Hierarchical Multi-Stage Modeling of Adaptive and Maladaptive Self-States in Social Media Timelines

Abir Naskar and Mike Conway
School of Computing and Information Systems, University of Melbourne
Parkville, Melbourne, 3053, VIC, Australia
anaskar@student.unimelb.edu.au, mike.conway@unimelb.edu.au

Abstract

We address the CLPsych 2026 Shared Task on modeling psychological self-states from longitudinal social media data. We propose (i) a hierarchical multi-stage framework that integrates a multi-task transformer encoder and (ii) a four stage instruction-tuned large language model finetuning pipeline for subelement classification, presence estimation, and evidence extraction. Our approach incorporates element-conditioned label masking and cross-stage encoder transfer, enabling structured prediction aligned with the ABCD psychological framework. Experiments show improvements over the baseline on the development setup, with RoBERTa achieving an 8.3% gain in macro-F1 and improved root mean square error (RMSE), while a fine-tuned Qwen3 model attains the best overall performance. These results demonstrate the effectiveness of combining hierarchical multi-task learning with structured generation for interpretable mental health analysis. All code for this task is publicly available in a GitHub repository[1].

1 Introduction

Mental health disorders remain a critical global challenge, motivating the need for scalable methods for early detection and monitoring. Social media provides a valuable source of longitudinal behavioral data, enabling the analysis of individuals' evolving emotional and cognitive states in naturalistic settings (Uban et al., 2021). While early work in clinical NLP focused on static classification tasks such as depression detection, recent research emphasizes modeling fine-grained and dynamic psychological processes over time (Atzil-Slonim, 2026). The Workshop on Computational Linguistics and Clinical Psychology (CLPsych) 2026 Shared Task (Ali et al., 2026) further advances this direction by introducing a structured framework for identifying adaptive and maladaptive self-states grounded in the ABCD (Affect, Behavior, Cognition, Desire) model, requiring both subelement prediction and intensity estimation.

To address these challenges, we propose a hierarchical multi-stage framework that combines discriminative and generative modeling. First, a multi-task transformer jointly predicts ABCD subelements and their presence scores, leveraging shared representations for structured prediction (Tian et al., 2019; Ferraro and Benedetti, 2025). Next, an instruction-tuned language model is used to extract supporting evidence and generate interpretable outputs, enabling transparent reasoning over predictions. Our contributions are as follows: First, we propose a hierarchical multi-stage architecture that decomposes self-state modeling into sequential prediction stages, enabling explicit modeling of dependencies between detection, classification, and evidence extraction. Second, we introduce element-conditioned label masking and cross-stage encoder transfer, which enforce psychologically consistent predictions and improve representation learning. Third, we combine discriminative and generative modeling by integrating a multi-task transformer with an instruction-tuned LLM, enabling both accurate prediction and interpretable structured outputs. Finally, we design a computationally efficient system that supports CPU-based inference while retaining competitive performance.

2 Task and Dataset

The CLPsych 2026 Shared Task extends the longitudinal modeling framework introduced in CLPsych 2022 (Tsakalidis et al., 2022) and 2025 (Tseriotou et al., 2025). It uses an expanded subset of the CLPsych 2022 "Reddit-New" dataset and is grounded in the MIND framework (Atzil-Slonim, 2025; Slonim, 2024), which represents self-states as structured combinations of six psychological

[1] https://github.com/AbeerNaskar/clpsych26-sharedtask1

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 472–481

dimensions: Affect, Behaviour toward Others, Behaviour toward the Self, Cognition of Others, Cognition of the Self, and Desire, each comprising adaptive and maladaptive elements. We focus on Subtask 1, which involves identifying ABCD subelements in posts, determining their composition into adaptive or maladaptive self-states with supporting evidence spans, and predicting their psychological centrality on a 1–5 scale (evaluated using RMSE). The dataset consists of longitudinal timelines, including 30 labelled training timelines (373 posts) and 10 unlabelled test timelines (92 posts). Table 2 summarises the key corpus statistics.

3 Method

We present a dual-approach system for hierarchical multi-task modeling of self-state signals, designed with a focus on low-resource deployment. Our primary approach is a hierarchical multi-task encoder based on pretrained transformers (e.g., BERT/RoBERTa), which jointly predicts state presence and intensity scores, and performs element classification and token-level evidence extraction for detected states. Complementing this, we employ an instruction-tuned large language model adapted via low-rank adapters, which follows a staged prompting pipeline for state detection, element identification, category and evidence generation, and final scoring. Our design is grounded in prior work on pretrained transformers (Vaswani et al., 2017), multi-task learning (Ruder, 2017), instruction tuning (Ouyang et al., 2022), chain-of-thought reasoning (Wei et al., 2022), and evidence extraction.

3.1 Hierarchical Multi-Task Transformer-based Encoder

We implement a two-stage, hierarchical multi-task model using a pretrained encoder-based transformer like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), etc. The architecture diagram is depicted in Figure 1.

3.1.1 Stage 1: Binary Detection and Presence Regression

Stage 1 consists of a binary classifier and a regression unit.

Given an input post $X = (x_1, x_2, \ldots, x_n)$, the encoder produces contextual representations:

$H = Encoder(X) \in R^{n \times d}$.

We extract the pooled representation $h_{cls} = H[0]$, which is passed through dropout and fed into two task-specific heads:

Classification logits:

$z_{cls} = W_{cls}.Dropout(h_{cls}) + b_{cls} \in R^2$

Regression output:

$\hat{p}_{pres} = W_{reg}.Dropout(h_{cls}) + b_{reg} \in R$.

The classification probabilities are obtained via softmax. The regression output is treated as a continuous estimate and clipped to the valid range [1,5] at inference time. The joint training objective is:

$\mathcal{L}_{S1} = \mathcal{L}_{CE}(z_{cls}, y_{cls}) + \lambda \mathcal{L}_{MSE}(\hat{p}_{pres}, y_{pres})$,

where λ is a constant (set to 0.5 in our experiments), $\mathcal{L}_{CE}$ is cross-entropy loss and $\mathcal{L}_{MSE}$ is mean squared error.

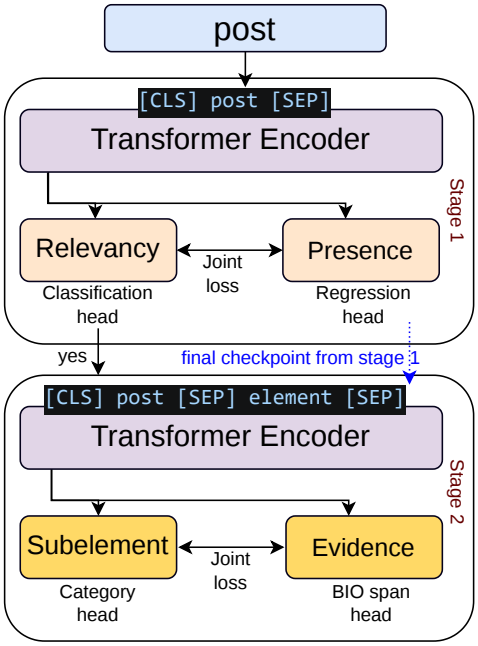

Figure 1: Two-stage hierarchical multi-task transformer-based encoder approach for solving the task.

3.1.2 Stage 2: Element-Conditioned Subelement Classification and Evidence Detection

Stage 2 initialises its encoder from the Stage 1 checkpoint, providing a domain-adapted starting point for the more challenging subelement classification task. Each training example pairs a post x with an ABCD element identifier,

$e \in$ {Adaptive-A, Adaptive-B-O, Adaptive-B-S, Adaptive-C-O, Adaptive-C-S, Adaptive-D, Maladaptive-A, Maladaptive-B-O, Maladaptive-B-S, Maladaptive-C-O, Maladaptive-C-S, Maladaptive-D}

as a text segment, producing the concatenated input: $X' = [CLS] \oplus X \oplus [SEP] \oplus e \oplus [SEP]$.

This element-conditioning allows a single model to serve all six elements, with the element token acting as a task prompt that shifts the [CLS] representation toward the relevant MIND dimension. Two task-specific heads operate on the Stage-2 encoder output. The classification head predicts the subelement category over the full label vocabulary L ($|L_{adaptive}| = 16$, $|L_{maladaptive}| = 16$, and a none label):

$z_{cat} = W_{cat}.Dropout(h'_{cls}) + b_{cat} \in R^{|\mathcal{L}|}$

We enforce a structural constraint by masking invalid labels for each element, ensuring predictions remain within the valid subelement set. Let $V(e)$ denote the set of valid label indices for element e (plus the none index). The masked logits are:

$$\tilde{z}_{cat}[i] = \begin{cases} z_{cat}[i], & if \ \ i \in V(e) \\ -\infty, & \text{otherwise} \end{cases}$$

Predictions are obtained by:

$\hat{y}_{cat} = \arg\max Softmax(\tilde{z}_{cat})$

For the evidence (BIO [2]) span detection, a token-level classifier operates over the sequence representations:

$Z_{BIO} = W_{BIO}H' + b_{BIO} \in R^n$

producing per-token logits indicating evidence span membership. Reference BIO labels are created by aligning spans to tokens. The joint loss is defined as:

$\mathcal{L}_{S2} = \mathcal{L}_{CE}(\tilde{z}_{cat}, y_{cat}) + \mathcal{L}_{BCE}(Z_{BIO}[active], y_{BIO}[active])$

where active masks out padding positions from the BIO loss, and $\mathcal{L}_{BCE}$ is binary cross-entropy loss.

3.1.3 Inference

At inference time, Stage 1 predicts state presence and intensity for each post. Posts predicted as positive are forwarded to Stage 2, where all six element-conditioned predictions are computed in parallel using their respective masks. Posts predicted as negative are not processed further, reducing error propagation to downstream modules. This filtering reduces unnecessary computation and limits error propagation from downstream modules.

[2]BIO = Beginning – Inside – Outside

3.2 Instruction-Tuned LLM Pipeline with Transformer-based Decoder

We propose a multi-stage framework for psychological state prediction using an instruction-tuned LLM with LoRA, decomposing the task into sequential stages: state detection, element identification, category and evidence extraction, and presence refinement. This approach reduces task complexity and enforces intermediate representations, aiming to improve interpretability through intermediate outputs than one-shot prompting. Each stage is trained independently with a causal language modeling objective and executed in a chained manner at inference. While effective for fine-grained prediction, the pipeline remains susceptible to error propagation across stages. Further details of the stages, prompts, and training procedure are provided in Appendix A.2. However, this pipeline may accumulate errors across stages.

3.3 Ablation Study

The ablation study is conducted on Method 1 using a controlled hold-out setup with an 85/15 train–validation split and a fixed random seed for fair comparison. Stage 1 is trained on the training portion for state prediction, while Stage 2 operates only on positively labeled instances expanded to element-level data. A0 denotes the full model with all components enabled. Starting from this baseline, we systematically modify one component at a time to assess the impact of key design choices, including multi-task supervision, class imbalance handling, cross-stage knowledge transfer, span-level supervision, structural constraints, input preprocessing, and fine-tuning strategy. This controlled setup enables precise attribution of performance changes to individual components. Results are reported on a single split and may vary across different random seeds. We do not perform statistical significance testing due to limited data.

4 Experiments and Results

We evaluate two complementary approaches. For Method 1, we use pretrained transformer encoders, BERT-base-uncased and RoBERTa-base, implemented with Hugging Face Transformers (Wolf et al., 2020) and trained in PyTorch. Models are trained with sequence lengths of 256 (Stage 1) and 300 (Stage 2), batch size 16, AdamW optimizer (learning rate 2×10^{-5}), and dropout (0.1). Class imbalance is handled using a weighted random

Model	Subelement Avg Macro F1	RMSE
RoBERTa	0.331	1.146
BERT	0.313	1.255
Qwen3	**0.333**	**1.06**
CLPsych2026 - Baseline	0.247	1.424

Table 1: Subelement-level average Macro-F1 and RMSE results. BERT (submission ID 662787) was our official competition submission during the competition phase; RoBERTa (submission ID 706280) and Qwen3 (submission ID 705090) were evaluated during the post-competition analysis phase and are reported here for completeness. Our official submission ranked in the competition under team name NoviceTrio. The CLPsych 2026 baseline uses one-shot structured prompting with LLaMA-3.1-8B-Instruct.

sampler, with gradient clipping and fixed seeds for stability. We prefer a weighted random sampler over a weighted loss function because it ensures balanced class exposure at the batch level without modifying the loss landscape, which we found more stable under the small dataset regime of this shared task; weighted loss reweighting can destabilise regression heads when classification and regression losses are jointly optimised.

For Method 2, we use Qwen3-4B-Instruct implemented via Unsloth [3], fine-tuned with LoRA (Hu et al., 2022) in 4-bit precision for efficiency. Training follows a four-stage supervised fine-tuning pipeline with an effective batch size of 8, AdamW (8-bit) optimizer (learning rate 2×10^{-4}), and a maximum sequence length of 2048. Each stage is trained for a fixed number of steps, with loss computed only on response tokens.

We submit our system under the team name **NoviceTrio**. Table 1 shows that Qwen3 achieves the best overall performance (Macro-F1 = 0.333, RMSE = 1.06), slightly outperforming RoBERTa (0.331, 1.146), while both models significantly surpass BERT (0.313, 1.255) and the CLPsych 2026 baseline (0.247, 1.424). RoBERTa, however, provides a strong trade-off between performance and computational efficiency.

Ablation results in Table 5 show that most components are critical for performance. In particular, removing the presence regression head, weighted sampling, or element constraints leads to notable degradation, while freezing the encoder in Stage 2 results in the largest drop, highlighting the importance of joint optimization across stages.

[3] https://github.com/unslothai/unsloth

5 Discussion

Stronger and better-optimized models consistently yield superior performance: Qwen3 achieves the highest Macro F1 and lowest RMSE, followed by RoBERTa, with BERT trailing. Results suggest that architecture and task design may play an important role, although further experiments are needed.

The ablation study indicates that performance is primarily driven by Stage-2 design and end-to-end optimization. Multi-task learning in Stage 1 improves representations (A1), while addressing class imbalance is critical for stability (A2). Hierarchical transfer from Stage 1 to Stage-2 is beneficial (A3). In Stage-2, the BIO span head and validity mask provide effective inductive biases (A4–A5). The encoder shows some robustness to input noise (A6), but freezing it leads to severe degradation (A7), highlighting the importance of full fine-tuning. Overall, representation learning, structural constraints, and fine-tuning are key drivers of performance.

6 Conclusion

We propose a relatively lightweight hierarchical framework for fine-grained self-state modeling designed for CPU-only and low-compute environments. Instead of relying on large-scale model exploration, the focus is on efficiency while maintaining competitive performance across different backbones, including BERT, RoBERTa, and a quantized Qwen3 model.

Experimental results show consistent gains across models. Qwen3 achieves the highest Macro-F1, although performance differences are small. While RoBERTa offers a strong balance between accuracy and efficiency, including an 8.3% improvement in Macro-F1 over the CLPsych 2026 baseline. Despite Qwen3 being much larger, its performance is close to RoBERTa, indicating that the hierarchical design effectively captures task structure without requiring large parameter growth.

The method leverages hierarchical multi-task learning to exploit shared representations across interdependent subtasks, improving efficiency and modularity. In terms of sustainability, encoder-based models are significantly more energy-efficient, with BERT and RoBERTa producing far lower carbon emissions than Qwen3, highlighting

the environmental benefits of lightweight architectures.

However, there remains a performance gap of about 0.11 compared to the top-ranked system. Future work will explore joint modeling of timelines and reducing error propagation in multi-stage pipelines to further improve accuracy and computational efficiency.

Limitations

The study has several limitations. First, the exploration of model space is limited to a small set of backbones, restricting generalizability to other architectures. Second, a notable performance gap to the state-of-the-art remains, suggesting that the proposed framework does not capture all relevant signals. Third, the ablation study is limited in scope and does not examine alternative architectural designs. Fourth, evaluation on a single benchmark raises concerns about domain-specific overfitting and limits claims of generalization. Finally, the lack of detailed interpretability analysis constrains understanding of component-level contributions, which is particularly important in clinical applications. Additionally, encoder models were trained with fixed hyperparameters without search; exploring alternative learning rates and batch sizes remains future work.

Ethics Statement

This study uses the CLPsych 2026 shared task dataset, an extension of the "Reddit-New" corpus from the CLPsych 2022 Shared Task (Tsakalidis et al., 2022): Capturing Moments of Change in Longitudinal User Posts, which contains sensitive, de-identified mental health–related content. We adhere to the shared task Data Access Agreement and follow established ethical guidelines such as Ethical Research Protocols for Social Media Health Research (Benton et al., 2017). The dataset is used solely for research purposes within the approved scope, stored securely, and will be deleted after the permitted period. We do not redistribute the data. We do not attempt to re-identify users, and all examples presented in this paper are paraphrased to protect privacy. In compliance with task rules, we use only open-source models (e.g., BERT, RoBERTa) and avoid proprietary systems such as ChatGPT, Claude, and Gemini.

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.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Giulia Ferraro and Luca Benedetti. 2025. Hierarchical multi-task learning for fine-grained and coarse text classification. *Frontiers in Interdisciplinary Applied Science*, 2(2):184–190.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder,

Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098.

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

Bing Tian, Yong Zhang, Jin Wang, and Chunxiao Xing. 2019. Hierarchical inter-attention network for document classification with multi-task learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3569–3575. International Joint Conferences on Artificial Intelligence Organization.

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, and 1 others. 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.

Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, 124:480–494.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendices

A.1 Task and Dataset

The dataset statistics are provided in Table 2. A large proportion of posts contain both adaptive and maladaptive states (approximately 47%), with maladaptive states rated as more psychologically central than adaptive ones (mean presence 3.23 vs. 2.39). Maladaptive self-states are most strongly expressed through negative Affect (A; 68%) and Self-Criticism (C-S; 57%), while adaptive states are predominantly characterised by relational Behaviour (B-O; 53%) and Desire for connection (D; 50%). Longitudinal change events — self-state switches (21%) and escalations (20%) — are distributed across 90% of training timelines, motivating temporal modelling as a core system component.

Statistic	Train	Test
Timelines / Posts	30 / 373	10/92
Annotated posts	236 (63%)	0 (0%)
Posts per timeline (mean/max)	12.43 / 25	9.2 / 11
Post length in words (mean)	114	115.46
Both states present	177 (47.4%)	NA
Adaptive state	203 (54%)	NA
Maladaptive state	210 (56%)	NA
Most freq. adaptive element	B-O	NA
Most freq. maladaptive element	A	NA

Table 2: Training and test data statistics

A.2 Instruction-Tuned LLM Pipeline with Transformer-based Decoder

We propose a multi-stage framework for structured psychological state prediction from social media posts as depicted in Figure 2. Given an input post x, the goal is to generate a hierarchical output y capturing (i) adaptive and maladaptive states, (ii) associated psychological elements, (iii) fine-grained categories with supporting evidence, and (iv) overall presence scores.

Instead of solving this as a single complex prediction problem, we decompose it into four sequential stages, each implemented using an instruction-tuned large language model with parameter-efficient fine-tuning.

We use an instruction-tuned causal language model fine-tuned with Low-Rank Adaptation (LoRA). All stages share the same base model, which is sequentially fine-tuned on stage-specific

instruction–response data. This allows the model to progressively specialize while maintaining a unified parameter space.

A.2.1 Step 1: State Detection

In the first stage, the model determines whether the post expresses adaptive and/or maladaptive states, and assigns a presence score (1–5) to each. The output is a structured JSON object containing binary labels and ordinal scores. Which is, $z_1 = (s_a, p_a, s_m, p_m) = f_1(x)$.

A.2.2 Step 2: Element Identification

Given the detected states, the model identifies which psychological elements are present. We consider a predefined set of elements $\{A, B\text{-}O, B\text{-}S, C\text{-}O, C\text{-}S, D\}$. The model outputs only those elements that are explicitly expressed in the post. Which is, $z_2 = \mathcal{E} = f_2(x, z_1)$.

A.2.3 Step 3: Category and Evidence Extraction

For each identified element, the model predicts a fine-grained subcategory and extracts the corresponding supporting text span from the post. The evidence is required to be an exact substring of the input, ensuring interpretability and traceability of predictions, i.e., $z_3 = \{(c_e, \tilde{x}_e) \mid e \in \mathcal{E}\} = f_3(x, z_2)$.

A.2.4 Step 4: Presence Estimation

Finally, the model predicts an overall presence score (1–5) for each detected state, conditioned on the elements and their associated evidence. This step refines the initial estimate from Step 1 using more granular information, that is, $z_4 = (p'_a, p'_m) = f_4(x, z_2, z_3)$.

Each stage is trained independently using supervised fine-tuning with instruction–response pairs constructed from annotated data. We use the standard causal language modeling objective, where the model learns to generate the target response given the input prompt.

The loss is computed only over response tokens, masking out the instruction part of the input. No task-specific loss functions are introduced.

A.2.5 Inference

At inference time, predictions are generated sequentially in a chained manner. The output of each stage is used to construct the prompt for the next stage. This decomposition simplifies the overall task and improves interpretability, as intermediate predictions (e.g., elements and evidence) are explicitly exposed. So the final output we receive from, $y = (z_1, z_2, z_3, z_4)$ where,

$$
\begin{aligned}
z_1 &= f_1(x) \\
z_2 &= f_2(x, z_1) \\
z_3 &= f_3(x, z_2) \\
z_4 &= f_4(x, z_2, z_3)
\end{aligned}
$$

The prompts used in four stages of Method 2 are given in Table A.2.

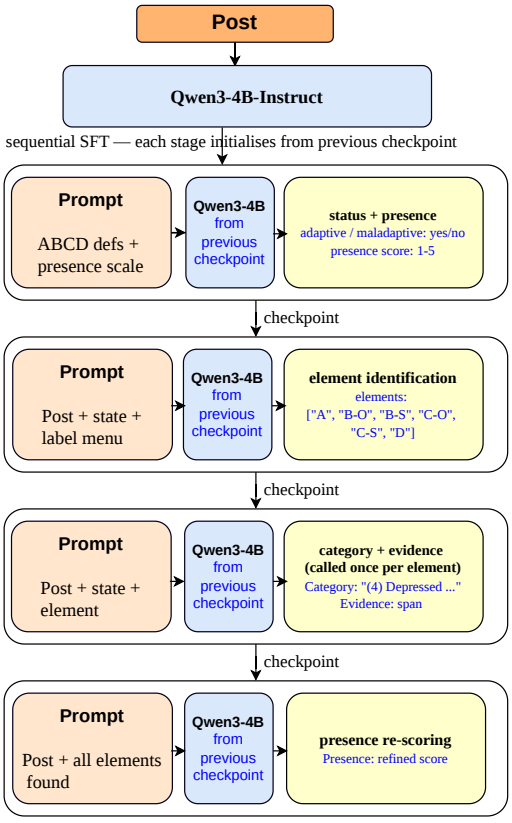

Figure 2: Four-stage approach using the Qwen3 model to solve the task.

A.3 Ablation study

The ablation study conducted only on the first method, which uses a controlled hold-out validation setup, where the training data is split into 85% train and 15% validation with a fixed seed, and all variants are trained and evaluated on the same splits for fair comparison. Stage 1 is trained on the full training set for binary classification and evaluated on the validation set, while Stage-2 is

Stage	Inputs injected	Prompt text
Stage 1	post — the raw social media text DEFINITIONS_BLOCK — full adaptive + maladaptive sub-element taxonomy PRESENCE_DEF — 1–5 scale definition	You are a clinical psychologist. Read this post and determine whether it expresses an ADAPTIVE state (conducive to fulfilling basic needs) and/or a MALADAPTIVE state (hindering basic needs), plus the presence score (1-5) for each. [DEFINITIONS_BLOCK injected here] Presence scale: 1=Not present … 5=Highly present. Post: """[post]""" Answer with valid JSON only (no prose): {"adaptive":"yes/no","adaptive_presence":1-5, "maladaptive":"yes/no","maladaptive_presence":1-5}
Stage 2	post — social media text state — "adaptive" or "maladaptive" sub-element list — dynamically selected for the given state	You are a clinical psychologist. Post: """[post]""" For the [ADAPTIVE/MALADAPTIVE] state, identify which ABCD elements are expressed. Sub-elements: A: (1) Calm/laid back, (3) Sad/Emotional pain … B-O: (1) Relating behavior … … [state-specific sub-elements injected] List only elements that are clearly present. Answer with valid JSON only: {"elements": ["A","B-O",...] or []}
Stage 3	post — social media text state — "adaptive" or "maladaptive" element — one of {A, B-O, B-S, C-O, C-S, D} options list — sub-labels valid for this element × state	You are a clinical psychologist. Post: """[post]""" For the [ADAPTIVE/MALADAPTIVE] state, element [e], choose the best sub-category from: (1) Self care and improvement (2) Self harm, neglect and avoidance … [element-specific options injected] Also copy the EXACT span from the post that best supports this. Answer with valid JSON only: {"Category":"<sub-category>", "highlighted_evidence":"<exact span>"}

Stage 4	post — social media text
	state — "adaptive" or "maladaptive"
	elements_info — JSON dict of {element → {Category, highlighted_evidence}} from Stage 3 output
	PRESENCE_DEF — 1–5 scale definition
		You are a clinical psychologist.
		
		Presence scale: 1=Not present … 5=Highly present.
		
		Post: """[post]"""
		
		The [ADAPTIVE/MALADAPTIVE] state was identified with these elements:
		{
		"A": {"Category": "(1) Calm/laid back",
		"highlighted_evidence": "..."},
		"C-S": {"Category": "(1) Self-acceptance…",
		"highlighted_evidence": "..."}
		}
		[elements_info JSON injected from Stage 3]
		
		Rate the overall PRESENCE of the [ADAPTIVE/MALADAPTIVE] state (1-5).
		
		Answer with valid JSON only:
		{"presence": <1-5>}

Table 3: Prompts used across the four stages in Method 2.

trained and evaluated only on positively labeled instances expanded into element-level records. Starting from the full model (A0) as the baseline, we introduce controlled modifications targeting: (i) Stage 1 learning design, including removal of the auxiliary regression head (A1) and disabling class-balanced sampling (A2), to assess the impact of multi-task supervision and imbalance handling on representation learning; (ii) cross-stage knowledge transfer, where Stage-2 is trained without initialization from Stage 1 (A3), to quantify the benefit of hierarchical pretraining; (iii) Stage-2 task formulation, by removing the token-level BIO evidence span head (A4), to examine the role of span supervision; (iv) structural constraints, by disabling element-specific validity masking over label space (A5), to evaluate the importance of inductive bias; (v) input preprocessing, by omitting text cleaning (A6), to measure robustness to noisy input; and (vi) fine-tuning strategy, by freezing the encoder during Stage-2 training (A7), to analyze the necessity of full model adaptation. Each variant alters exactly one component while keeping all other settings fixed, enabling precise attribution of performance differences to individual design choices.

The summary of results of ablation study is provided in Table 5.

A.4 Carbon Footprint

We measure the environmental impact of our models using CodeCarbon [4]. Training is conducted on NVIDIA A100 80GB PCIe, and emissions are estimated based on energy consumption and regional carbon intensity. Our results show that BERT and RoBERTa incur significantly lower emissions (0.018 kg and 0.032 kg CO_2 eq, respectively) compared to Qwen3 (0.111 kg), reflecting the efficiency advantages of encoder-based architectures. The result depicted in Table 4.

Model	Hardware	Training Time	CO_2 (kg)
BERT	A100	492.44	0.018
RoBERTa	A100	660.89	0.032
Qwen3	A100	2749.42	0.111

Table 4: Carbon emissions measured for different models using the CodeCarbon Python package.

A.5 Codabench Submission

We name our submission **NoviceTrio**. Submission ID 662787, submitted during the competition phase, was produced using Method 1 with BERT (bert-base-uncased). Submission ID 706280 in the analysis phase also used Method 1 but with RoBERTa (roberta-base), while Submission ID 705090 in the analysis phase used Method 1 with Qwen3 (Qwen3-4B-Instruct-2507). The code for training, preprocessing, post-processing, and ablation studies is publicly available on our GitHub repository.

[4] https://github.com/mlco2/codecarbon

DELTA vs BASELINE (A0)					
ID	Variant	S1-Adap	S1-Malad	S2-Adap-Cat	S2-Malad-Cat
A1	No presence regression head	-0.01	-0.0404	-0.1137	-0.0137
A2	No weighted sampling	0.0072	-0.1027	-0.0444	0.0648
A3	No Stage 1 pre-training for Stage-2	0	0	-0.0789	0.0142
A4	No evidence span head (Stage-2)	0.0171	0	-0.034	-0.0452
A5	No element validity mask (Stage-2)	0.0171	-0.0177	-0.109	-0.0156
A6	No text cleaning	0.0171	-0.0012	-0.0714	0.0167
A7	Frozen encoder in Stage-2	0	0	-0.3517	-0.1322

Table 5: Change (delta) in performance compared to the full model or baseline (A0).