EAIonFlux_2 (k=10)	Delta	Best System (Score)
Task A.1	Recall$^{\uparrow}$	0.498	0.139	0.517	0.120	uOttawa (0.637)
	Weighted Recall$^{\uparrow}$	0.480	0.018	0.471	0.027	uOttawa (0.498)
Task A.2	MSE$^{\downarrow}$	2.08	0.16	2.87	0.95	BULUSI (1.920)
	F1 Macro$^{\uparrow}$	0.321	0.072	0.320	0.073	BLUE (0.393)
Task B	Mean Consistency$^{\uparrow}$	0.884	0.026	0.888	0.022	BLUE (0.910)
	Max Contradiction$^{\downarrow}$	0.780	0.247	0.782	0.249	BLUE (0.533)
Task C	Mean Consistency$^{\uparrow}$	0.906	0.040	0.913	0.033	BLUE (0.946)
	Max Contradiction$^{\downarrow}$	0.774	0.420	0.760	0.406	PsyMetric (0.354)

$^{\uparrow}$ Higher values are better. $^{\downarrow}$ Lower values are better.
Delta shows the absolute difference between our system and the best system for each metric.

Table 1: Results of EAIonFlux submissions compared to the best-performing systems in the CLPsych 2025.

tensive clinical validation, careful consideration of bias, and appropriate safeguards to prevent misuse and comply with regulatory standards. We also recognize that computational models of mental health states may reflect biases present in training data. While our in-context learning approach aims to mitigate some biases by explicitly incorporating clinical frameworks, more work is needed to ensure fair and equitable performance across diverse populations.

While our system demonstrated strong performance across all tasks, several limitations should be noted:

Dependency on Example Quality The effectiveness of our ICL approach depends heavily on the quality and representativeness of the examples in the vector database. Our implementation prioritizes examples with human-verified summaries when available, as seen in the timeline similarity retrieval method in Task C, but future versions should incorporate more sophisticated filtering to eliminate potentially misleading examples.

Computational Requirements The use of vector databases and running Llama 3.3 70B locally requires substantial computational resources, which could limit accessibility. Our code includes configurable parameters for model selection and batch processing delays, but the core implementation still requires high-end hardware for optimal performance, potentially creating barriers to adoption in resource-constrained environments.

Limited Clinical Validation While our system was evaluated against expert annotations, broader clinical validation would be necessary before any real-world deployment. The shared task evaluation metrics may not fully capture all aspects of clinical utility, and real-world application would require additional validation studies with mental health professionals.

Potential for Hallucination LLMs can sometimes generate plausible-sounding but incorrect information, which is particularly concerning in clinical contexts. Although our prompts explicitly instruct the model to "Include only EXACT text spans from the post, without any modifications," we observed that the model sometimes struggled with adhering to this constraint. To address these issues, our implementation includes:

- Structured JSON response formats that constrain the model's outputs
- JSON response cleaning methods that validate and sanitize model outputs
- Explicit instructions in prompts to reference only content present in the input text
- Post-processing that validates evidence spans against original post content

Despite these measures, hallucination remains a challenge requiring ongoing research and potential integration of human oversight in critical applications.

Cultural and Demographic Biases The system may inherit biases present in the training data of the underlying LLMs, which could affect its performance across different demographic groups. Mental health expressions vary across cultures, and our current approach does not explicitly account for these differences. For example, the ABCD framework may not adequately capture culturally-specific expressions of psychological distress that fall outside Western clinical paradigms. Our vector database implementation does not include specific mechanisms to ensure diverse representation across cultural contexts.

Cascading Error Propagation and Lack of Human Oversight Our cascading architecture, while efficient, creates the potential for error propagation through the pipeline and subsequently severe ethical risks. Our code analysis revealed that errors in evidence identification from Task A directly affect the input to Tasks B and C, as seen in the data flow between the task implementations. While our implementation includes exception handling and validation steps, it lacks explicit mechanisms for detecting or correcting propagated errors. Future work should explore incorporating human-in-the-loop validation checkpoints between stages to prevent error cascades and to provide corrective feedback that could further improve the system's accuracy and reliability.

A Task-Specific Prompts

This appendix contains the core prompt templates used for each task in our system. These prompts were dynamically combined with retrieved examples during in-context learning. The system message instructing the model to act as "an expert in clinical psychology analyzing social media posts" was consistent across all tasks.

A.1 Task A Prompt: Post-level Evidence Identification and Well-being Assessment

The following is the complete Task A prompt implementation with in-context learning examples as used in our code:

> **System Message:** You are an expert in clinical psychology analyzing social media posts.
>
> **User Message:**
> You are analyzing social media posts for the CLPsych 2025 shared task. Your task is to:
>
> 1. Identify evidence of adaptive and maladaptive self-states in the post. 2. Rate the overall well-being presented in the post on a scale from 1 (low) to 10 (high).
>
> ## Definitions of Self-States
> Self-states constitute identifiable units characterized by specific combinations of Affect, Behavior, Cognition, and Desire/Need (ABCD) that tend to be coactivated in a meaningful manner for limited periods of time.
>
> - An adaptive self-state pertains to aspects of Affect, Behavior, and Cognition towards the self or others, which is conducive to the fulfillment of basic desires/needs (D), such as relatedness, autonomy and competence.
>
> - A maladaptive self-state pertains to aspects of Affect, Behavior, and Cognition towards the self or others, that hinder the fulfillment of basic desires/needs (D).
>
> ## ABCD Elements with Examples
>
> ### Affect: Type of emotion expressed by a person
> - Adaptive Examples: Calm/Laid back, Emotional Pain/Grieving, Content/Happy, Vigor/Energetic, Justifiable Anger/Assertive Anger, Proud. - Maladaptive Examples: 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 (B-O): The person's main behavior(s) toward the other
> - Adaptive Examples: Relating behavior, Autonomous behavior - Maladaptive Examples: Fight or flight behavior, Overcontrolled/controlling behavior
>
> #### Behavior toward the Self (B-S): The person's main behavior(s) toward the self
> - Adaptive Examples: Self-care behavior - Maladaptive Examples: Self-harm/Neglect/Avoidance behavior
>
> ### Cognition
>
> #### Cognition of the Other (C-O): The person's main perceptions of the other
> - Adaptive Examples: Perception of the other as related, Perception of the other as facilitating autonomy needs - Maladaptive Examples: Perception of the other as detached or over attached, Perception of the other as blocking autonomy needs
>
> #### Cognition of the Self (C-S): The person's main self-perceptions
> - Adaptive Examples: Self-acceptance and self-compassion - Maladaptive Examples: Self-criticism
>
> ### Desire: The person's main desire, need, intention, fear or expectation
> - Adaptive Examples: Relatedness, Autonomy and adaptive control, Competence, Self-esteem, Self-care - Maladaptive Examples: Expectation that relatedness need will not be met, Expectation that autonomy needs will not be met, Expectation that competence needs will not be met
>
> ## Well-being Scale (1-10)
>
> - 10: No symptoms and superior functioning in a wide range of activities - 9: Absent or minimal symptoms (e.g., 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 (e.g., difficulty concentrating after family argument). Slight impairment in social, occupational or school functioning. - 7: Mild symptoms (e.g., 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 (e.g., 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 (e.g., 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 (e.g., no job, home, or friends). - 2: In danger of hurting self or others (e.g., 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.

The clinical cutoff score is 6, meaning that individuals scoring below 6 may be experiencing significant distress.

## Similar Examples for Reference
Here are some examples of similar posts with their annotations:

Example 1:

Post: "{example_post_1}"

Annotation: { "adaptive_evidence": { "A": { "highlighted_evidence": "{adaptive_evidence_span}", "Category": "{adaptive_category}" }, ... }, "maladaptive_evidence": {...}, "well_being_score": {score} }

Example 2:

Post: "{example_post_2}"

Annotation: { ... }

[ADDITIONAL EXAMPLES UP TO k=5 OR k=10]

Please analyze the target post following a similar approach to these examples, but make your own assessment based on the specific content.

## Post to Analyze

Here is the post to analyze: "{post_content}"

## Response Format

Respond in JSON format with the following structure: { "adaptive_evidence": { // Include only the categories where evidence is found "A": { "highlighted_evidence": "exact text span", "Category": "Specific affect category (e.g., 'Content/Happy')" }, // Other categories as needed (B-O, B-S, C-O, C-S, D) }, "maladaptive_evidence": { // Same structure as adaptive_evidence }, "well_being_score": integer from 1-10, "reasoning": "brief explanation of your assessment" }

Important: 1. Include only EXACT text spans from the post, without any modifications. 2. Only include categories where you found clear evidence. 3. Be specific about the subcategory (e.g., "Content/Happy" not just "Affect"). 4. Make sure your well-being score aligns with the detailed scale provided. 5. If you find no clear evidence of any self-states, return empty objects for the evidence. 6. Your response should be ONLY the JSON. No other text before or after.

At runtime, the system retrieves k similar posts from our vector database using the Linq-Embed-Mistral embeddings and dynamically formats them as examples using the pattern shown above.

A.2 Task B Prompt: Post-level Summary of Self-state Dynamics

The following shows how the Task B prompt is augmented with in-context learning examples:

System Message: You are an expert in clinical psychology analyzing social media posts.

User Message:
You are analyzing a social media post for the CLPsych 2025 shared task, focusing on Task B - Post-level summary of self-state's inner dynamics.

Your task is to generate a summary of the interplay between adaptive and maladaptive self-states identified in the post. You need to:

1. Determine which self-state is dominant (adaptive/maladaptive) and describe it first. 2. For each self-state, identify the central organizing aspect (A, B, C, or D) that drives the state. 3. Structure the summary around this central aspect, describing how it influences the rest. 4. Emphasize potential causal relationships between the aspects. 5. Then, proceed to the second self-state and follow the same approach. 6. If the post contains only one self-state, summarize only that state.

## Self-State Definitions
Self-states constitute identifiable units characterized by specific combinations of Affect, Behavior, Cognition, and Desire/Need (ABCD) that tend to be coactivated in a meaningful manner for limited periods of time.

- An adaptive self-state pertains to aspects of Affect, Behavior, and Cognition towards the self or others, which is conducive to the fulfillment of basic desires/needs (D). - A maladaptive self-state pertains to aspects of Affect, Behavior, and Cognition towards the self or others, that hinder the fulfillment of basic desires/needs (D).

## Similar Examples for Reference
Here are some examples of similar posts with their evidence and summaries:

Example 1:

Post: "{example_post_1}"

Evidence:
- Adaptive:
A: "{adaptive_evidence_span}" ({adaptive_category})
...

- Maladaptive:
...

Summary:
"{example_summary_1}"

Example 2:

Post: “{example_post_2}”

Evidence:
...

Summary:
“{example_summary_2}”

[ADDITIONAL EXAMPLES UP TO k=5 OR k=10]

Please analyze the target post following a similar approach to these examples, but make your own assessment based on the specific content.

## Post to Analyze
Here is the post: “{target_post}”

## Evidence Identified in the Post
Adaptive evidence:
{formatted_adaptive_evidence}

Maladaptive evidence:
{formatted_maladaptive_evidence}

## Response Instructions
Write a cohesive paragraph summary (200-300 words) that:

1. First describes the dominant self-state (whichever has more significant evidence). 2. Identifies which ABCD aspect (Affect, Behavior-Self, Behavior-Other, Cognition-Self, Cognition-Other, or Desire/Expectation) is central to each self-state. 3. Explains how this central aspect influences other aspects, focusing on causal relationships. 4. Naturally integrates ABCD elements into the description without explicitly highlighting them. 5. Uses clinical language appropriate for psychological assessment.

Do not use bulleted lists or headers in your summary. Write in a fluid, paragraph style.

At runtime, our system retrieves k similar posts through vector similarity, prioritizing examples that have both evidence annotations and existing high-quality summaries. This selective filtering ensures that the examples provided to the model demonstrate appropriate summary creation. Unlike Task A, which only requires evidence identification, Task B examples must showcase how evidence is integrated into coherent summaries that identify central organizing aspects and causal relationships.

A.3 Task C Prompt: Timeline-level Summary of Self-state Dynamics

The following shows how the Task C prompt is augmented with in-context learning examples, specifically using timeline-level representations:

System Message: You are an expert in clinical psychology analyzing social media posts.

User Message:
You are analyzing a social media timeline for the CLPsych 2025 shared task, focusing on Task C - Timeline-level summary of self-state’s dynamics.

Your task is to generate a summary focusing on the interplay between adaptive and maladaptive self-states along the timeline. You need to:

1. Emphasize temporal dynamics focusing on concepts such as flexibility, rigidity, improvement, and deterioration. 2. Describe the extent to which the dominance of the self-states changes over time. 3. Explain how changes in aspects (Affect, Behavior, Cognition, and Desire) contribute to these transitions.

## Timeline to Analyze

This timeline contains {len(posts_with_evidence)} posts spanning from {posts_with_evidence[0]["date"]} to {posts_with_evidence[-1]["date"]}.

Here are the posts with their Well-being scores and evidence:

{chronological_post_listing}

## Post-level Summaries (if available)

{post_level_summaries}

## Similar Timelines for Reference
Here are some examples of similar timelines with their summaries:

Example 1 (Timeline ID: {timeline_id_1}):

Timeline Characteristics:
ID: {timeline_id_1}
Time span: {duration_text}
Post count: {post_count}
Average well-being: {avg_well_being}
Well-being range: {min_well_being} to {max_well_being}
Well-being trend: {trend}
Self-state pattern: {state_dynamics}

Timeline Summary:
“{example_summary_1}”

Example 2 (Timeline ID: {timeline_id_2}):

Timeline Characteristics:
...

Timeline Summary:
“{example_summary_2}”

[ADDITIONAL EXAMPLES UP TO k=5 OR k=10]

Please analyze the target timeline following a similar approach to these examples, but make your own assessment based on the specific content.

## Response Instructions

Write a cohesive paragraph summary (200-400 words) that:

1. Describes the overall pattern of self-states across the timeline (e.g., predominantly adaptive, predominantly maladaptive, fluctuating). 2. Identifies any shifts or transitions between dominance of adaptive and maladaptive states. 3. Analyzes temporal dynamics using concepts like flexibility, rigidity, improvement, and deterioration. 4. Explains how changes in ABCD aspects (Affect, Behavior, Cognition, Desire) contribute to these transitions. 5. Uses clinical language appropriate for psychological assessment.

Emphasize how the individual’s self-states evolve over time. Do not use bulleted lists or headers in your summary. Write in a fluid, paragraph style.

For Task C, our system uses a specialized timeline-level vector database different from the post-level database used in Tasks A and B. The system creates a comprehensive timeline representation that includes:

- Chronological timeline metadata (duration, post frequency, date range)
- Well-being statistics (average scores, range, trends)
- Self-state pattern analysis (predominantly adaptive, maladaptive, mixed, or balanced)
- Sequence of individual posts with their dominant self-states

This timeline representation is embedded using Linq-Embed-Mistral and used to retrieve similar timelines for in-context learning. The system specifically filters for examples that have existing high-quality timeline summaries, providing effective demonstrations of temporal analysis. Task C builds upon the outputs of Tasks A and B, using the evidence spans, well-being scores, and post-level summaries as inputs for the timeline-level analysis.

B Example System Outputs

B.1 Task B Example Output

The following is an example of a post-level summary generated by our system for Task B:

> "The dominant self-state is maladaptive, primarily organized around feelings of anxiety and tension (Affect). The individual expresses significant worry about an upcoming presentation, manifested through rumination and catastrophic thinking. This anxiety drives avoidant behavior, as they mention repeatedly postponing preparation and distracting themselves with less important tasks. Their cognition reveals self-criticism, particularly regarding their perceived incompetence and inability to handle pressure. Despite these maladaptive elements, there is a minor adaptive self-state present, characterized by moments of self-awareness when they acknowledge the need to address their procrastination and apply coping strategies they've learned previously."

B.2 Task C Example Output

The following is an example of a timeline-level summary generated by our system for Task C:

> "This timeline demonstrates a fluctuating pattern between maladaptive and adaptive self-states over a three-month period, with a gradual improvement trend. Initially dominated by maladaptive states characterized by anxiety and depressive affect, the individual's posts show rigid negative cognitions about themselves and their future. Around the midpoint, a significant transition occurs following engagement with therapy, where adaptive self-states begin emerging with greater frequency. This shift is marked by increasing cognitive flexibility, with the individual demonstrating greater self-compassion and employing healthier coping behaviors. While maladaptive states still recur during stressful periods, they become less intense and persistent. The timeline reveals a dynamic interplay between affect and cognition as primary drivers of state transitions, with improvements in cognitive patterns (reduced self-criticism, increased perspective-taking) typically preceding positive affect changes."

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2022. An empirical study on neural methods for legal judgment prediction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4783–4798. Association for Computational Linguistics.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Jiawei Liu, Zhiwei Tu, William Yang Wang, Dongkuan Zhang, Yiquan Cui, and Philip S. Yu. 2023. Retrieval-augmented generation for knowledge-intensive nlp tasks: A survey. In *Proceedings of the 2023 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 10616–10649. Association for Computational Linguistics.

Yury A. Malkov and Dmitry A. Yashunin. 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.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316.

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, New Orleans, LA. Association for Computational Linguistics.

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

Talia Tseriotou, Jenny Chim, Ayal Klein, Aya Shamir, Guy Dvir, Iqra Ali, Cian Kennedy, Guneet Singh Kohli, Anthony Hills, Ayah Zirikly, Dana Atzil-Slonim, and Maria Liakata. 2025. Overview of the clpsych 2025 shared task: Capturing mental health dynamics from social media timelines. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

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

Jingqing Wang, Rupert He, Amos Koker, Zhihong Ren, and Percy Liang. 2023. A survey on retrieval-augmented text generation. *ACM Computing Surveys*, 55(12):1–36.

World Health Organization. 2022. Mental disorders. *World Health Organization Fact Sheets*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

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*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.