A Hypothesis-Driven NLI Framework for Clinical Mental Health Assessment in Conversational Settings

Zafar Sarif[1,2], Abhishek Das[1] and Dipankar Das[2]

[1]Aliah University, Kolkata, India
[2]Jadavpur University, Kolkata, India

Abstract

To conduct a mental health assessment through Natural Language Processing (NLP) therapy cues, different psychological constructs through therapy dialogues are necessary, such as PHQ-9, GAD-7, and CompACT-10, among other psychological modalities. This work proposes a hypothesis-driven Natural Language Inference (NLI) framework for automatically inferring item level severity scores and selecting appropriate therapist responses from multi-turn therapeutic dialogues. This methodology re-contextualizes both tasks in the field of semantic inference. Conversational contexts are paired with clinically grounded hypotheses and processed using a pretrained multilingual NLI model. The framework operates entirely in a zero-shot setting without task-specific training. Experimental evaluation demonstrates a stronger performance for PHQ-9 (Mean Absolute Error or MAE: 1.477) compared to CompACT-10 (MAE: 1.637) and GAD-7 (MAE: 2.103). We achieved Mean Zero-One Error (MZOE) score of 0.802 for CompACT-10, 0.813 for PHQ-9, and 0.948 for GAD-7. For therapist's response selection, the proposed framework achieved an overall accuracy of 0.3556. The findings demonstrate that hypothesis-driven NLI provides an interpretable and flexible framework for conversational mental health assessment and therapeutic dialogue reasoning.

Keywords

Mental Health Assessment, Conversational Analysis, Natural Language Inference, PHQ-9 and GAD-7

1. Introduction

Learning about mental health from conversational data by mental health professionals is an established process in the field of natural language processing. This process aims at designing a computer-aided support systems to help detect advanced psychological distress. It is different from regular sentiment analysis or emotion detection systems, as predictive models must create delineated and structured, though qualitative, clinical predictions, such as PHQ-9, GAD-7, and CompACT-10. One of the greatest difficulties is that modelers must quantify implicit and context-sensitive psychological hints and cues, interspersed in a greater number of rounds of conversations between the patient and the therapist, and provide interpretable assessment estimates, along with justification for the levels of severity assessed. Another major difficulty in this work is to deal with Spanish texts, it lacks better models in comparison with the English language.

Addressing Psychological Needs of Conversational Agents (APNCA) in combination with a hypothesis-driven NLI framework allows us to define the therapist response, as well as mental health assessment from a different semantic perspective. We model the therapist response selection challenge, and then mental health assessment selection as a combination of construct definitions. We offer a convenient method to affect, and then score, the level of severity of different psychological constructs present in the context of a piece of latent therapy dialogue.

We evaluate our approach in the context of a shared task involving multi-round conversational data, where systems are required to produce predictions at each turn. Experimental results demonstrate that the proposed method is capable of capturing depression-related and psychological flexibility signals with reasonable effectiveness, while highlighting limitations in modeling anxiety severity. These findings underscore both the potential and the challenges of using NLI-based reasoning for structured mental health evaluation.

IberLEF 2026, September 2026, León, Spain
zsarifau@gmail.com (Z. Sarif); adasus@gmail.com (A. Das); dipankardipnil2005@gmail.com (D. Das)

The remainder of the paper is organized as follows. Section 2 describes the task setup and dataset. Section 3 reviews related work on mental health detection and inference-based modeling. Section 4 presents the proposed methodology. Section 5 outlines the experimental setup, results achieved, and analysis with respect to evaluation metrics and ranking. Finally, Section 6 concludes the paper and discusses future directions.

2. Task Description

This work is conducted within the MentalRiskES shared task at IberLEF 2026, which focuses on the early detection of mental health risk from conversational data [1, 2] in Spanish language. The task involves creating a realistic context for an advanced therapeutic session, where systems must deconstruct and gradually create and implement clinical predictions from patient-therapist interactions. The organizers offer data in a multi-round format, where each dialogue builds gradually within a series of interaction rounds. In each round, participants receive the up-to-date conversation history via an evaluation server, as opposed to a static dataset. The system must evaluate and forecast within the existing context of dialogue before the next round. This enforces a streaming evaluation protocol by not providing predictive information, while creating and maintaining a design within realistic restrictions.

The evaluation framework is server-based, as participants interact through an API to both receive test data and provide their predictions. As each round is contended, the system uses an authentication token to query the server and retrieves the state of dialogue within the given context. Predictive submissions must conform to a specified JSON format. Following submission, the server is prompted to provide the next submissions within a dialogue round series. In addition to the input and output format, predictive submissions, and time restrictions, the server operates in order to create and enforce clear and unambiguous evaluations of the systems.

The shared task consists of two primary objectives.

- `Task1`- The first task involves predicting item-level scores for three clinically validated instruments: PHQ-9 (depression), GAD-7 (anxiety), and CompACT-10 (psychological flexibility). Each instrument comprises multiple items, and each item is associated with an ordinal response scale (e.g., 0–3 for PHQ-9 and GAD-7, and 0–6 for CompACT-10). Systems must generate predictions for each item at every dialogue turn, reflecting the inferred severity or presence of specific psychological symptoms;
- `Task2`- The second task focuses on response selection. Given the dialogue context and a set of candidate therapist responses provided by the organizers, the system must select the most appropriate response. This task evaluates the model's ability to maintain conversational coherence and produce contextually relevant interventions.;

A significant hurdle inherent in this task is its incremental multi-turn format, wherein significant clinical insights are implicitly context-specific and allocated over multiple turns of the conversation. As a result, systems are required to do progressive, context-specific reasoning and update their analysis with new turns of dialogue.

3. Related Work

The MentalRiskES task described in the previous section emphasizes the need for models that can extract clinically meaningful information from multi-turn conversational data while producing structured predictions aligned with standardized assessment instruments. Recent advances in NLP have significantly improved automated mental health assessment from conversational and textual data. The increasing availability of dialogue-based datasets and transformer-based architectures has enabled the development of systems capable of detecting depression, anxiety, suicidal ideation, and psychological distress from natural language interactions [3, 4].

Early approaches in computational mental health primarily relied on traditional machine learning techniques using handcrafted lexical, syntactic, and psycholinguistic features [5]. Although these methods demonstrated the relationship between linguistic behavior and psychological conditions, they were limited in modeling contextual dependencies and semantic variation across conversations. The emergence of transformer-based language models such as BERT [6] and XLM-RoBERTa [7] substantially improved contextual representation learning for mental health applications. Recent studies have explored the use of transformer architectures for depression severity estimation, anxiety detection, and conversational mental health screening [8, 9]. Transformer-based conversational systems have also shown promising results in detecting clinically relevant cues from free-text interactions while maintaining strong psychometric reliability [10, 11].

Several recent works have focused specifically on conversational mental health assessment. Imel et al. [4] demonstrated that transformer-based models can effectively analyze counseling conversations to identify therapeutic patterns and behavioral signals. Similarly, conversational agent frameworks for mental health screening have incorporated PHQ-9 and GAD-7 questionnaires into interactive dialogue systems [11]. Systematic reviews further indicate that conversational agents are becoming increasingly important for scalable and accessible mental health support [12, 13]. Recent research has also explored adaptive and inference-driven assessment strategies. The ALBA framework [14] introduced adaptive language-based assessment techniques for psychological trait estimation using conversational responses. In parallel, large language models (LLMs) are increasingly being investigated for dialogue-based depression screening and psychological distress assessment [15, 16]. These studies highlight the growing importance of reasoning-based approaches capable of capturing subtle contextual and emotional patterns beyond conventional classification frameworks.

NLI has emerged as an effective paradigm for semantic reasoning, where a model determines whether a hypothesis is entailed by a premise. Recent work has shown that reformulating classification problems as inference tasks enables strong zero-shot generalization and improved interpretability [17]. However, the application of NLI to structured mental health assessment in multi-turn conversational settings remains relatively underexplored.

4. Methodology

We propose a hypothesis-driven framework that reformulates both clinical scoring and response selection as NLI tasks. An overview of the system is illustrated in Figure 1. The pipeline consists of four main stages: dialogue context construction, hypothesis formulation, entailment scoring, and prediction aggregation.

4.1. Dialogue Context Construction

At each round t, the system receives a dialogue consisting of sequential patient–therapist interactions. The context $x_{1:t}$ is constructed by concatenating the most recent utterances:

$$x = \text{concat}(u_{t-1}, u_t) \tag{1}$$

To ensure compatibility with the NLI model, the context is truncated to a fixed maximum length.

4.2. Clinical Hypothesis Formulation

To enable inference-based reasoning, each questionnaire item is represented as a natural language hypothesis. All hypotheses are expressed in a consistent declarative form (“The person ...”), allowing the NLI model to evaluate whether the dialogue context entails the presence of a specific symptom or behavioral pattern. The hypotheses used for different tasks are clearly mentioned below.

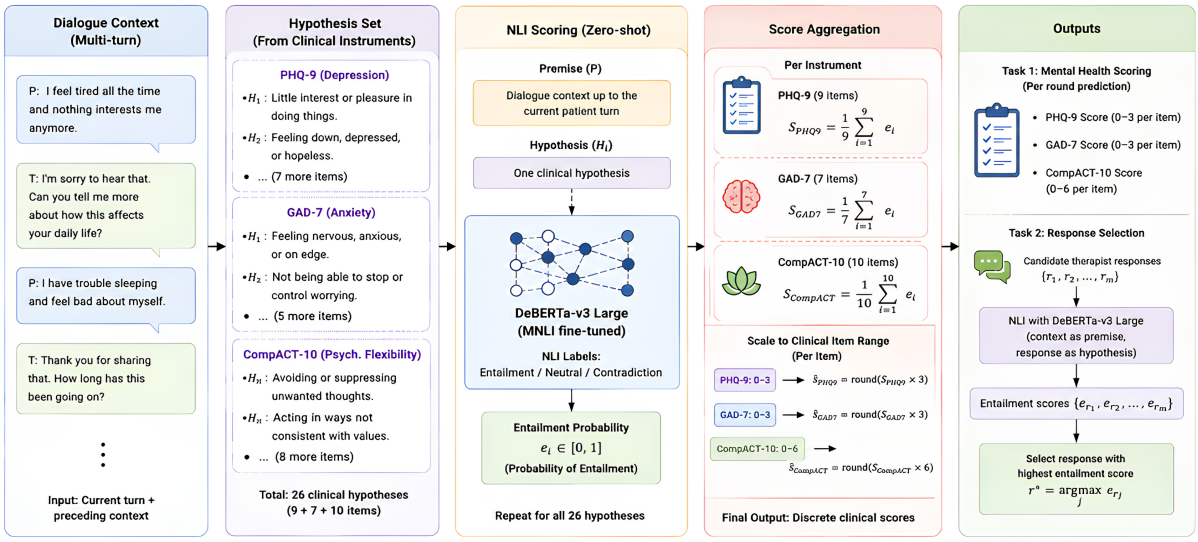

Figure 1: Overview of the proposed hypothesis-driven NLI framework

PHQ-9 Hypotheses (Depression, scale 0–3):

- The person has little interest or pleasure in doing things.
- The person feels down, depressed, or hopeless.
- The person has trouble falling or staying asleep, or sleeps too much.
- The person feels tired or has little energy.
- The person has poor appetite or overeats.
- The person feels bad about themselves or feels like a failure.
- The person has trouble concentrating on things.
- The person moves or speaks slowly, or is unusually restless.
- The person has thoughts of self-harm or that they would be better off dead.

GAD-7 Hypotheses (Anxiety, scale 0–3):

- The person feels nervous, anxious, or on edge.
- The person is unable to stop or control worrying.
- The person worries excessively about different things.
- The person has trouble relaxing.
- The person is so restless that it is hard to sit still.
- The person becomes easily annoyed or irritable.
- The person feels afraid that something awful might happen.

CompACT-10 Hypotheses (Psychological Flexibility, scale 0–6):

- The person is open to experiencing difficult thoughts and feelings.
- The person accepts unpleasant emotions without trying to avoid them.
- The person remains aware of the present moment.
- The person can distance themselves from distressing thoughts.
- The person adapts behavior according to the situation.
- The person takes actions aligned with personal values.
- The person allows emotions to come and go without resistance.
- The person avoids or suppresses negative experiences. (reverse-oriented)

- The person remains committed to meaningful actions.
- The person demonstrates psychological flexibility in daily life.

For Task 1, each clinical item is represented as a hypothesis h_i. The dialogue context x is paired with each hypothesis to form premise–hypothesis pairs: (x, h_i). For Task 2, each candidate therapist response r_j is treated as a hypothesis, forming pairs: (x, r_j)

4.3. NLI-Based Inference

We employ the pretrained multilingual NLI model `joeddav/xlm-roberta-base-xnli`, implemented via the Hugging Face Transformers library. The model is used through a text-classification pipeline, which outputs probabilities for the labels: {entailment, neutral, contradiction}.

For each input pair, the model produces a probability distribution:

$$f_\theta(x, h) \rightarrow \{p_e, p_n, p_c\} \tag{2}$$

where p_e corresponds to the entailment probability. We extract:

$$e = p_e \tag{3}$$

as the primary signal indicating whether the hypothesis is supported by the dialogue.

4.4. Entailment Scoring

For each pair (x, h_i), a pretrained NLI model f_θ computes the probability of entailment:

$$e_i = P(y = \text{entailment} \mid x, h_i) \tag{4}$$

where $e_i \in [0, 1]$ reflects the likelihood that the symptom described by h_i is supported by the dialogue context.

4.5. Intensity Mapping

The continuous entailment scores are mapped to discrete ordinal scales using a threshold function $g(\cdot)$:

$$\hat{y}_i = g(e_i) \tag{5}$$

For PHQ-9 and GAD-7 items, predictions are mapped to a 4-point scale:

$$\hat{y}_i \in \{0, 1, 2, 3\} \tag{6}$$

representing increasing severity levels.

For CompACT-10 items, predictions are mapped to a 7-point scale:

$$\hat{y}_i \in \{0, 1, 2, 3, 4, 5, 6\} \tag{7}$$

capturing degrees of psychological flexibility.

4.6. Score Aggregation

Instrument-level scores are obtained by summing item-level predictions:

$$S = \sum_{i=1}^{N} \hat{y}_i \tag{8}$$

where N is the number of items (9 for PHQ-9, 7 for GAD-7, and 10 for CompACT-10).

4.7. Response Selection

For Task 2, each candidate response r_j is evaluated using the same NLI model:

$$s_j = P(y = \text{entailment} \mid x, r_j) \tag{9}$$

The selected response is:

$$r^* = \arg\max_j s_j \tag{10}$$

4.8. Output Formatting and Submission

Predictions are formatted according to the evaluation server specifications and submitted at each round. The system then proceeds to the next round, repeating the pipeline until all interactions are processed. The pipeline is implemented using the Hugging Face `pipeline` API with GPU acceleration (NVIDIA T4). Input truncation and batch processing are applied to ensure stable execution. The system operates deterministically across runs, producing consistent outputs for identical inputs.

This pipeline enables a unified, zero-shot framework that leverages pretrained NLI capabilities for both structured clinical scoring and response selection.

5. Results and Analysis

5.1. Submission and Evaluation Protocol

The proposed system was evaluated using the official MentalRiskES evaluation server [1, 2]. During evaluation, the server incrementally released dialogue data across multiple conversational rounds. At each round, the system fetched the updated dialogue context through authenticated API requests and generated predictions for both tasks.

For Task 1, the system produced item-level predictions for PHQ-9, GAD-7, and CompACT-10 in the required JSON format. For Task 2, the system selected the most appropriate therapist response from a predefined candidate set. Predictions were submitted after every round, after which the server released the next segment of the dialogue. The complete evaluation consisted of 82 conversational rounds. Three independent runs were submitted for both tasks. Since the framework operates deterministically without stochastic decoding, all runs generated highly consistent outputs. That's why, we have considered the best run of each participating team including ours.

5.2. Evaluation Metrics

5.2.1. Task 1

Task 1 evaluated the ability of systems to predict clinically aligned questionnaire scores from conversational data. The official evaluation metrics included Mean Absolute Error (MAE), Macro MAE, and Mean Zero-One Error (MZOE). MAE measures the average absolute deviation between predicted and gold-standard scores, while Macro MAE evaluates prediction quality uniformly across severity levels. MZOE was used for evaluating categorical predictions associated with CompACT-10.

5.2.2. Task 2

Task 2 focused on therapist response selection. The official evaluation metrics included Overall Accuracy, Macro Precision, Macro Recall, Macro F1-score, and Error Recovery Rate. These metrics evaluate the system's ability to identify semantically and contextually appropriate therapist responses from candidate sets.

5.3. Results

5.3.1. Task 1 Results

The proposed framework achieved comparatively stronger performance on PHQ-9 than on GAD-7, indicating that depression-related conversational cues were captured more effectively by the entailment-based hypothesis formulation. Depression symptoms are often expressed more explicitly in conversational settings, enabling stronger alignment with predefined clinical hypotheses. In contrast, anxiety-related symptoms in GAD-7 were more implicit and context-dependent, resulting in comparatively higher prediction deviations. Similarly, CompACT-10 constructs related to psychological flexibility are more abstract and behavior-oriented, making them more difficult to infer from short conversational contexts.

Table 1
Performance of the proposed framework on Task 1

Type	MAE	Macro MAE	MZOE
PHQ-9	1.477	1.463	0.813
GAD-7	2.103	1.556	0.948
CompACT-10	1.637	2.122	0.802

5.3.2. Task 2 Results

The proposed response selection framework achieved competitive performance in identifying contextually appropriate therapist responses. The strong performance observed across precision, recall, and F1-score indicates that the entailment-based inference mechanism was effective in modeling semantic coherence between dialogue context and candidate responses. Unlike Task 1, which required fine-grained ordinal severity estimation, Task 2 primarily involved semantic compatibility assessment, making it better suited for zero-shot NLI-based reasoning.

Table 2
Performance of the proposed framework on Task 2

Metric	Score
Overall Accuracy	0.3556
Macro Precision	0.3531
Macro Recall	0.3546
Macro F1-score	0.3530

5.4. Ranking and Comparison

For Task 1, our team *JUNLP-MH* achieved a rank of 15th among 19 participating teams (considering the best submission of each team). Although the ranking is moderate, the proposed framework operated entirely in a zero-shot setting without supervised fine-tuning, domain adaptation, or ensemble modeling. Most higher-ranked systems likely employed supervised transformer architectures, large-scale fine-tuning, or hybrid ensemble approaches. In contrast, our framework relied solely on pretrained multilingual NLI inference with manually designed clinical hypotheses. makecell

For Task 2, our team achieved a rank of 12th among 18 participating teams. Notably, the best-performing system achieved an overall accuracy of 0.3926, while our proposed framework achieved 0.3556, resulting in a relatively small performance gap. This result demonstrates the effectiveness of entailment-based semantic reasoning for the selection of conversational responses. Despite the simplicity of the proposed architecture and the absence of supervised optimization, the framework remained competitive with more complex systems.

Table 3
Performance comparison of our team with others for Task 1.

Team	MZOE_GAD7	MAE_GAD7	MZOE_PHQ9	MAE_PHQ9	MZOE_CompACT10	MAE_CompACT10	MAE_Combined
FBKillers (Best Team)	0.527	0.595	0.623	0.749	0.704	1.293	0.879
BASELINE-Gemma	0.515	0.582	0.599	0.724	0.806	1.7	1.002
JUNLP_MH (Our Team)	0.948	2.103	0.813	1.477	0.802	1.637	1.739
yaregresepekas (Lowest Ranked Team)	0.983	2.253	0.820	1.516	0.989	4.175	3.616

Table 4
Performance comparison of our team with others for Task 2.

Team	Overall Accuracy	Macro Precision	Macro Recall	Macro F1
NLP Innovators (Best Team)	0.393	0.392	0.392	0.392
JUNLP_MH (Our Team)	0.356	0.323	0.335	0.229
BASELINE-Gemma	0.180	0.187	0.182	0.176
yaregresepekas (Lowest Ranked Team)	0.327	0.109	0.333	0.164

Overall, the results demonstrate that zero-shot NLI-based reasoning constitutes a viable and interpretable baseline for conversational mental health assessment, particularly for therapist response selection and dialogue-level semantic understanding.

5.5. Analysis

The experimental findings reveal distinct performance characteristics across the two tasks. Task 1 required precise clinical severity estimation aligned with structured psychometric instruments, whereas Task 2 focused primarily on semantic response ranking and contextual coherence. The proposed framework demonstrated comparatively stronger effectiveness in Task 2, suggesting that pretrained NLI models are particularly well-suited for dialogue-level semantic reasoning. In contrast, fine-grained severity estimation in Task 1 requires more nuanced modeling of long-range conversational dependencies and subtle psychological indicators. Another important observation is the interpretability of the proposed system. Unlike black-box classification approaches, the hypothesis-driven framework explicitly aligns conversational evidence with clinically meaningful statements, enabling symptom-level reasoning and transparent prediction generation.

6. Conclusion and Future Scope

This work presented a hypothesis-driven NLI framework for structured mental health assessment from multi-turn conversational data. Unlike conventional sentiment- or emotion-based approaches, the proposed system aligns conversational evidence with clinically grounded questionnaire items from PHQ-9, GAD-7, and CompACT-10 through explicit hypothesis evaluation. A unified inference pipeline was developed to address both clinical score prediction and therapist response selection using a pretrained multilingual NLI model in a zero-shot setting. The system was evaluated within the MentalRiskES shared task framework, where it demonstrated moderate effectiveness in identifying depression-related and psychological flexibility signals while maintaining a fully interpretable prediction process. The study further highlights the feasibility of using inference-based reasoning for clinically aligned conversational analysis without relying on task-specific training data.

Future work will focus on improving severity prediction through domain-specific fine-tuning, better context modeling, and clinically optimized hypothesis design. Incorporating LLMs and instruction-tuned architectures may further enhance contextual reasoning and response selection performance. Additionally, hybrid frameworks combining NLI-based reasoning with supervised transformer models could improve fine-grained clinical assessment in multi-turn conversational settings.

Declaration on Generative AI

The authors acknowledge limited assistance from ChatGPT (free version), a generative AI tool for text editing and image generation.

References

[1] A. M. Mármol-Romero, F. Suárez-Maroto, R. García-Viedma, M. Hernández-López, F. M. Plaza-del Arco, M. D. Molina-González, L. A. Ureña-López, A. Montejo-Ráez, Overview of mentalriskes at iberlef 2026: Early detection of mental disorders risk in spanish, Procesamiento del Lenguaje Natural 77 (2026).

[2] A. Bonet-Jover, J. A. González-Barba, L. Chiruzzo, Overview of iberlef 2026: Natural language processing challenges for spanish and other iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.

[3] M. Khan, et al., Transformer-based language models for mental health issues: A survey, Pattern Recognition Letters 167 (2023) 204–211.

[4] Z. E. Imel, M. J. Tanana, C. S. Soma, et al., Mental health counseling from conversational content with transformer-based machine learning, JAMA Network Open 7 (2024) e2352590.

[5] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn, The development and psychometric properties of liwc2015, University of Texas at Austin (2015).

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[7] A. Conneau, K. Khandelwal, N. Goyal, et al., Unsupervised cross-lingual representation learning at scale, in: Proceedings of ACL, 2020, pp. 8440–8451.

[8] C. Lau, X. Zhu, W.-Y. Chan, Automatic depression severity assessment with deep learning using parameter-efficient tuning, Frontiers in Psychiatry 14 (2023) 1160291.

[9] J. Wright-Berryman, et al., Virtually screening adults for depression, anxiety, and suicide risk using machine learning and language from an open-ended interview, Frontiers in Psychiatry 14 (2023) 1143175.

[10] I. R. Podina, A.-M. Bucur, L. Fodor, R. Boian, Screening for common mental health disorders: a psychometric evaluation of a chatbot system, Behaviour & Information Technology 44 (2024) 2160–2169.

[11] R. Boiana, et al., A conversational agent framework for mental health screening: design, implementation, and usability, Behaviour & Information Technology 44 (2024) 2364–2378.

[12] Y.-M. Cho, S. Rai, L. Ungar, J. Sedoc, S. C. Guntuku, An integrative survey on mental health conversational agents to bridge computer science and medical perspectives, in: Proceedings of EMNLP, 2023.

[13] G. Cameron, et al., Conversational agents for depression screening: A systematic review, International Journal of Medical Informatics 181 (2024) 105272.

[14] V. Varadarajan, S. Sikström, O. N. E. Kjell, H. A. Schwartz, Alba: Adaptive language-based assessments for mental health, in: Proceedings of NAACL, 2024.

[15] Z. Zhong, Z. Wang, Intelligent depression prevention via llm-based dialogue analysis, arXiv preprint arXiv:2504.16504 (2025).

[16] J. Tang, Y. Shang, Advancing mental health pre-screening: A new custom gpt for psychological distress assessment, arXiv preprint arXiv:2408.01614 (2024).

[17] W. Yin, J. Hay, D. Roth, Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, in: Proceedings of EMNLP-IJCNLP, 2019, pp. 3914–3923.