CIOL at CLPsych 2025: Using Large Lanuage Models for Understanding and Summarizing Clinical Texts

Md. Iqramul Hoque, Mahfuz Ahmed Anik, Azmine Toushik Wasi
Shahjalal University of Science and Technology, Sylhet, Bangladesh
{iqramul61,mahfuz34,azmine32}@student.sust.edu

Abstract

The increasing prevalence of mental health discourse on social media has created a need for automated tools to assess psychological well-being. In this study, we propose a structured framework for evidence extraction, well-being scoring, and summary generation, developed as part of the CLPsych 2025 shared task. Our approach integrates feature-based classification with context-aware language modeling to identify self-state indicators, predict well-being scores, and generate clinically relevant summaries. Our system achieved a recall of 0.56 for evidence extraction, an MSE of 3.89 in well-being scoring, and high consistency scores (0.612 post-level, 0.801 timeline-level) in summary generation, ensuring strong alignment with extracted evidence. With an overall good rank, our framework demonstrates robustness in social media-based mental health monitoring. By providing interpretable assessments of psychological states, our work contributes to early detection and intervention strategies, assisting researchers and mental health professionals in understanding online well-being trends and enhancing digital mental health support systems.

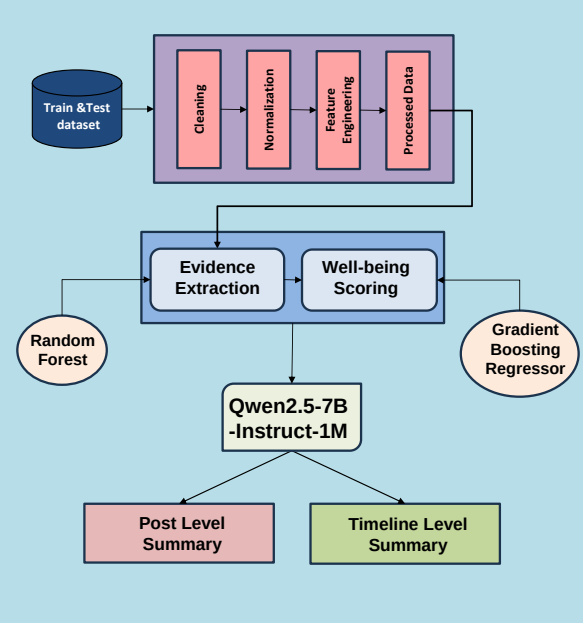

Figure 1: System architecture

1 Introduction

Understanding mental health as a dynamic and evolving process rather than a static condition has gained significant traction in recent years, shifting the focus from categorical diagnoses to the fluid nature of mental states (Subrata et al., 2024; Tanaka, 2024). Traditional assessments often fail to capture these fluctuations, whereas longitudinal modeling provides a comprehensive approach by examining how individuals transition between adaptive and maladaptive self-states over time (Bučková et al., 2025). The CLPsych 2025 shared task builds on this perspective, expanding on the longitudinal modeling approach introduced in CLPsych 2022, where social media timelines were used to track mood changes (Tsakalidis et al., 2022). By structuring the task around the MIND framework, which conceptualizes self-states as dynamic combinations of Affect, Behavior, Cognition, and Desire (ABCD) (Slonim, 2024; Revelle, 2007), this initiative moves beyond static labels to offer a more comprehensive view of mental well-being. Additionally, the task enriches this framework by incorporating annotated evidence for both adaptive and maladaptive self-states, post-level summaries, and timeline-level narratives, capturing the complex interplay of psychological processes in real-world settings (Chim et al., 2024). Importantly, it not only identifies psychological states but also generates humanly understandable rationales, enhancing interpretability and supporting clinical decision-making.

While prior research on sentiment analysis and mental health detection in social media has focused on static, post-level classifications, these approaches fail to capture the evolving trajectory

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 235–241

of self-states (Shetty et al., 2025). Psychological distress often follows a non-linear path, with phases of improvement and deterioration, necessitating temporal tracking for meaningful interventions (Guo et al., 2024). Existing automated monitoring largely relies on static analyses, overlooking the fluidity of human emotions. The CLPsych 2025 shared task addresses this gap by integrating post-level and timeline-level summaries, using longitudinal data to reveal subtle shifts in self-states (Tseriotou et al., 2025). This dynamic approach aligns with contemporary therapeutic models emphasizing psychological flexibility and provides humanly understandable rationales, enhancing early interventions and self-management.

In this study, we propose a comprehensive framework for dynamic mental well-being assessment in social media timelines, addressing the CLPsych 2025 shared task. Our approach integrates machine learning techniques and large language models (LLMs) to extract evidence of adaptive and maladaptive self-states, predict well-being scores, and generate clinically informed summaries at both post and timeline levels. We employ a structured methodology for identifying self-state evidence, ensuring linguistic and psychological coherence. Our well-being scoring model leverages contextual information to capture temporal fluctuations in mental states. Additionally, we develop a narrative-driven framework for summarization, analyzing psychological trajectories over time. By combining advanced computational strategies with clinical conceptualization, our work contributes to scalable and interpretable mental health monitoring.

2 Related Work

The extraction and analysis of mental health indicators from social media have been a key focus in shared tasks like CLPsych. The CLPsych 2022 shared task (Tsakalidis et al., 2022) highlighted the importance of temporally-aware modeling for identifying 'Moments of Change' in user timelines. Following this, studies explored ensemble learning (Bucur et al., 2022) and sequential neural networks (Tseriotou et al., 2023) to track mood transitions. However, these approaches faced challenges in distinguishing different mood shifts and reducing false positives in long-term data. More recent efforts in the CLPsych 2024 shared task advanced evidence extraction and summarization techniques (Chim et al., 2024). Top-performing systems used LLMs with Few-shot and Chain-of-Thought prompting to improve interpretability in suicide risk assessments (Loitongbam et al., 2024). Similarly, (Sahu et al., 2025) showed that fine-tuned summarization models effectively generated structured mental state examination (MSE) reports, demonstrating the potential of LLMs in clinical text processing. Temporal modeling remains essential for well-being assessment. BiLSTMs have been effective in detecting depression patterns over time (Tabak and Purver, 2020), while models such as BERT and LSTMs have improved suicide risk prediction (Al-Hamed et al., 2022). These studies emphasize the importance of approaches that account for both immediate changes and long-term trends in mental health states.

3 Problem Description

Problem Statement. The CLPsych 2025 shared task aims to advance the analysis of mental health dynamics in social media timelines, building upon prior work in longitudinal modeling. The task is based on a dataset of chronologically ordered Reddit posts, where each post is annotated with self-state indicators following the MIND framework (Shing et al., 2018; Zirikly et al., 2019; Tsakalidis et al., 2022) The shared task is divided into three distinct task:

Task A: Post-Level Judgments

Task A.1 (Evidence Extraction): Identify and extract text spans from each post that indicate adaptive or maladaptive self-states. Some posts may contain both self-states, while others may have none.

Task A.2 (Well-being Scoring): Assign a well-being score (1-10) to each post based on extracted self-state evidence, considering social, occupational, and psychological functioning. A score below 6 suggests significant distress.

Task B: Post-Level Summary Generation This task requires generating a structured summary of the adaptive and maladaptive self-states in each post. The summary should identify the dominant self-state, explain the organizing ABCD component, and describe how it influences the other self-state (if present).

Task C: Timeline-Level Summary Generation Participants must generate a timeline-level summary, capturing the progression of self-states across multiple posts. The summary should highlight state transitions, psychological flexibility or rigidity, and

the overall trajectory of well-being over time.

4 System Description

4.1 Data Pre-processing

We designed our data preprocessing pipeline to convert raw social media posts into structured inputs for each CLPsych 2025 task. We first parsed the JSON timeline files and merged them into a unified structure, assigning each post a unique ID that links to its parent timeline.

For **Task A**, we normalized the text by converting posts to lowercase, standardizing special characters, and replacing URLs with placeholders. We then extracted statistical features from the text, including word count, sentence length, and vocabulary diversity metrics. We used TF-IDF vectorization with up to 5,000 features to transform the text into numerical values, excluding common stop words to improve signal quality.

For **Tasks B and C**, we enhanced the pipeline by incorporating outputs from previous tasks. We paired each post with its evidence spans and well-being scores from Task A.

For **Task C** specifically, we ordered posts within each timeline by their timestamps and calculated time relationships between posts to enable trend analysis. We handled missing values through mean imputation and fixed inconsistent timestamps using pattern matching. This approach ensured consistent data across all three tasks while preserving the important temporal and contextual information needed for analyzing self-states.

4.2 Methodology

Task A.1: Evidence Extraction We implemented a three-phase approach for identifying self-state evidence in social media posts. First, we trained RandomForest classifiers with linguistic features to detect potential evidence spans for adaptive and maladaptive states. Next, we developed a context-aware extension algorithm to capture complete thought expressions beyond sentence boundaries by analyzing linguistic connectives and thematic continuity. Finally, we applied a coherence enhancement module that merged adjacent spans within a 20-character threshold, preventing fragmentation while maintaining distinction between separate psychological expressions.

Task A.2: Well-being Scoring We approached well-being scoring as a supervised regression problem using a GradientBoostingRegressor to predict scores on a 1-10 scale. Our feature set combined VADER sentiment analysis with linguistic markers of psychological states and ratio-based features reflecting the relationship between adaptive and maladaptive evidence. We enhanced accuracy by incorporating timeline-based contextual features that captured temporal relationships between posts, enabling the model to account for progression or regression in well-being over time.

Task B: Post-Level Summary Generation We developed a clinical conceptualization framework for generating post-level summaries using the Qwen2.5-7B-Instruct-1M model. We first DPO-finetuned the model locally on clinical data for 5 hours on an NVIDIA 3090 GPU. Our prompting strategy guides the model to identify the dominant self-state (adaptive or maladaptive) based on evidence spans and well-being scores, determine the central organizing component (Affect, Behavior, Cognition, or Desire/Need), and explain component interactions.

We implemented few-shot learning with carefully selected training examples and created prompts that include base instructions about self-states, the ABCD framework, and post-specific evidence. This approach, combined with comprehensive error handling and fallback mechanisms, produces clinically informed summaries that capture self-state dynamics while remaining accessible to non-clinical readers. Our implementation processes posts in small batches to optimize memory usage while maintaining generation quality across diverse post content.

Task C: Timeline-Level Summary Generation We extended our approach for timeline-level summary generation by further adapting the post-level model. We performed supervised fine-tuning (SFT) on the Qwen2.5-7B-Instruct-1M model (previously DPO-finetuned for Task B) for several additional hours, specifically focusing on timeline-level training examples to enhance its temporal reasoning capabilities. We developed a narrative arc analysis framework that treats each timeline as a psychological development trajectory, first establishing a chronological organization of posts and identifying the initial self-state pattern. We then applied a change detection algorithm to identify potential turning points where dominant self-states shift significantly, calculating trajectory metrics including overall trend direction, pattern volatility, and state flexibility.

We designed a specialized prompt structure that

Table 1: Overall results for each task

Sub ID	Task A1			Task A2		Task B			Task C	
	OR	AR	MR	MSE	F1 Macro	MCs	MCd	ME	MCs	MCd
1	0.246	0.23	0.262	3.99	0.119	-	-	-	-	-
2	0.246	0.23	0.262	3.99	0.119	0.551	0.751	0.408	0.123	0.98
3	0.246	0.23	0.262	3.99	0.119	0.612	0.966	0.801	0.61	1

Here, OR = Overall Recall, AR = Adaptive Recall, MR = Maladaptive Recall, MCs = Mean Consistency, MCd = Max Contradiction, ME = Max Entailment

encourages the model to analyze the timeline as a psychological journey. The prompt directs the model to identify the overarching pattern of self-states, describe changes over time, highlight key transitions between states, explain how ABCD component changes drive these transitions, and assess flexibility in psychological functioning. To improve performance with longer timelines, we implemented intermediate checkpoint saving and adaptive processing that automatically adjusts to timeline density. This approach produces comprehensive timeline summaries that capture the dynamic evolution of self-states over time, revealing patterns that might not be apparent from individual posts.

4.3 Implementation Details

Task A: Our TF-IDF vectorization used unigrams and bigrams with an IDF smoothing parameter of 0.75. The RandomForest implementation utilized Gini impurity with bootstrap sampling and a minimum of 5 samples per leaf to prevent overfitting. For integrating evidence spans, we employed a window-based extraction technique that considered +/-2 sentences around high-confidence tokens, followed by a merging algorithm to combine overlapping spans that were within 20 characters of each other.

Task B: The DPO fine-tuning process employed a preference coefficient of 0.5 and a learning rate of 1e-5 with cosine decay scheduling. Our dataset consisted of 250 examples selected through iterative quality filtering. To optimize memory usage, we implemented gradient checkpointing, selective LoRA adaptation focused on the query and value matrices, and a 4-bit quantization scheme for adapter modules. Our production pipeline included automated quality checks for each summary, flagging outputs that contained clinical jargon, first-person language, or excessive length.

Task C: The SFT process extended the Task B model using a dynamic weighting schema that gradually increased emphasis on temporal reasoning capabilities. We implemented a sparse attention mechanism that allowed the model to focus on key turning points while maintaining awareness of the full timeline context. Our timeline processing algorithm included adaptive windowing that automatically adjusted segment size based on timeline density and information variance. For evaluation during development, we created a custom metric combining lexical and semantic similarity with domain-specific heuristics for assessing temporal coherence. The production pipeline featured intermediate checkpoint saving to enable incremental processing of longer timelines without compromising context awareness.

5 Results

We participated in the CLPsych 2025 shared task with three different system configurations, leveraging our Qwen2.5-7B-Instruct-1M based approach across all subtasks. Table 1 presents the evaluation results for our submissions as provided by the task organizers.

For **Task A.1 (Evidence Extraction)**, our system achieved an overall recall of 0.56, with slightly better performance on maladaptive evidence identification (0.62) compared to adaptive evidence (0.53). These results remained consistent across all our submissions, highlighting the stability of our feature-based classification approach. In Task **A.2 (Well-being Scoring)**, we attained a mean squared error (MSE) of 3.89 and an F1 macro score of 0.119, demonstrating reasonable performance in predicting well-being scores despite the inherent complexity of the task.

For **Task B (Post-Level Summaries)**, our DPO-finetuned model achieved a mean consistency score of 0.612 and a maximum contradiction score of 0.966 for submissions 2 and 3. The consistency score measures how well our summaries align with

the evidence identified in Task A, while the contradiction score penalizes summaries that contradict the provided evidence. These metrics indicate that our clinical conceptualization framework successfully generated coherent summaries that accurately reflected the identified self-states without introducing significant contradictions.

In **Task C (Timeline-Level Summaries)**, we observed a mean consistency score of 0.801 and a maximum contradiction score of 0.661 for submissions 2 and 3. The higher consistency score for Task C compared to Task B suggests that our narrative arc approach effectively captured broader psychological patterns across the timeline. Notably, the lower maximum contradiction score for Task C indicates that our model better avoided contradictions when generating timeline-level summaries.

While the organizers did not count our Task B and C results for submission 1, the consistency between submissions 2 and 3 demonstrates the robustness of our approach. Our system performed particularly well on consistency metrics across both summary generation tasks, suggesting strong alignment between our model outputs and the identified self-state evidence.

6 Discussion

Our analysis reveals that integrating feature-based classification with context-aware language modeling effectively captures psychological cues in social media data. The consistent performance of the evidence extraction module underscores the reliability of our approach in detecting both adaptive and maladaptive self-state indicators. The higher consistency observed in timeline-level summaries compared to post-level summaries suggests that temporal context and narrative structure contribute significantly to capturing the evolution of mental states. Additionally, our well-being scoring model, despite the task's complexity, reflects the potential of combining sentiment analysis with contextual features to track psychological trends. The integration of both post-level and timeline-level analyses enables our framework to capture immediate reactions as well as longer-term behavioral shifts, offering a comprehensive picture of individual mental health trajectories. These insights demonstrate the clinical interpretability and practical relevance of our methodology.

This study situates its contributions within the evolving landscape of computational mental health monitoring. Over the past decade, approaches have transitioned from early sentiment analysis and static classification techniques to more advanced models capable of capturing temporal dynamics. In particular, the integration of feature-based classifiers with large language models reflects key milestones in the field, including the shift towards longitudinal analysis and the increased use of deep learning methods. This evolution underscores the importance of tracking psychological well-being over time, providing a framework for more sophisticated, time-sensitive assessments. The clinical relevance of the model is reinforced through comparisons with established psychiatric scales such as the Global Assessment of Functioning (GAF) and the Patient Health Questionnaire (PHQ-9). Preliminary expert feedback indicates that the well-being scores and generated summaries hold promise for practical application. Although clinical validation remains in its early stages, this alignment with clinical standards offers a strong foundation for future trials and real-world implementation, ensuring that the approach can be refined based on direct input from mental health professionals.

By addressing both short-term and long-term psychological trends, our approach bridges the gap between automated analysis and clinical relevance. Future work will focus on refining validation methods and enhancing model adaptability to diverse populations, ensuring broader applicability in mental health monitoring.

7 Conclusion

In this study, we introduced a comprehensive framework for dynamic mental health assessment using social media data. Our approach integrates evidence extraction, well-being scoring, and summary generation to provide a multi-level understanding of psychological states. The results indicate that our methodology yields robust and interpretable outputs, capturing both immediate cues and long-term trends in mental health. By combining feature-based classifiers with context-aware language modeling, our framework offers a scalable solution for digital mental health monitoring. Overall, our work paves the way for future innovations in automated mental health assessments, supporting both research and practical applications in mental health care. This study highlights the potential of advanced NLP techniques.

Limitations

Our approach faced computational and methodological constraints. Using 8-bit quantization and batch processing for Qwen2.5 led to occasional quality tradeoffs, especially with complex psychological patterns. Binary classification of self-states may oversimplify nuanced mental states, and consistency scores (0.612 for Task B, 0.801 for Task C) suggest room for improvement in summary accuracy. The model struggled with temporal reasoning in Task C and lacks direct clinical validation, limiting generalizability. Future work should explore multi-modal inputs and clinically aligned evaluations.

Challenges in temporal analysis include data sparsity, non-stationarity, and evolving behavior. To improve long-term assessments, we propose memory-enhanced models and reinforcement learning for sequence prediction. Bias mitigation strategies include data augmentation and bias-aware training. Detailed documentation of hyperparameters, dataset splits, and ablation studies ensures reproducibility.

Ethical Statement

Secure access to the CLPsych 2025 shared task dataset was provided with appropriate IRB approvals and data use agreements. Our system was designed to analyze sensitive psychological content privately, operating entirely locally without external API dependencies to enhance data protection. We acknowledge the ethical implications of automated analysis of mental health data and emphasize that our approach is intended as a research tool to explore computational methods for self-state detection, not as a clinical diagnostic instrument. Our work aims to support mental health research while maintaining strict protections for the sensitive personal information contained in the dataset.

Acknowledgement

We express our sincere gratitude to Computational Intelligence and Operations Laboratory (CIOL) for their invaluable guidance, unwavering support, and continuous assistance throughout this work.

References

Falwah AlHamed, Julia Ive, and Lucia Specia. 2022. Predicting moments of mood changes overtime from imbalanced social media data. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 239–244.

Barbora Rehák Bučková, Charlotte Fraza, Rastislav Rehák, Marián Kolenič, Christian Beckmann, Filip Španiel, Andre Marquand, and Jaroslav Hlinka. 2025. Using normative models pre-trained on cross-sectional data to evaluate intra-individual longitudinal changes in neuroimaging data. *eLife*, 13.

Ana-Maria Bucur, Hyewon Jang, and Farhana Ferdousi Liza. 2022. Capturing changes in mood over time in longitudinal data using ensemble methodologies. Association for Computational Linguistics.

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.

Yunmei Guo, Ming Zhou, Xin Yan, Ying Liu, and Lianhong Wang. 2024. Latent class analysis and longitudinal development trajectory study of psychological distress in patients with stroke: a study protocol. *Frontiers in Psychiatry*, 15:1326988.

Gyanendro Loitongbam, Junyu Mao, Rudra Mutalik, and Stuart E Middleton. 2024. Extraction and summarization of suicidal ideation evidence in social media content using large language models.

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

Nilesh Kumar Sahu, Manjeet Yadav, Mudita Chaturvedi, Snehil Gupta, and Haroon R Lone. 2025. Leveraging language models for summarizing mental state examinations: A comprehensive evaluation and dataset release. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2658–2682.

Nisha P Shetty, Yashraj Singh, Veeraj Hegde, D Cenitta, and Dhruthi K. 2025. Exploring emotional patterns in social media through nlp models to unravel mental health insights. *Healthcare Technology Letters*.

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.

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

Sumarno Adi Subrata, Heba Mohamed Abdelaal, and Mira Naguib Abd-Elrazek. 2024. Innovation in mental health services: Where are we now? *Innovation in Health for Society*, 4(2):60–68.

Tom Tabak and Matthew Purver. 2020. Temporal mental health dynamics on social media. *arXiv preprint arXiv:2008.13121*.

Masaru Tanaka. 2024. Beyond the boundaries: Transitioning from categorical to dimensional paradigms in mental health diagnostics. *Advances in Clinical and Experimental Medicine*, 33(12):1295–1301.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, et al. 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.

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.

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.

Ayah Zirikly, Philip Resnik, Ozlem 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.

A Appendix

A.1 Explanations

Here we explain several terms for clinical readers:

1. **GradientBoostingRegressor:** This is a machine learning technique that builds predictions by combining multiple weak decision-making models in sequence, where each new model corrects the errors of the previous ones. It is particularly useful in predictive modeling tasks that require fine-tuned adjustments over time.
   Clinical relevance: In mental health applications, this method can refine predictions of psychological well-being by iteratively learning from past misclassifications, helping to detect patterns in mood fluctuations or early signs of distress with improved accuracy.

2. **Temporal Aggregation:** Temporal aggregation refers to grouping data over specific time intervals to observe trends and changes over time. Instead of analyzing individual data points in isolation, it allows for the identification of broader patterns across hours, days, weeks, or months.
   Clinical relevance: In mental health assessments, aggregating social media activity or self-reported symptoms over time can provide deeper insights into long-term behavioral shifts, enabling clinicians to differentiate between short-term fluctuations and sustained changes in mental health.

3. **Intermediate Checkpoint Saving:** This refers to saving the progress of a machine learning model at various stages of training, allowing for recovery in case of failure and enabling assessment of performance at different points. This ensures that models are not trained from scratch if issues arise.
   Clinical relevance: In healthcare AI systems, intermediate checkpoints help track a model's development in real time, ensuring that training is progressing as expected. This is particularly valuable in clinical applications where prolonged training times are common, and periodic evaluations are necessary to validate reliability before deployment.

4. **Narrative Arc Analysis:** This technique examines the structure and evolution of a storyline, mapping key transitions such as rising and falling trends. In data analysis, it is used to understand changes in emotional expression or behavioral patterns over time.
   Clinical relevance: For mental health applications, narrative arc analysis can help detect critical shifts in a patient's emotional state based on their social media activity or self-reported narratives. It allows researchers and clinicians to identify potential crises or significant improvements by analyzing the trajectory of sentiment and language use.