From Evidence Mining to Meta-Prediction: a Gradient of Methodologies for Task-Specific Challenges in Psychological Assessment

Federico Ravenda[1], **Fawzia-Zehra Kara-Isitt**[2],
Stephen Swift[2], **Antonietta Mira**[1,3], **Andrea Raballo**[1,4]
federico.ravenda@usi.ch, fuzzy.kara-isitt@brunel.ac.uk ,
stephen.swift@brunel.ac.uk, antonietta.mira@usi.ch, andrea.raballo@usi.ch
[1]Università della Svizzera italiana, [2]Brunel University,
[3]Insubria University, [4]Cantonal Sociopsychiatric Organisation

Abstract

Large Language Models are increasingly used in the medical field, particularly in psychiatry where language plays a fundamental role in diagnosis. This study explores the use of open-source LLMs within the MIND framework. Specifically, we implemented a mixed-methods approach for the CLPsych 2025 shared task: **(1)** we used a combination of retrieval and few-shot learning approaches to highlight evidence of mental states within the text and to generate comprehensive summaries for post-level and timeline-level analysis, allowing for effective tracking of psychological state fluctuations over time **(2)** we developed different types of ensemble methods for well-being score prediction, combining Machine Learning and Optimization approaches on top of zero-shot LLMs predictions. Notably, for the latter task, our approach demonstrated the best performance within the competition[1].

1 Introduction

Recent advancements in NLP have enabled the development of new and complex models across various areas, particularly in digital and mental health. Transformer-based models (Vaswani et al., 2017) have significantly advanced mental health analysis on social media platforms. While models initially primarily used BERT-based architectures fine-tuned in supervised contexts to predict the presence of symptoms related to mental disorders (Yang et al., 2021; Bucur et al., 2021), recently the use of LLMs in psychology has proven promising (Ravenda et al., 2025; De Grandi et al., 2024; Varadarajan et al., 2024). The advantage of LLMs is that they can be employed even in contexts with limited or absent training data, leveraging their capabilities as few- or zero-shot models.

The CLPsych 2025 (Tseriotou et al., 2025) shared task addresses the significant challenge of generating supporting evidence and predicting well-being for clinical assessments, with a specific focus on well-being assessment. The shared task builds upon the foundation established by CLPsych 2019 and 2022 (Shing et al., 2018; Zirikly et al., 2019; Tsakalidis et al., 2022). In particular, the 2025 competition extends the 2022 work by incorporating evidence generation (Chim et al., 2024), thereby promoting the development of humanly interpretable rationales for recognizing dynamic mental states. The task employs the MIND framework (Slonim, 2024), a pan-theoretical paradigm that conceptualizes human experience as fluctuating self-states rather than static conditions. Self-states are defined as identifiable units characterized by specific combinations of Affect, Behaviour, Cognition, and Desire/Need (ABCD) (Revelle, 2007) that coactivate meaningfully for limited periods (Lazarus and Rafaeli, 2023).

The shared task comprises four primary components: (1) Task **A.1** focuses on post-level judgments, requiring participants to identify evidence of adaptive and maladaptive self-states, while Task **A.2** rate overall well-being using the Global Assessment of Functioning (GAF) scale (American Psychiatric Association et al., 1994) associated to each post; Task **B** involves generating post-level summaries of self-state dynamics, identifying dominant states and their central organizing aspects, while Task **C** requires timeline-level summaries capturing temporal dynamics between self-states.

The main contributions of this work are:

(1.) We designed a comprehensive approach for predicting well-being scores from Reddit posts. The final prediction, constructed from the predictions of various open sources LLMs, is generated by a tool we call *"aggregator"*, which can be implemented as a simple average ensemble, a machine learning meta-model (which we call "*Oracle*"), or a weighted average of predictions from different LLMs where the weights are mathematically opti-

[1] Code available at the following link: https://github.com/Fede-stack/BULUSI-CLPsych

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 242–248
May 3, 2025 ©2025 Association for Computational Linguistics

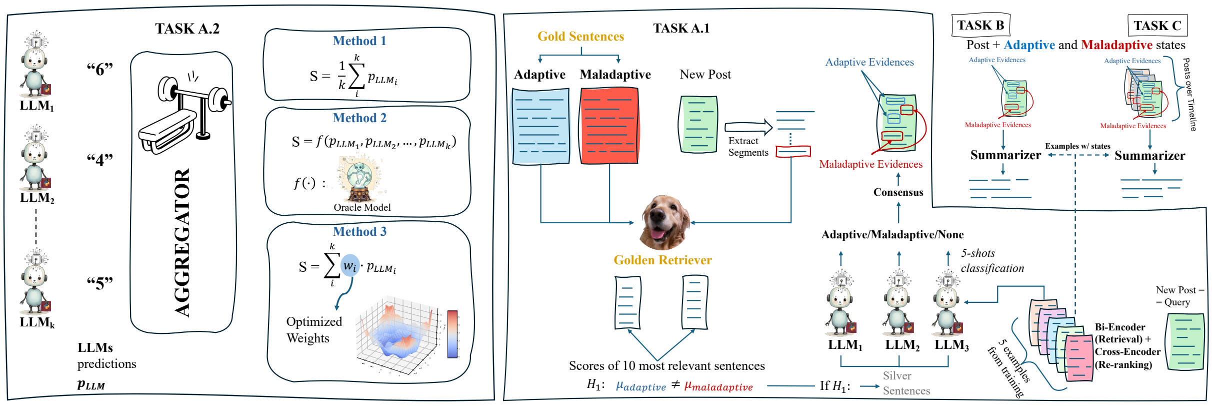

Figure 1: Multi-component pipeline for mental health state detection and assessment in the CLPsych 2025 Shared Task, illustrating our approach to evidence extraction (Task **A.1**), well-being score prediction with three aggregation methods (Task **A.2**), and summarization at post and timeline levels (Tasks **B** & **C**).

mized to minimize Mean Squared Error (MSE).
(2.) We implemented an approach to highlight evidence of adaptive and maladaptive states, using a multi-step procedure based on an initial retrieval stage that identifies potential segments within posts where adaptive and maladaptive states emerge, followed by a second stage where LLMs supervise and classify these candidate segments, ensuring accurate identification of psychological states while reducing false positives through consensus-based validation. These evidence are then used to generate comprehensive summaries for post-level and timeline-level analysis, allowing for effective tracking of psychological state fluctuations over time.

2 Methods

2.1 Predict Well-Being Score

Task **A.2** consists in assigning a well-being score to each post within user timelines. For each user, we have a chronological sequence of Reddit posts, and our goal is to rate each post's overall well-being on a scale from 1 (low well-being) to 10 (high well-being) based on the GAF framework.

For this Task, we employed an ensemble approach using six open-source LLMs in a zero-shot setting to predict well-being scores. In particular, we used the following models: `gemma-2-9b`, `qwen-2.5-72b`, `deepseek-V3`, `phi-4`, `mixtral-8x22b`, `llama-3.3-70b`. Additionally, our prompt instructions explicitly directed the models to return null (NaN) values when insufficient evidence was available to make a confident assessment. This approach allowed us to optimize predictions based on training data by tuning few meta-model's parameters rather than updating the large number of LLMs parameters, creating a solution that remains scalable and efficient.

We explored three distinct aggregation strategies (for a visual interpretation see Figure 1) to calibrate the importance of the different LLMs to predict the final score. For each Reddit post RP, we have a vector of predictions p_{LLMs} of dimension six: $p_{LLMs} = (p_{LLM_1}, \dots, p_{LLM_6})$.

Simple Ensemble (submission_1): For each post, the final score S is computed as the rounded average of predictions from all LLMs, $S = \frac{1}{k}\sum_{i=1}^{k} p_{LLM_i}$, where $k = 6$

Meta-Learning (submission_2): We trained a LightGBM Regressor (Ke et al., 2017) that uses LLM predictions as features to calibrate final scores. This meta-model learns the relationship between model outputs and ground truth on training data, effectively functioning as a sort of stacking ensemble or, as we called it, as an "*Oracle Model*". We chose LightGBM as it is able to handle missing values by default. The final score is calculated as: $S = f(p_{LLM_1}, \dots, p_{LLM_6})$, where $f(\cdot) = LightGBM$ with default parameters as in LightGBM python package.

Optimized Weighting (submission_3): We mathematically optimized model weights by minimizing mean squared error between the weighted sum of predictions and ground truth. The optimization procedure handles NaN values through dynamic weight renormalization and enforces non-negative weights that sum to 1:

$$\min_{\mathbf{w}} \quad \text{MSE}(\mathbf{y}, \hat{\mathbf{y}}(\mathbf{w})), \text{ s.t.} \sum_{i=1}^{M} w_i = 1, \; w_i \geq 0 \; \forall i$$

The weighted prediction for each post with NaN handling:

$$S = \frac{\sum_{i=1}^{M} w_i \cdot p_{LLM_i} \cdot I(p_{LLM_i} \neq \text{NaN})}{\sum_{i=1}^{M} w_i \cdot I(p_{LLM_i} \neq \text{NaN})}$$

where p_{LLM_i} is model i's prediction for the specific post and $I(\cdot)$ is the indicator function.

For posts where most LLMs returned NaN (indicating insufficient evidence), we defaulted to scores of 7 based on empirical patterns observed in training data.

2.2 Evidence and Summarization

In this subsection we summarise the methods of the other three tasks, also shown in Figure 1.

For Task **A.1**, we implemented a multi-stage pipeline called *Golden-Retrieval Augmented Generation* (G-RAG). Each post was systematically segmented based on punctuation markers, specifically periods ('.') and the conjunction '*but*', which typically signal natural breaks in thought patterns.

Our initial phase employed a retrieval-based approach to identify relevant segments within the test set posts by comparing them against the training data evidence. This process involved calculating embedding distances between each segment and the available evidence (*gold sentences*). For every segment, we identified the 10 most relevant pieces of evidence (those with the highest embedding cosine similarity) and subsequently filtered for segments exhibiting significant differential distances between adaptive and maladaptive states evidence.

These filtered segments served as "*proposals*" for our pipeline. We then employed three different open-source Large Language Models (LLMs) (`qwen-2.5-72b`, `mixtral-8x22b`, `llama-3.3-70b`) in inference mode to classify whether each evidence segment corresponded to an adaptive state, a maladaptive state, or neither category. To enhance contextual understanding, we augmented the LLMs' input with five examples from the training set — specifically selecting posts most similar to the target post along with their corresponding annotated evidence. To resolve classification discrepancies among the three primary LLMs, we used a fourth open-source LLM (`llama-3.1-405b`) as an "*arbitrator*", but only in cases where no majority consensus was reached.

It is important to note that, due to time constraints, we made only one submission and did not optimize the results with respect to the evaluation metrics considered. We observed that using LLMs without the retrieval step resulted in highlighting an excessive number of sentences. Therefore, the retrieval stage was implemented specifically to mitigate this behavior. In general, we focused more on being strict and conservative regarding the number of sentences to highlight, paying particular attention to not include sentences that were neither adaptive nor maladaptive. This conservative approach was further justified by the fact that for Tasks **B** and **C**, we used these identified evidence segments to generate summaries at both post and timeline levels. Including excessive or inaccurate states would have negatively impacted the quality of these summaries, potentially introducing noise and reducing the coherence of the generation.

For a detailed discussion on retrieval models used for the tasks discussed we refer to Section A in Appendix.

3 Experiments

3.1 Metrics

The CLPsych 2025 shared task evaluation used specific metrics for each subtask (we refer to (Tseriotou et al., 2025) for an in depth-explanation):

Task A.1 (Evidence Identification): Semantic overlap between submitted and expert-annotated evidence was evaluated using recall (via maximum recall-oriented BERTScore) and weighted recall (adjusting for evidence length differences), with separate measurements for adaptive and maladaptive spans.

Task A.2 (Well-being Score Prediction): The main metric is Mean Squared Error (MSE) over all posts in a timeline, averaged across all timelines. Additional MSE calculations is performed for specific score ranges: posts indicating serious impairment (1-4), impaired functioning (5-6), and minimal impairment (7-10), providing insight into performance across different well-being levels. F1 macro at post level is also measured.

Task B (Post-level Summaries): This task evaluates consistency with expert-written summaries using Natural Language Inference models. Two metrics are used: mean consistency (measuring the absence of contradiction between submitted and expert summaries) and maximum contradiction (evaluating the worst-case contradictions between predicted and gold summaries).

Task C (Timeline-level Summaries): The same

TASK A.1						
Approach	Recall			Weighted Recall		
	overall	*adaptive*	*maladaptive*	*overall*	*adaptive*	*maladaptive*
G-RAG	0.433	0.339	0.526	0.37	0.339	0.402

TASK A.2					
Approach	MSE (↓)				F1
	overall	*min. impairment*	*impaired*	*ser. impairment*	*macro*
Average Ensemble	2.13	1.19	1.1	3.65	0.416
Oracle-Meta	2.12	0.55	0.82	3.98	0.365
Optimized Ensemble	1.92	0.65	1.19	3.04	0.351

TASK B			
Approach	gold summary		evidence
	mean consistency	*max contradiction* (↓)	*max entailment*
`qwen-2.5-72b`	0.868	0.805	0.808
`mixtral-8x22b`	0.822	0.880	0.562
`llama-3.1-405b`	0.845	0.768	0.553

TASK C		
Approach	gold summary	
	mean consistency	*max contradiction (↓)*
`qwen-2.5-72b`	0.890	0.898
`mixtral-8x22b`	906	0.992
`llama-3.1-405b`	0.941	0.714

Table 1: Results of our approaches w.r.t. all metrics considered in the shared task, conditioned on the four different tasks. Metrics highlighted in blue indicate the best result for that specific metric in the competition considering all the submissions from all the teams. The symbol (↓) indicate metrics for which a lower value is preferable.

consistency metrics as Task B is applied to evaluate how well the system-generated timeline summaries aligned with expert timeline analyses, focusing on capturing the temporal dynamics of mental states.

3.2 Results

Our BULUSI team's approach showed promising results across different CLPsych 2025 shared tasks, with particularly strong performance in Task **A.2**. Table 1 presents our results across all metrics for the four different tasks. The prompts used for the LLMs across different Tasks are reported in in the Github Repository: https://github.com/Fede-stack/BULUSI-CLPsych.

For Task **A.1**, our G-RAG approach achieved a recall of 0.433 overall, with stronger performance on maladaptive state evidence (0.526) compared to adaptive state evidence (0.339). The weighted recall metrics showed similar patterns, with an overall weighted recall of 0.37. For this task, we submitted only one solution that was implemented to be highly conservative in identifying evidence, without optimizing the evaluation metric. This conservative approach is reflected in the fact that the scores for adaptive recall and adaptive weighted recall remain the same, demonstrating our cautious strategy to include only strong evidence and avoid false positives in our evidence list.

In Task **A.2**, we compared three different aggregation strategies. The Optimized Ensemble method demonstrated the best performance with an overall MSE of 1.92, outperforming both the Average Ensemble (2.13) and Oracle approaches (2.12), as well as being the best result within the competition. Additionally, the F1 macro score for the average ensemble approach (0.416), and the minimal impairment metric for the Oracle model (0.55) achieved the best performance within the competition. Figure 2 shows the distribution of scores across all submissions and metrics in the competition for Task **A.2**, highlighting the scores obtained by our three approaches. We observe that for all metrics, our scores often represent either the best results or fall within the top results.

For Task **B**, we tested three different LLMs to generate summaries. The *qwen-2.5-72b* model showed the best performance with a mean consistency of 0.868 and max entailment of 0.808. This latter result is the highest for this specific metric within the competition. On the other hand, for Task **C**, the `llama-3.1-405b` model performed well with a mean consistency of 0.941 (close to the best performance within the competition of 0.946), demonstrating that our approach can effectively captured the temporal dynamics of mental states across user timelines.

We observe that in the last two tasks, despite

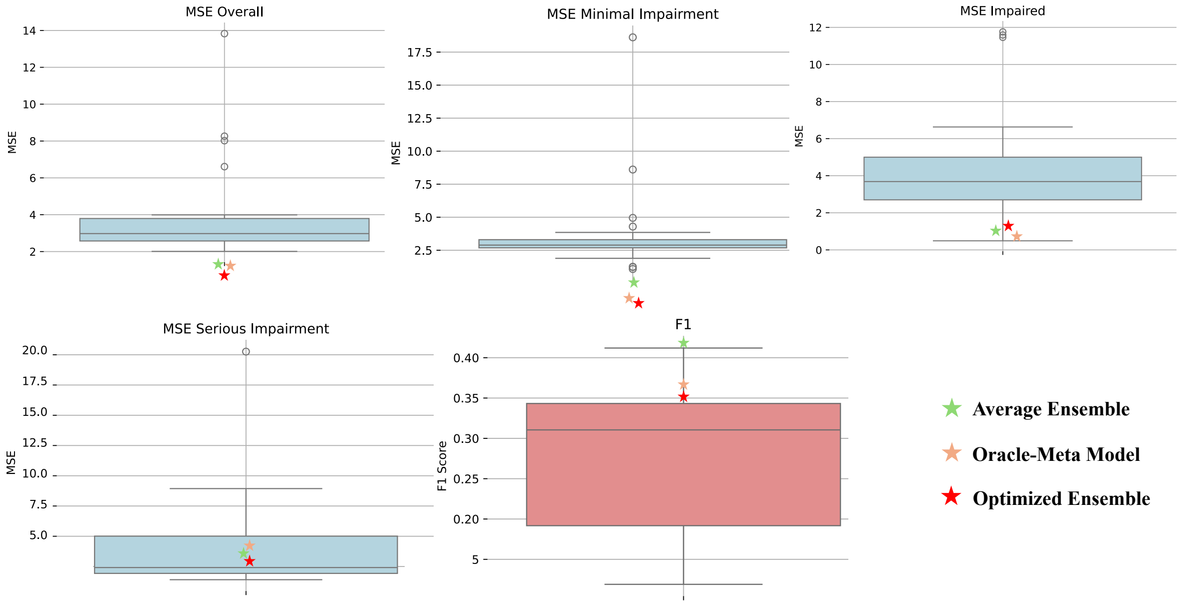

Figure 2: The boxplots show the score distribution of all competition submissions for Task **A.2** metrics. Colored stars ⋆ indicate the performance of three ensemble approaches (Average, Oracle-Meta, and Optimized), showing how they compare to the overall distribution across different impairment severity levels using MSE and F1 metrics.

using identical prompts, performances vary considerably between different LLMs. This variation may be attributed to the type of metrics used for evaluating summary quality. Specifically, the evaluation relies on Natural Language Inference (NLI) models based on BERT architectures to measure consistency and contradiction. Different LLMs may produce summaries that align differently with how these NLI evaluation models conceptualize contradictions and entailments. This suggests that the performance variations could stem not only from differences in the LLMs' generation capabilities but also from their alignment with the specific linguistic patterns that the BERT-based evaluation models were trained to recognize.

4 Conclusions and Future Works

Our multi-stage pipeline implemented for the CLPsych 2025 shared task combined retrieval-augmented evidence identification with ensemble methods for well-being prediction, achieving top performance in the competition. Our approach effectively demonstrates the potential of open-source LLMs in psychological assessment. While challenges remain in detecting severe impairment cases, this work establishes a promising foundation for computational tools that could support mental health monitoring through social media analysis.

Future work could focus on improving the evidence mining task. In studies like (Ravenda et al., 2025; Pérez et al., 2022), the initial step for predicting specific symptom scores within psychological questionnaires, based on Reddit posts, involves retrieving the most relevant posts for each questionnaire item. Task **A.1** could therefore be extended to retrieve evidence related to specific symptoms of various psychological conditions, while the LLM ensemble approach from Task **A.2** could be leveraged to enhance prediction accuracy.

5 Limitations

Our approach faces several limitations that should be considered when interpreting the results.

First, our retrieval-augmented approach for evidence identification depends on the quality and coverage of the training dataset. If the training data lacks representation of certain mental state expressions or cultural contexts, our system may fail to identify relevant evidence in these cases.

Second, while our ensemble approach for well-being score prediction demonstrated the best performance in the task, it still struggles with accurately assessing posts indicating serious impairment

Third, a limitation of this work is the relatively small number of users. Therefore, there is no guarantee that similar results will be replicated across new data.

Finally, our implementation faced time constraints that limited optimization efforts, particularly for Tasks **A.1**, **B** and **C**. With additional time, we could have explored more sophisticated approaches for generating summaries and potentially improved performance across all tasks.

6 Ethics Considerations

Mental health assessments derived from computational models should never replace professional clinical judgment. Our system is designed as a supportive and screening tool that can assist mental health professionals rather than as an autonomous diagnostic system. The well-being scores and identified evidence of mental states should be considered as preliminary insights that require professional validation.

Additionally, there is potential for algorithmic bias in mental health assessment systems. Language models may perpetuate biases present in their training data, potentially leading to disparities in assessment quality across different demographic groups (Basta et al., 2019). We acknowledge this limitation and emphasize the importance of ongoing evaluation for fairness and bias mitigation.

Acknowledgements. This research has partly been founded by the SNSF (Swiss National Science Foundation) grant 200557.

References

A American Psychiatric Association, American Psychiatric Association, et al. 1994. *Diagnostic and statistical manual of mental disorders: DSM-IV*, volume 4. American psychiatric association Washington, DC.

Christine Basta, Marta R Costa-Jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*.

Ana-Maria Bucur, Adrian Cosma, and Liviu P Dinu. 2021. Early risk detection of pathological gambling, self-harm and depression using bert. *arXiv preprint arXiv:2106.16175*.

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.

Alessandro De Grandi, Federico Ravenda, Andrea Raballo, and Fabio Crestani. 2024. The emotional spectrum of llms: Leveraging empathy and emotion-based markers for mental health support. *arXiv preprint arXiv:2412.20068*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

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

Anxo Pérez, Neha Warikoo, Kexin Wang, Javier Parapar, and Iryna Gurevych. 2022. Semantic similarity models for depression severity estimation. *arXiv preprint arXiv:2211.07624*.

Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, Antonietta Mira, and Noriko Kando. 2025. Are llms effective psychological assessors? leveraging adaptive rag for interpretable mental health screening through psychometric practice. *arXiv preprint arXiv:2501.00982*.

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

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.

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.

Vasudha Varadarajan, Allison Lahnala, Adithya V Ganesan, Gourab Dey, Siddharth Mangalik, Ana-Maria Bucur, Nikita Soni, Rajath Rao, Kevin Lanning, Isabella Vallejo, et al. 2024. Archetypes and entropy: theory-driven extraction of evidence for suicide risk. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 278–291.

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.

Feifan Yang, Tao Yang, Xiaojun Quan, and Qinliang Su. 2021. Learning to answer psychological questionnaire for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1131–1142.

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 Retrieval Approaches

Regarding section 2.2, various retrieval approaches were used. For proposing text segments as evidence of adaptive and maladaptive states, the following dense retrieval models were used as golden retrievers (see Figure 1): `msmarco-distilbert-base-v4`, `msmarco-MiniLM-L12-cos-v5`, and `GIST-large-Embedding-v0`. Each of these returned a list of evidence, obtained as described in Section 2.2, and the union of all unique evidence items was then taken as “*proposals*”.

For Task **B** and **C** in Section 2.2, when processing a new test post, we identified the 5 most similar posts from the training set to provide examples that would assist the LLM in generation. To select these 5 most relevant posts, we first retrieved the 50 most relevant posts using a zero-shot retrieval approach, `contriever`, and then obtained the 5 most similar posts from this initial set by using a re-ranking model, specifically `ms-marco-MiniLM-L-6-v2`. The retrieved posts were added to the LLMs prompts as examples to follow.