A Multi-Strategy Fusion Framework for Dynamic Mental State Modeling

Mengjia Zhang Rui Chen Haonan Xiao Yi Yang
Chongqing Technology and Business University
{zhangmengjia1, chenrui1, xiaohaonan1, yangchongyi}@ctbu.edu.cn

Abstract

Targeting the CLPsych 2026 (Ali et al., 2026) shared task, this study proposes a unified framework based on three complementary strategies. We implement a pipeline integrating post-level psychological state recognition, timeline-level change detection, and evidence-based clinical summarization. For Task 1, we evaluate three paradigms for extracting ABCD elements: zero-shot prompting, Qwen-3 fine-tuning, and TextCNN-based hierarchical classification. For Task 2, we design a cascaded temporal model that combines Mental-BERT features, log-time encoding, and Bi-LSTM to capture mental state switches and escalations under irregular posting intervals. For Task 3, we adopt a dynamic retrieval-augmented generation strategy to produce structured clinical summaries. Experimental results show that our system ranks second on the official leaderboard for Task 2, demonstrating the effectiveness of time-aware modeling for change detection. Error analysis further reveals over-prediction bias and valence inversion in Tasks 1 and 3, highlighting pipeline error propagation as a key challenge for future work.

1 Introduction

The Computational Linguistics and Clinical Psychology (CLPsych) workshop has hosted shared tasks focused on longitudinal mental state modeling from social media timelines (Tsakalidis et al., 2022a). Building on previous editions, the CLPsych 2026 Shared Task (Ali et al., 2026) focuses on capturing fine-grained mental health changes under the MIND framework (Atzil-Slonim, 2025), requiring psychological element extraction, temporal change detection, and evidence-based clinical summarization.

However, most existing methods do not adequately model the temporal dynamics of mental states and struggle with sparse, irregular posting sequences, while clinical interpretability remains limited. Automated detection of such dynamic changes is also critical for understanding the mechanisms underlying therapeutic change (Atzil-Slonim, 2026),where changes are operationalized as Switches (abrupt shifts) and Escalations (gradual intensifications) following(Tsakalidis et al., 2022b).

This work addresses three core questions: (1) how to accurately extract complex ABCD elements with limited annotations; (2) how to identify state fluctuations and deteriorations in irregular timelines; and (3) how to improve clinical interpretability through evidence-based summarization (Atzil-Slonim, 2026). Our contributions include a systematic comparison of ABCD extraction paradigms, a time-aware cascaded temporal model, and a dynamic RAG framework for clinical summary generation.

2 Related Work

Computational mental health research has transitioned from early bag-of-words and shallow classifiers (Zantvoort et al., 2023) to dense representations powered by pre-trained language models (Ji et al., 2022). However, general-purpose language models often exhibit semantic bias and struggle with highly specialized clinical expressions (Omar et al., 2025). While domain-specific models like Mental-BERT (Ji et al., 2022) mitigate this gap by encoding deep psychological patterns, their application has historically been limited to static, post-level classification rather than longitudinal trajectory modeling.

Longitudinal modeling from social media timelines has gained traction since CLPsych 2022 (Tsakalidis et al., 2022a). Nonetheless, sequential modeling in this domain faces two severe bottlenecks. First, data scarification and highly irregular posting intervals present substantial noise for standard recurrent architectures. Second, be-

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 422–434

cause critical psychological change points (e.g., crises) are inherently sparse, conventional sequence optimizers frequently collapse into majority-class predictions, missing pivotal turning points (Tsakalidis et al., 2022a).

Furthermore, traditional systems typically treat diverse mental states as isolated, orthogonal variables. This directly violates clinical reality, particularly the systemic interweaving and dynamic coupling among Affect, Behavior, Cognition, and Desire (ABCD) highlighted by the MIND framework (Atzil-Slonim, 2025). In the LLM era, although zero-shot prompting and retrieval-augmented methods have been introduced for clinical summarization (Tseriotou et al., 2025), bridging the gap between fine-grained leaf-node classifiers and hallucination-free generative reasoning remains a challenge.To address these limitations, our system explores cascaded temporal modeling with log-time differences, multi-strategy ABCD extraction, and dynamic retrieval-augmented generation for clinical summarization.

3 Methodology

As illustrated in Figure 2, our framework strictly follows the three CLPsych 2026 sub-tasks: (1) identification and existence scoring of adaptive/maladaptive ABCD combinations; (2) detection of Moments of Change (MoC), distinguishing *switches* and *escalations*; and (3) generation of structured clinical summaries grounded in temporal information.

3.1 Task 1: ABCD Element Extraction

We explore three paradigms for transforming raw text into structured MIND annotations. First, we adopt zero-shot prompting with Llama-3.1-8B and constrained decoding to generate structured outputs. Second, we perform instruction-based fine-tuning by adapting Qwen-3-8B with 4-bit QLoRA under the LLaMA-Factory framework. Third, we employ a hierarchical discriminative CNN based on TextCNN, implemented via the NeuralClassifier (Liu et al., 2019) toolkit, to identify fine-grained leaf-node labels including Affect, Behavior, Cognition, and Desire, while capturing local n-gram features critical for clinical text analysis.

3.2 Task 2: Moment of Change Detection

To capture mental state transitions within sparse and irregular social media time series, we design a time-aware temporal modeling framework that extracts domain-specific semantic representations using Mental-BERT and captures irregular posting intervals through Log-scale Temporal Encoding. The model employs a hierarchical cascaded architecture where the predicted probability of a mental state switch is integrated as an auxiliary feature to guide and intensify the detection of risk escalations.

Domain-enhanced Encoding. We employ *Mental-BERT (base-uncased)* (Ji et al., 2022), pre-trained on large-scale mental health corpora, to extract semantic representations. Each post p_i is encoded into a 768-dimensional vector $\mathbf{v}_i$ via the `[CLS]` token, capturing domain-specific psychological patterns that general-purpose models might overlook.

Log-scale Temporal Encoding. To handle the inherent irregularity of posting intervals in longitudinal data, we explicitly model the temporal distance between adjacent posts. Let $\Delta t_i = t_i - t_{i-1}$ be the time interval (in hours) between consecutive posts. We compute a log-scale temporal feature f_{temp} as:

$$f_{temp} = \log(\Delta t_i + 1) \tag{1}$$

This scalar is concatenated with the semantic vector $\mathbf{v}_i$, resulting in a 769-dimensional time-aware input $\mathbf{x}_i = [\mathbf{v}_i; f_{temp}]$.

Cascaded Bi-LSTM. A three-layer Bidirectional LSTM (Bi-LSTM) is used as the backbone to capture long-range dependencies across the timeline. Following the clinical intuition that state escalations often coincide with transitions, we design a hierarchical prediction strategy that first predicts the probability of a state switch, then integrates this switch probability as an auxiliary feature into a second prediction head to guide escalation detection, so that the model can better focus on critical transition points when identifying intensified risks.

Training and Inference. We utilize *Focal Loss* (Lin et al., 2017) combined with dynamic positive weight balancing to mitigate the extreme class imbalance (i.e., the sparsity of change points). During inference, the initial post of each timeline is assigned a 'no-change' status by default to serve as a stable baseline anchor for longitudinal comparison.

3.3 Task 3.1: Evidence-Grounded Summary Generation

While initial experiments for Task 1 utilized the Qwen-3-8B-Instruct model, subsequent evaluations

sub ID	task1.1	task1.2	task2	task3.1					task3.2	
	MF1↑	RMSE↓	Comb MF1↑	CS↑	CT↓	R-L↑	BS↑	Avg↑	Imp↑	Det↑
1	0.118	**1.501**	0.470	0.622	0.859	0.239	——	——	——	——
2	0.210	1.506	0.397	**0.615**	**0.848**	**0.232**	**0.317**	**0.329**	——	——
3	**0.236**	1.585	**0.588**	0.633	0.842	0.200	——	——	0.125	0.500
Rank	14	16	2	12	12	8	7	10	7	5

Table 1: Overall submission results across all tasks and sub-tasks, including primary/secondary metrics and comparative official rankings. Here ↑ indicates higher scores are better; ↓ indicates lower scores are better. Metrics in **bold with arrows** (e.g., **MF1↑**, **RMSE↓**) are the **primary metrics used for official ranking**.
Metrics Definition: MF1: Subelement Macro F_1; **RMSE**: Avg RMSE for Maladaptive + Adaptive (**Lower is better**); **Comb MF1**: Combined Post/Timeline Macro F_1; **CS**: Consistency Score; **CT**: Contradiction Score (**Lower is better**); **R-L**: ROUGE-L Recall; **BS**: BERTscore Recall;**Avg**: Score Average; **Imp/Det**: Improvement/Deterioration Overall Score.

revealed suboptimal performance in extracting fine-grained psychological elements. Consequently, we shifted to the Qwen-2.5-7B-Instruct architecture for Task 3 fine-tuning via the Unsloth framework, to better align the generative process with clinical narrative logic and ABCD interaction dynamics, facilitating a more structured representation of the MIND framework.

To ensure the model's robustness during inference, the input context concatenates the raw text of each post with its corresponding predicted Task 1 ABCD annotations and Task 2 change labels. This pipeline ensures that the generated summaries rely strictly on predicted evidence rather than ground-truth labels, mirroring real-world application.

To mitigate overfitting and enhance clinical coherence, we implemented a Retrieval-Augmented Generation (RAG) strategy using `BGE-large-en-v1.5` (Xiao et al., 2023). Specifically, for each input query sequence, we map its textual representation into a high-dimensional dense vector space and compute the semantic similarity against the pre-computed training repository. The similarity score between a query vector $\mathbf{u}$ and a candidate instance vector $\mathbf{v}$ is evaluated via cosine similarity, formulated as:

$$\text{Sim}(\mathbf{u}, \mathbf{v}) = \cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2} \qquad (2)$$

where d denotes the embedding dimension of the dense encoder. Based on these calculated scores, we dynamically retrieve the K ($K = 2$) most semantically relevant historical sequences from the training set and prepend their corresponding gold summaries to the context window as in-context exemplars. The final clinical summaries are generated via greedy decoding, truncated to a maximum of 350 words, and further constrained by a post-hoc filtering layer to ensure all cited ABCD elements are present in the source text.

3.4 Task 3.2: Dynamic Signature Extraction

For Task 3.2, we identify recurrent dynamic signatures of deterioration and improvement using a multi-stage hybrid strategy. We first conduct statistical profiling via frequency analysis on training summaries to capture dominant ABCD state transitions. We then perform representative clustering using K-Means on BGE-encoded sequence embeddings to select the most typical evidence sequences for each signature type. Finally, we carry out thematic distillation using a local Qwen-2.5-7B model to extract core academic terms and behavioral patterns from the clustering results. These extracted components are synthesized into structured academic templates, with each signature description constrained to 350 words to satisfy submission requirements.

4 Experiments

4.1 Dataset, Metrics, and Compliance

Dataset. The official CLPsych 2026 dataset comprises 373 posts across 30 user timelines, where 236 posts (63.27%) are labeled with fine-grained MIND annotations. Given the frequent co-occurrence of adaptive and maladaptive states, the task demands high-resolution discrimination. Text preprocessing involves removing hyperlinks and special symbols. For Task 2, timelines are chronologically ordered, and variable-length sequences are handled using masking.

Metrics. To ensure a fair, rigorous, and standardized comparison, our evaluation strictly adheres to the official metrics defined by the CLPsych 2026 shared task. Crucially, all quantitative results reported herein were directly evaluated by the shared

task organizers on their official platform using our blind submissions, thereby establishing the independent validity and credence of our system's performance. For Task 1.1 use **Precision**, **Recall**, and **Macro-F_1** over 12 element-valence pairs, ranking by the mean of Adaptive and Maladaptive Macro-F_1. Task 1.2 employs **MAE**, **RMSE**, and **Spearman correlation** on 1–5 scale ratings, with ranking determined by the mean of Adaptive and Maladaptive RMSE. Task 2 is a binary classification task for *Switch* and *Escalation*, evaluated via **Macro-F_1** at post and timeline levels, with the final score defined as:

$$\text{Score} = \frac{F_{1\text{post}} + F_{1\text{time}}}{2} \tag{3}$$

Task 3.1 uses ROUGE-L, BERTScore, Consistency Score (CS), and Contradiction Top (CT). Task 3.2 is qualitatively evaluated by experts regarding *evidence fit*, *recurrence*, and *specificity*.

Open-source Compliance. All experiments strictly adhere to the CLPsych 2026 Data Access Form. We rely exclusively on open-source models, including `Llama-3.1-8B`, `Qwen-3-8B`, `Qwen-2.5-7B`, `Mental-BERT`, and `BGE-large-en-v1.5`, which are deployed locally or via Hugging Face. Notably, no proprietary APIs (e.g., ChatGPT, Claude, Gemini) were employed at any stage.

4.2 Baselines

All experiments in this work compare our proposed method against the official baselines released by the shared task organizers.

5 Results and Analysis

Main Results. Table 1 summarizes the overall performance of our system across all tasks, while Tables 3 – 6 provide detailed, task-specific results. All scores reported in these tables are from the official evaluation conducted by the shared task organizers, ensuring the validity of our quantitative results. Each task-specific table compares our submitted system against the official shared-task baseline, reporting both the primary ranking metric and supporting secondary metrics.

In **Task 1**, the TextCNN-based model achieves a mean F_1 score of 0.236 and an RMSE of 1.585. For **Task 2**, our Mental-BERT cascaded architecture attains a combined Macro-F_1 score of **0.588**, ranking **second** on the official leaderboard. In **Task 3.1** (Table 5), our RAG-enhanced model reaches a BERTScore of 0.317, though its consistency score (CS) of 0.615 and ROUGE-L of 0.232 remain below the official baselines. However, in **Task 3.2** (Table 6), our approach achieves an overall score of 0.500 for the deterioration scenario, outperforming the official baseline (0.483). This result provides preliminary evidence that our model can capture certain patterns of clinical decline.

Ablation. To evaluate the contribution of each module, we conduct a series of ablation experiments on the official validation split of the CLPsych 2026 dataset.Table 7 validates each Task 2 component. Removing Bi-LSTM causes a catastrophic drop (-0.158), confirming the necessity of sequential context. Replacing Focal Loss with standard BCE degrades performance by -0.098, and removing `pos_weight` further drops it to 0.430, underscoring the importance of hard-example weighting for sparse clinical labels. Eliminating log-time diff incurs a modest but non-negligible penalty (-0.018), indicating that irregular intervals (e.g., bursts of posts or long silences) carry hidden temporal cues. Interestingly, decoupling Switch guidance from Escalation slightly improves score (+0.012), suggesting the cascaded design may introduce excessive dependency propagation; future work will explore gated fusion.

Discussion. While our second-place finish in Task 2 underscores the efficacy of domain-specific temporal modeling, the sub-baseline performance in Tasks 1 and 3 reveals substantial error propagation. Task 1 suffers from prediction pattern rigidity, where the model consistently over-predicts a fixed set of maladaptive elements (particularly C-S=2, C-O=2) across diverse posts, failing to discriminate between metaphorical language and genuine clinical symptoms. These recognition inaccuracies propagate to Task 3, manifesting as framework hallucinations where the generative model infers ABCD elements not explicitly present in the source text. This suggests that while cascaded models excel at macro-level change detection, bridging the gap between hierarchical classifiers and nuanced clinical reasoning remains a primary challenge.

Error Analysis for Task 1 Task 1 exhibits **over-prediction bias** and **label combination rigidity**. The model defaults to stereotyped configurations (C-S=2, C-O=2) across diverse posts, regardless of actual clinical content. **Example 1: Non-clinical over-prediction.** A casual movie discussion (*"Gol-*

lum is exactly what's going on in my head") receives six maladaptive elements (C-S=2, C-O=2, A=8, D=4, B-S=2, B-O=4), identical to severe depression posts. The model fails to recognize metaphorical language. **Example 2: Lack of discrimination.** An acute panic attack post and a chronic social drift post receive **identical predictions** (C-S=2, C-O=2, A=8, D=4, B-S=2, B-O=4), despite requiring clinically distinct ABCD profiles. **Root causes:** (1) low uniform threshold (0.15) admitting false positives; (2) absence of top-k constraints; (3) hierarchical classification bias toward complete label paths. These inference-level deficiencies suggest that post-hoc calibration may substantially improve performance.

Error Analysis for Task 3.1. Task 3.1 exhibits structural framework discrepancies (CS: 0.615; CT: 0.848) driven by two primary failure modes: (1) *Subelement Confusion and Valence Inversion*—The generative model misinterprets upstream predictions at the fine-grained level and fails to maintain polar consistency. As typified in sequence `seq8161a42a4`, while Task 1 correctly identifies adaptive coping markers (`D=1`), the text generation pipeline suffers a valence inversion, mapping it to a negative clause associated with `D=2` (*"expectation that it would persist ('will never be able')"*). Furthermore, driven by an inherent "completeness bias," the LLM over-generates non-predicted elements to synthetically balance the ABCD matrix. (2) *Temporal Misattribution and Signal Defiance*—The model fails to preserve longitudinal alignment across the timeline sequence. In `seq8161a42a4`, specific functional impairments such as severe avolition and somatic crisis (`B-S=2`, `B-O=4`), which were uniquely predicted in the late-stage post (*"sleeping 12 hours... tension headache"*), are erroneously cross-contaminated and anchored onto the summary of the initial checkpoint. This temporal drift is compounded by instruction defiance: the generation framework rejects the explicit upstream Task 2 transition token (`Switch=S`), instead fabricating a contradictory `Escalation=E` narrative. Consequently, while the metaphorical discourse in the subsequent post (*"Gollum is exactly what's going on in my head"*) receives highly coherent paraphrasing, the cumulative tracking yields a high BERTScore (0.317) but low factual consistency, highlighting the trade-off between narrative fluency and structural fidelity.

6 Conclusion

We presented a multi-strategy framework for CLPsych 2026 that integrates Mental-BERT domain embeddings, log-scale temporal encoding, and dynamic RAG. While the cascaded Bi-LSTM model demonstrated efficacy in capturing macro-level psychological switches and escalations, Task 1 and 3 results underscore the persistent challenges in fine-grained clinical element recognition and hallucination control in generative summaries. This work highlights the potential of temporal-aware modeling and identifies key failure modes for error propagation in automated mental health trajectory monitoring.

Ethics Statement

The dataset consists of anonymized Reddit posts with personally identifiable information removed by the organizers. This paper contains **no verbatim examples** from the raw dataset; any illustrative content is paraphrased or synthesized from official guidelines. All models (Llama-3.1-8B, Qwen-3-8B, Qwen-2.5-7B, Mental-BERT, BGE-large-en-v1.5) are **strictly open-source** and deployed locally or via Hugging Face, fully complying with the CLPsych 2026 Data Access Form. No proprietary APIs (e.g., ChatGPT, Claude, Gemini) are used. The system is intended for research assistance only; it does **not** constitute clinical diagnosis or treatment advice. Automated mental-health assessments carry risks of false positives/negatives and must be reviewed by qualified clinicians before any clinical application.

Limitations

Our work is constrained by the small-scale dataset (30 timelines, 373 posts), which limits cross-demographic generalization. All experiments are monolingual (English), leaving cross-cultural applicability unverified. Hardware restrictions (8GB VRAM) prevent exploration of larger backbones (e.g., 13B+ parameters) that might improve representation capacity. Task 1 did not surpass the official baseline, suggesting that fine-grained clinical element extraction remains an open challenge. Task 3.2 relies on heuristic rules and clustering, which may not generalize to out-of-distribution narrative patterns.

Acknowledgments

We thank the CLPsych 2026 organizers for the dataset and evaluation framework. We acknowledge the open-source communities behind Hugging Face, Unsloth, and LLaMA-Factory.

References

Iqra Ali, Talia Tseriotou, Guy Dvir, Callum Chan, Yuxiang Zhou, Juan Antonio Lossio-Ventura, Ayal Klein, Aya Shamir, Dan Sayda, Anthony Hills, Aya Zirikly, Diana Inkpen, Dana Atzil-Slonim, and Maria Liakata. 2026. Overview of the clpsych 2026 shared task: Capturing and characterizing mental health changes through social media timeline dynamics. In *Proceedings of the 11th Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

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

Dana Atzil-Slonim. 2026. Leveraging theoretical and technological innovations to study the mechanisms that underlie therapeutic change in psychotherapy. In Louis G. Castonguay, Dana Atzil-Slonim, Michael Barkham, and Wolfgang Lutz, editors, *Practice-Based Evidence in the Psychological Therapies: Toward Policy Implications for Research, Training, and Clinical Guidelines*. Oxford University Press, New York.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *proceedings of the thirteenth language resources and evaluation conference*, pages 7184–7190.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Liqun Liu, Funan Mu, Pengyu Li, Xin Mu, Jing Tang, Xingsheng Ai, Ran Fu, Lifeng Wang, and Xing Zhou. 2019. Neuralclassifier: An open-source neural hierarchical multi-label text classification toolkit. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–92.

Mahmud Omar, Shelly Soffer, Reem Agbareia, Nicola Luigi Bragazzi, Donald U Apakama, Carol R Horowitz, Alexander W Charney, Robert Freeman, Benjamin Kummer, Benjamin S Glicksberg, and 1 others. 2025. Sociodemographic biases in medical decision making by large language models. *Nature Medicine*, 31(6):1873–1881.

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. 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, Seattle, USA. Association for Computational Linguistics.

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, Jenny Chim, Ayal Klein, Aya Shamir, Guy Dvir, Iqra Ali, Cian Kennedy, Guneet Singh Kohli, Anthony Hills, Ayah Zirikly, Dana Atzil-Slonim, and Maria Liakata. 2025. Overview of the CLPsych 2025 shared task: Capturing mental health dynamics from social media timelines. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 193–217, Albuquerque, New Mexico. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Kirsten Zantvoort, Jonas Scharfenberger, Leif Boß, Dirk Lehr, and Burkhardt Funk. 2023. Finding the best match—a case study on the (text-) feature and model choice in digital mental health interventions. *Journal of Healthcare Informatics Research*, 7(4):447–479.

Appendix

A Data Specification and Framework Mapping Exemplars

To ensure that the zero-shot open-source large language models (e.g., `Llama-3.1-8B`) strictly adhere to the hierarchical MIND coding manual without encountering parsing failures, we engineered a deterministic, role-based prompting schema.

As detailed in the template below, the prompt architecture functions as a gatekeeper for three core operational constraints:

1. **Clinical Alignment:** It explicitly enforces the structural dichotomy of the MIND framework, where even-numbered sub-elements map onto *Maladaptive* behaviors and odd-numbered sub-elements correspond to *Adaptive* coping strategies.

2. **Presence Score Calibration:** To prevent the pipeline from crashing on posts devoid of

active psychological cues, the prompt constraints the model to guarantee a baseline output (`"presence": 1`) instead of producing null fields.

3. **JSON Schema Adherence:** It restricts the generation path exclusively to valid JSON tokens via constrained decoding instructions, eliminating conversational verbiage or chain-of-thought metadata that could impede automated string parsing.

A.1 Task 1 Zero-Shot Annotation Prompt Template

In this section, we provide the specific prompt templates used for Task 1: ABCD Element Extraction using the zero-shot Llama-3.1-8B model.

Listing 1: System prompt and formatting constraints for Task 1 ABCD extraction.

```
You are a clinical psychology
    annotation expert for CLPsych
     2026.
Your task is to analyze social
    media posts using the MIND
    framework (ABCD model).

CRITICAL RULES:
1. Identify ABCD elements for
    BOTH Adaptive and Maladaptive
     states (if present in post)
2. For each element, output ONLY
     the subelement NUMBER (
    integer), NOT text
    descriptions
3. PRESENCE SCORE (1-5) rates
    how dominant the ENTIRE state
     is in the post:
    - 1 = Not present at all
    - 2 = Somewhat present (
       subtle role)
    - 3 = Moderately present (
       clearly contributes)
    - 4 = Much present (strongly
       influences)
    - 5 = Highly present (defines
        the overall experience)
4. Only include elements that
    are CLEARLY expressed in the
    post
5. For each element (A, B-O, B-S
    , C-O, C-S, D), you can
    identify at most ONE
    subelement per valence

VALID SUBELEMENT NUMBERS:
- A (Affect): 1-14
  Adaptive: 1,3,5,7,9,11,13 |
     Maladaptive:
     2,4,6,8,10,12,14
- B-O (Behavior-Others): 1-4
  Adaptive: 1,3 | Maladaptive:
     2,4
- B-S (Behavior-Self): 1-2
  Adaptive: 1 | Maladaptive: 2
- C-O (Cognition-Others): 1-4
  Adaptive: 1,3 | Maladaptive:
     2,4
- C-S (Cognition-Self): 1-2
  Adaptive: 1 | Maladaptive: 2
- D (Desire): 1-6
  Adaptive: 1,3,5 | Maladaptive:
      2,4,6

OUTPUT FORMAT (STRICT JSON - NO
    TEXT DESCRIPTIONS):
{
  "adaptive": {
    "presence": 1-5,
    "elements": {
      "A":1,
      ...
    }
  },
  "maladaptive": {
    "presence": 1-5,
    "elements": {
      ...
    }
  }
}
If a state is not present, use:
    {"presence": 1, "elements":
    {}}
```

A.2 Exploratory Multi-Task Joint Fine-Tuning Setup

During the exploratory phase for Task 1 and Task 2, we investigated a generative alternative: transforming the longitudinal social media streams into a unified, sequential instruction-following dataset for Supervised Fine-Tuning (SFT). This pipeline was implemented via the LLaMA-Factory toolkit

using an autoregressive `Qwen-3-8B` backbone.

The core engineering strategy relied on a **Rolling Chronological Window** to operationalize dynamic psychological trajectories:

- **Longitudinal Context Buffering:** For any target post p_t (*Current Post*), all prior historical posts $\{p_1, p_2, \ldots, p_{t-1}\}$ are formatted with precise timestamps and appended into the *Social Media Context* field. This provides the causal attention mechanism with the explicit baseline needed to distinguish static personality traits from acute psychological fluctuations.

- **Multi-Task Joint Reasoning:** Rather than optimizing separate task-specific heads, the target *Output* sequence is formulated to output a unified textual token stream. As formalized in the structured schema, the model is forced to jointly decode macro-level transitions (*Switch* and *Escalation*) alongside micro-level framework alignments (leaf-node ABCD segments and explicit *Presence* intensity ratings).

Table 2 provides a series of consecutive, anonymized training exemplars illustrating how our pipeline tracks a user's progressive shifts from severe distress to adaptive coping, and back to acute interpersonal anxiety.

A.3 Hierarchical Discriminative Classification Framework (NeuralClassifier)

To overcome the limitations of large language models in fine-grained clinical structural tracking, our configuration for Task 1 utilizes a discriminative, hierarchical multi-label text classification paradigm powered by the open-source `NeuralClassifier` framework.

A.3.1 Taxonomy Formalization

As illustrated in Figure 1, we mapped the clinical MIND manual into a strict three-tier structured hierarchy to formalize the semantic dependencies between high-level psychological dimensions and fine-grained behavioral representations:

- **Tier 1 (Root Nodes):** Formulates the 6 foundational dimensions of the ABCD model: Activating Event (`A`), Belief about Others (`B-O`), Belief about Self (`B-S`), Consequence regarding Others (`C-O`), Consequence regarding Self (`C-S`), and Disputation (`D`).

- **Tier 2 (Valence Dichotomy):** Branching from each root node into binary valences: *Adaptive* (`A`, e.g., `AA`, `B-SA`) representing positive coping, and *Maladaptive* (`M`, e.g., `AM`, `B-SM`) representing psychological vulnerability.

- **Tier 3 (Leaf Nodes):** Mapping directly onto explicit clinical sub-categories (e.g., `AM2` for anxious/fearful affect, `DM6` for self-criticism disputation) spanning 12 element-valence terminal pairs.

A.3.2 Data Synthesis Pipeline

To feed the discriminative hierarchical neural net (TextCNN backbone), our custom preprocessing pipeline flattens the clinical data into a specialized JSON token format. For any given text span, the label generator dynamically parses the leaf-node category and automatically expands the complete ancestor path from Tier 3 back to Tier 1, enforcing deterministic path constraints. A representative tokenized sample is formalized as follows:

```
{ "doc_label":  ["B-SA", "DA", "AA5",
"B-SA1"],
"doc_token":  ["after", "8", "years",
"of", "struggling", "with", "self",
"harm", "i", "am", "starting", "to",
"cover", "my", "older", "scars..."] }
```

By optimizing a localized hierarchical loss function rather than vanilla flat classifiers, the TextCNN backbone explicitly captures the directional entailment (e.g., predicting `AM2` implies absolute activation of `AM` and `A`), significantly stabilizing structural element boundaries under low-resource validation.

B System Architecture Diagram

Field	Content / Multi-Task Rolling Window Schema
Instruction	Analyze the psychological state of the 'Current Post' based on the provided 'Social Media Context'. Tasks: 1. Identify 'Switch' and 'Escalation'. 2. Extract ABCD framework evidence. 3. Assign 'Presence' scores.
Exemplar 1: Baseline Distress (Chronological Anchor)	
Input	**### Social Media Context (Previous Posts):** [02-09-2019] Post: I'm mean **### Current Post to Analyze:** Date: 04-09-2019 Content: I'm not okay... I wish someone actually liked me... I wish I didn't wish I were dead.
Output	**[Change Analysis]** - Switch: 0; - Escalation: 0; - Well-being: 5 **[Adaptive State]** - Dimension D: I wish that I want to make my life worth living... [Presence Score: 3] **[Maladaptive State]** - Dimension C-O: I wish someone wanted to hold my hand... [Presence Score: 4]
Exemplar 2: Transition Trigger (Altruistic Coping - State Switch Detected)	
Input	**### Social Media Context (Previous Posts):** [02-09-2019] Post: I'm mean [04-09-2019] Post: I'm not okay... I wish I didn't wish I were dead. **### Current Post to Analyze:** Date: 05-09-2019 Content: [HELP!] Calling All Animal Fosters... I found a bundle of kittens... I want to try my best to get these cats a better life...
Output	**[Change Analysis] - Switch: S**; - Escalation: 0; - Well-being: 7 **[Adaptive State]** - Dimension C-S: I want to try my best to get these cats a better life... [Presence Score: 4] **[Maladaptive State]** - Dimension A: I am in desperate need of some help right now. [Presence Score: 2]
Exemplar 3: Interpersonal Vulnerability (Relapse/Fluctuation Pattern)	
Input	**### Social Media Context (Previous Posts):** [04-09-2019] Post: I'm not okay... I wish I didn't wish I were dead. [05-09-2019] Post: [HELP!] Calling All Animal Fosters... **### Current Post to Analyze:** Date: 08-09-2019 Content: I think I need to work on my self esteem... So I went on a date recently... He didn't really respond and now I'm doubting myself...
Output	**[Change Analysis]** - Switch: 0; - Escalation: 0; - Well-being: 7 **[Adaptive State]** - Dimension B-S: So I went on a date recently and I actually... had a really great time. [Presence Score: 4] **[Maladaptive State]** - Dimension C-S: I'm doubting myself thinking I already messed something up. [Presence Score: 3]

Table 2: Chronological training sequence illustrating the multi-task rolling dynamic window used during the exploratory fine-tuning phase.

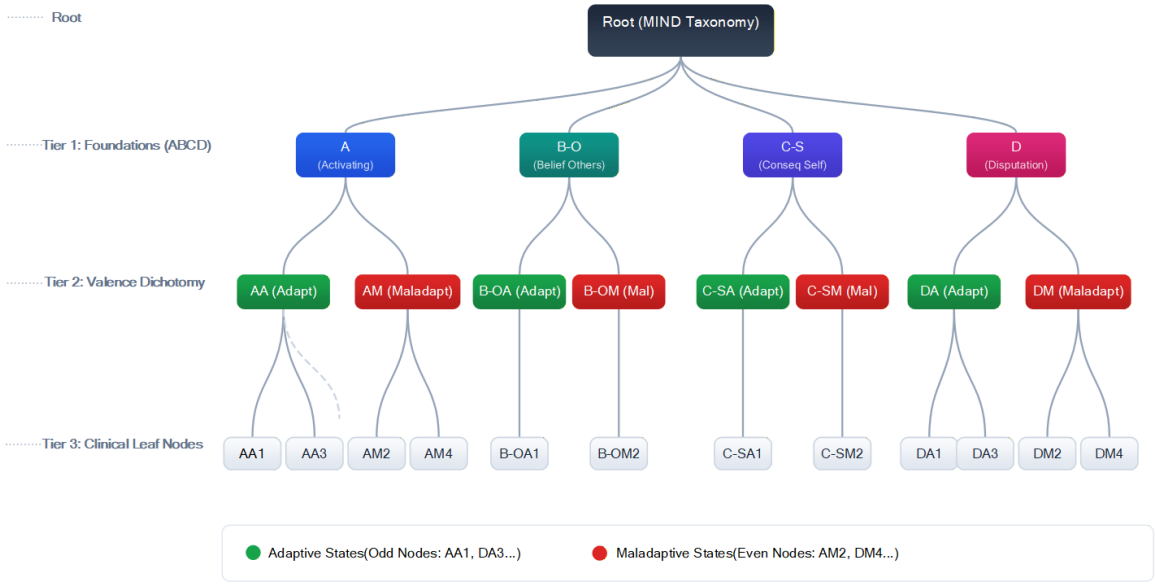


Figure 1: Hierarchical taxonomy tree architecture derived from the formalized clinical MIND framework. The tree explicitly operationalizes a three-tier structured dependency for the `NeuralClassifier` engine: Tier 1 establishes the foundational ABCD coordinates, Tier 2 bifurcates into dynamic binary valences (**Adaptive**vs. **Maladaptive**), and Tier 3 maps onto fine-grained clinical leaf representations (e.g., `AM2`, `DM4`) aligned with our empirical parsing pipeline.

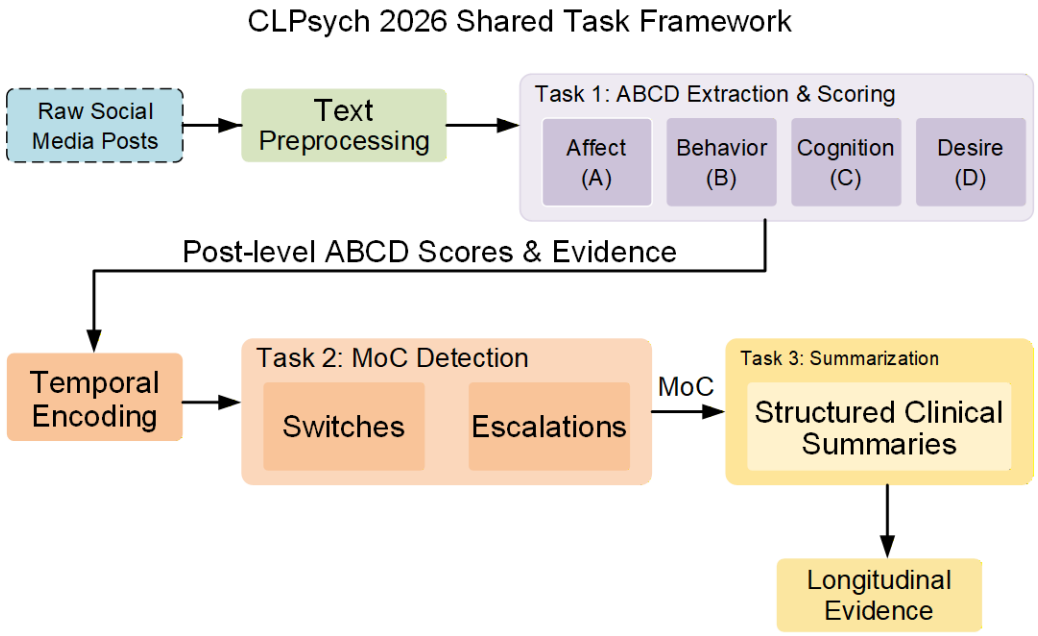


Figure 2: Overall three-stage framework of our proposed system. Clinical rule constraints are enforced throughout the entire pipeline.

sub ID	Core Technical Path	F1 (1.1)↑	Avg RMSE (1.2)↓
1	Ollama + Llama-ZeroShot	0.118	**1.501**
2	Qwen3-base + LLaMA-Factory fine-tuning	0.210	1.506
3	TextCNN + Loss optimization for imbalance	**0.236**	1.585
-	Official Baselines	0.247	1.424

Table 3: System Performance Evolution across Experimental Stages for Task 1.

Note: All scores reported in this table correspond to our official submissions during the formal shared task competition phase.Here, F_1 denotes the *Average Subelement Macro* F_1,while Avg RMSE represent *Avg RMSE Maladaptive + Adaptive*

sub ID	Method	Overall / Comb.↑	P-Macro F1↑	T-Macro F1↑	Rank
1	all-MiniLM + Bi-LSTM	0.470	0.465	0.475	—
2	Qwen-SFT (Task 1 feats)	0.397	0.340	0.455	—
3	Mental-BERT + Log-Time + Focal	**0.588**	**0.611**	**0.564**	**2**
-	Official Baselines 3	0.572	0.561	0.583	—
-	Official Baselines 2	0.365	0.383	0.346	—
-	Official Baselines 1	0.272	0.169	0.400	—

Table 4: Official Codabench submission results for Task 2. Best scores in bold.

Note: All scores reported in this table correspond to our official submissions during the formal shared task competition phase. Here, Overall / Comb., P-Macro F_1, and T-Macro F_1 denote *Combined (Post/Timeline) Macro* F_1, *Post Macro* F_1, and *Timeline Macro* F_1, respectively.

sub ID	Architecture	CS↑	CT↑	R-L↑	BS↑
1	Ollama + Few-shot	0.622	0.860	0.240	–
2	Qwen2.5-7B + LoRA	0.633	0.842	0.199	–
3	SFT + RAG + Alignment	**0.615**	**0.848**	**0.232**	**0.317**
–	Official Baseline 1	0.763	0.753	0.255	0.226
–	Official Baseline 2	0.767	0.745	0.269	0.235

Table 5: System performance for Task 3.1: Change Sequence Summarization. Best scores per column are indicated in bold.

Note: All scores reported in this table correspond to our official submissions during the formal shared task competition phase.**CS**: Consistency Score; **CT**: Contradiction Score (**Lower is better**); **R-L**: ROUGE-L Recall; **BS**: BERTscore Recall.

Method	Scenario	Fit↑	Recur.↑	Spec.↑	Overall↑
Ours	Improvement	0.250	0.250	0.000	0.125
	Deterioration	1.000	1.000	0.000	**0.500**
Baseline	Improvement	0.375	0.437	0.375	0.389
	Deterioration	0.562	0.375	0.437	0.483

Table 6: Qualitative assessment for Task 3.2: Signatures. Best overall scores in bold.

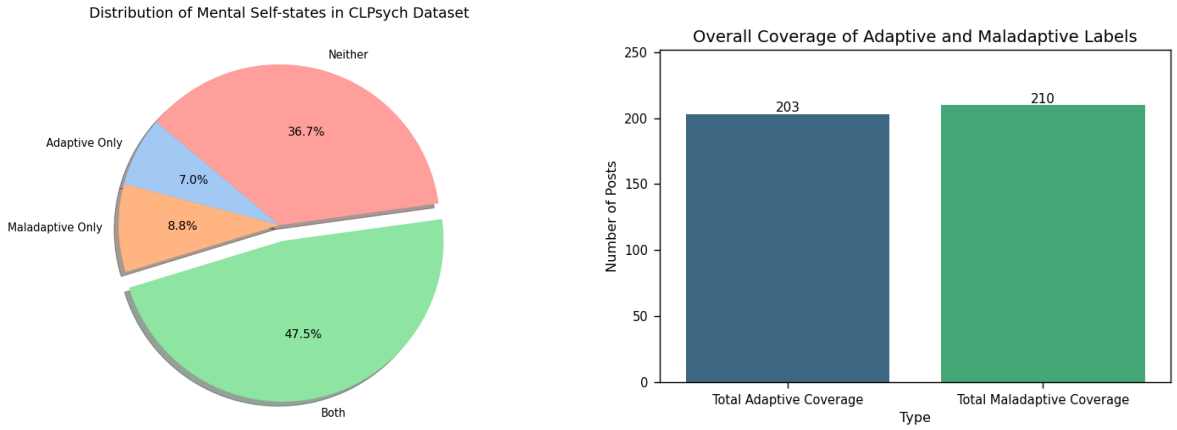
Note: All scores reported in this table correspond to our official submissions during the formal shared task competition phase. Here, Fit, Recur., Spec., and Overall denote *Fit Score*, *Recurrence Score*, *Specificity Score*, and *Overall Score*, respectively.

Table 7: Ablation study on Task 2 (Macro F_1). Δ is relative to the full model.

Variant	Key Modification	Score	Δ
Full Model	Best configuration (Bi-LSTM + Focal)	**0.588**	—
BERT-Only	Remove Bi-LSTM; use MeanPool + MLP	0.430	−0.158
w/o Focal	Replace with standard BCELoss	0.490	−0.098
No PosWeight	Focal loss with `pos_weight=None`	0.430	−0.158
NoSwitchGuide	Escalation detection w/o Switch logit	0.600	+0.012
NoTime	Remove log-time difference features	0.570	−0.018
Uni-LSTM	Use unidirectional LSTM layers	0.470	−0.118

C Dataset Characteristics and Label Distributions

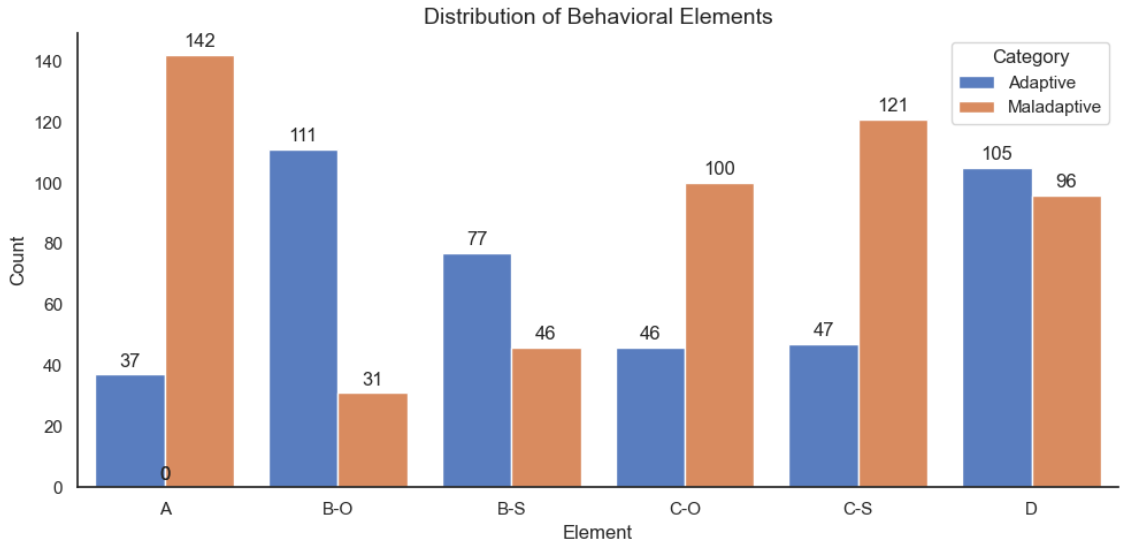
In this section, we present a comprehensive exploratory data analysis (EDA) of the official CLPsych 2026 dataset to provide deeper empirical insights into the distribution and co-occurrence patterns of the MIND framework annotations.As shown in Figure 3 and Figure 4, we systematically analyze the data characteristics from both macro and micro perspectives.



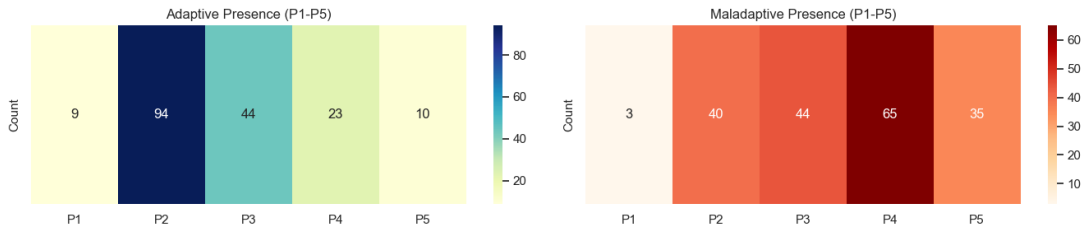
(a) Distribution of mental self-states.

(b) Overall label coverage.

Figure 3: Macro-level analysis of the CLPsych dataset. (a) Illustrates the severe overlap of adaptive and maladaptive states, where nearly half of the posts (47.5%) present both states simultaneously. (b) Demonstrates that adaptive and maladaptive signals maintain an approximately balanced representation across the corpus.



(a) Fine-grained distribution of ABCD sub-elements.



(b) Intensity distribution of Presence Scores (P1–P5).

Figure 4: Micro-level and intensity analysis of MIND annotations. (a) Compares the counts of Affect (A), Behavior-Others (B-O), Behavior-Self (B-S), Cognition-Others (C-O), Cognition-Self (C-S), and Desire (D) across valences, revealing a high concentration of maladaptive Affect and adaptive Behavior-Others. (b) Depicts the density of clinical presence ratings, indicating that adaptive cues lean towards moderate intensity (P2) while maladaptive deterioration presents a significant peak at high intensity (P4).