LLM Questionnaire Completion for Automatic Psychiatric Assessment

Gony Rosenman
Sagol School of Neuroscience
Tel Aviv University
gonyrosenman@mail.tau.ac.il

Lior Wolf
School of Computer Science
Tel Aviv University
liorwolf@gmail.com

Talma Hendler
Sagol School of Neuroscience
Tel Aviv University
hendlert@gmail.com

Abstract

We employ a Large Language Model (LLM) to convert unstructured psychological interviews into structured questionnaires spanning various psychiatric and personality domains. The LLM is prompted to answer these questionnaires by impersonating the interviewee. The answers obtained are coded as features, which are used to predict standardized psychiatric measures of depression (PHQ-8) and PTSD (PCL-C), using a Random Forest Regressor. Our approach is shown to enhance diagnostic accuracy compared to multiple baselines. Thus, it establishes a novel framework for interpreting unstructured psychological interviews, bridging the gap between narrative-driven and data-driven approaches for mental health assessment.

1 Introduction

Psychiatric evaluation, nowadays, is heavily based on patient verbal reports of disturbed feelings, thoughts, behaviors, and their changes over time. It consists of two main parts: unstructured interviews, where patients express themselves freely with open-ended questions, and structured questionnaires to standardize the assessment. These methods, outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM), aim to assign universal scores to individual experiences (American Psychiatric Association, 2013). However, the complexity of mental health, characterized by a positive manifold of symptoms and the subjective, often unreliable nature of self-reports (especially between sessions), as well as interviewer bias, makes accurate diagnosis difficult (Althubaiti, 2016; Bauhoff, 2011). The overlap of symptoms and unstable mental states, particularly in pathological conditions, adds to the challenge of achieving precision, limiting an objective quantitative account of subjective self-experience (Vanes and Dolan, 2021; Dwyer et al., 2018; Bzdok and Meyer-Lindenberg, 2018).

The evolution of psychiatric practice is increasingly shaped by integrating Natural Language Processing (NLP) and machine learning into traditional diagnostics (Malgaroli et al., 2023). This shift moves from a theory-driven to a data-driven paradigm, addressing the limitations of conventional psychiatric assessments. NLP plays a key role in overcoming the subjective interpretation of unstructured interviews and the rigidity of standardized questionnaires. By using large-scale text data and advanced language models, NLP adds nuanced, patient-specific insights to psychiatric evaluations. This data-driven approach complements existing theories, fostering a more precise and holistic understanding of mental health, aligning with the precision medicine trend in healthcare.

We propose a two-step method for psychiatric evaluation using unstructured interview text. First, the LLM is asked to complete multiple questionnaires while impersonating the interviewee. These include (i) established psychiatric questionnaires (PHQ-8, PCL-C), and (ii) custom questionnaires created with GPT-4, covering mental health, personality traits, and therapeutic aspects. In the second step, the LLM's responses are used as features for a Random Forest Regressor (Breiman, 2001), which predicts the interviewee's scores on the two clinical questionnaires.

2 Related Work

Earlier diagnostic efforts in mental health (Asgari et al., 2014; Al Hanai et al., 2018; Niu et al., 2021; Dai et al., 2021; Lu et al., 2022) incorporate both textual and speech intonation features, recognizing the role of tonal elements in depression assessment. Sun et al. (2017) manually extract six psychological dimensions (sleep quality, PTSD/Depression diagnosis, treatment history, introversion, personal preference, and feelings) from interview text, then apply a Random Forest Regressor.

Findings of the Association for Computational Linguistics: EACL 2024, pages 403–415
November 12-16, 2024

Recently, transformer architectures have driven significant progress in depression diagnosis. Milintsevich et al. (2023) use RoBERTa (Liu et al., 2019) to encode interview segments, focusing on symptom prediction, while Zhang and Guo (2024) introduce prompt learning and an attention mechanism for enhanced accuracy. These approaches rely on supervised learning, whereas our work utilizes zero-shot impersonation.

Galatzer-Levy et al. (2023) explored the Med-PaLM 2 LLM (Singhal et al., 2023) for psychiatric prediction, not by answering questionnaires but by predicting outcomes using prompts like, "Based on the following clinical interview, what do you estimate the Participant's [PHQ-8/PCL-C] score is?" Concurrently, (Yang et al., 2023) applies prompt-based few-shot learning (Brown et al., 2020) to analyze social media content, answering questions like "What mental disorder symptoms does this post show?" with a focus on textual explanation quality.

3 Method

This work leverages the Extended Distress Analysis Interview Corpus (E-DAIC) (DeVault et al., 2014), a dataset of semi-clinical interviews designed to assess psychological distress. To build the corpus, an autonomous AI interviewer was employed to minimize human bias and two clinical questionnaire-based scales, PHQ-8 and PCL-C, were used to measure both depression and post-traumatic stress disorder (PTSD). Our analyses adhere to the predefined training, development, and test splits as specified in the dataset documentation.

The E-DAIC dataset comprises $n = 275$ psychological transcripts, each associated with two psychiatric scores quantifying the severity of depression and PTSD, measured using the PHQ-8 and PCL-C standardized scales, respectively.

Given a text in the domain of psychological transcripts $t_i \in \text{PT}$, where $i \in 1, 2, \ldots, 275$, our objective is to design a feature extractor $F : \text{PT} \rightarrow \mathbb{R}^d$. This extractor converts a given transcript t_i into a d-dimensional feature vector $\mathbf{v}_i$ such that the two ground truth scores $s_i = [s_i^{\text{DEPRESSION}}, s_i^{\text{PTSD}}]$ can be accurately inferred. Once features are extracted, the next step involves employing a Random Forest Regressor, $R : \mathbb{R}^d \rightarrow \mathbb{R}^2$, which takes the feature vector $\mathbf{v}_i$ as input and predicts the pair of psychiatric scores s_i.

Our method, Language Model for Impersonation-based Questionnaire Completion (LMIQ), defines a feature extractor F as the transformation of unstructured interview transcripts into a structured array of impersonated responses to a list of questions. For each subject and question, the method: (1) generates a query by combining the transcript with a psychological questionnaire, (2) processes the query through an LLM to obtain responses, and (3) stores the responses in an array, which represents the feature vector for that subject.

The system prompt directed the LLM to interpret transcripts and simulate responses as the interview subject. The exact prompt used in our research is detailed in Appendix A.1, and a shorter reference prompt is as follows.

> Attached are a psychological interview transcript and psychological questionnaire.
> Analyze the conversation transcript and capture psychological insights about the speaker.
> Deduct answers to the accompanied questions as if you were the speaker.
> The answer should be a numerical value ranging from 1-5, with 1 depicting "don't agree at all" and 5 depicting "very much agree."
> Conversation Transcript: {transcript}
> Psychological Questionnaire: {question}

In our implementation, LMIQ utilizes GPT-3.5 Turbo Instruct, accessed via the OpenAI API (https://platform.openai.com/docs/models/gpt-3-5), for transcript analysis and predicting responses to psychological questionnaires. Additionally, we evaluated LMIQ using Mixtral 7Bx8 (Jiang et al., 2024) as its LLM backbone. Both LLMs were employed with their default configurations, including temperature settings, to ensure consistent and unbiased processing. The questions from the questionnaires were submitted to the LLM one at a time in separate sessions to prevent bias from question ordering.

The list of questions used by LMIQ includes the original eight-item PHQ-8 and seventeen-item PCL-C questionnaires, as well as several additional questionnaires we developed using OpenAI's ChatGPT-4. We prompted it to create multiple five-item sets across three domains: (i) clinical mental health (DSM-5 Guided), (ii) the five-factor personality model, and (iii) therapeutic-relevant domains. The broader range of questionnaires, beyond PHQ-

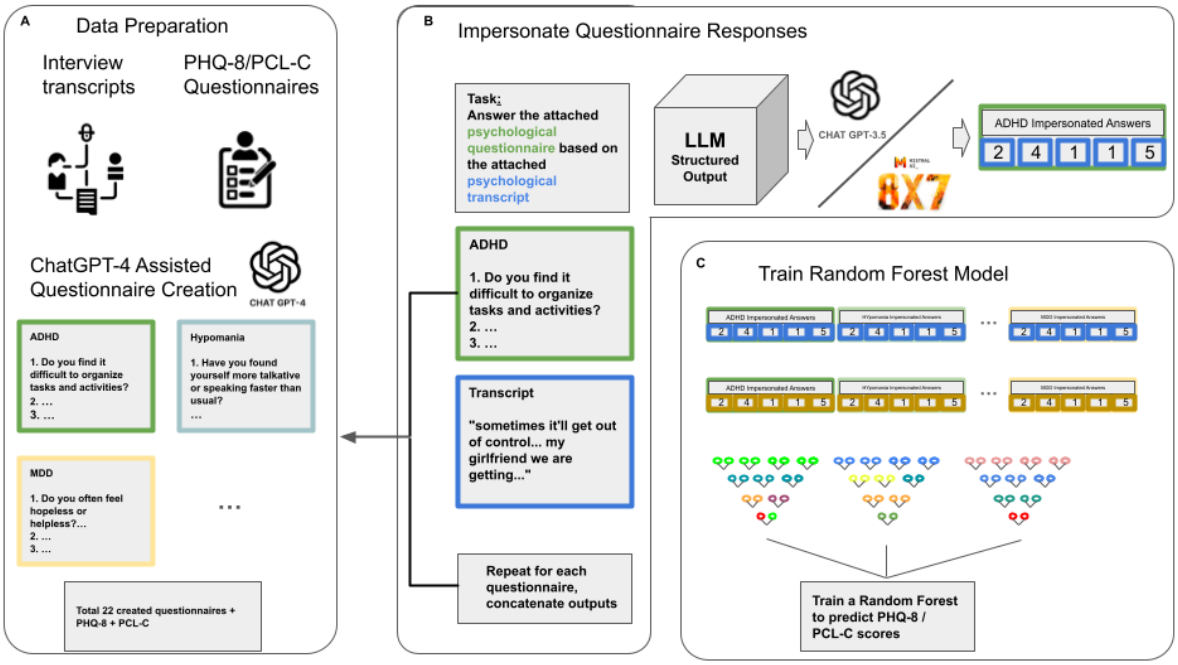

Figure 1: LMIQ pipeline overview. A. Top. The Extended-Daic dataset, comprising 275 psychological interviews with associated PTSD and Depression scores derived from the PCL-C and PHQ assessments. Bottom. 5-item questionnaires across multiple mental health and personality domains. B. Main prompt logic. For each instance, a task description is merged with a psychological interview transcript and a questionnaire, generating 5 impersonated responses. C. Feature Vector Formation: Aggregate the 5 simulated responses from each participant to construct a feature vector of dimension $d = 135$, which is then used to train a Random Forest model for accurate prediction of the original assessment scores.

8 and PCL-C, addresses psychiatric comorbidity and symptom overlap.

Within each domain, GPT-4 was prompted to define multiple topics and within each topic to generate exactly five questions, each aiming to quantify symptoms on a scale between 1 and 5. These initial outputs were manually refined to ensure higher relevance and precision. The obtained domains are listed in Tab. 1, the full list of questions is provided in Appendix B; as a sample, the generalized anxiety disorder questions include: (1) Do you find it hard to control your worrying? (2) Does your anxiety interfere with your work, school, or family responsibilities?

The feature vector was formed by concatenating responses (each an integer between 1 and 5) to five questions across twenty-four topics, resulting in a vector with $d = 135$ dimensions. This vector was then used as input for a Random Forest Regressor implemented with Scikit-Learn (Pedregosa et al., 2011). Model optimization involved a hyper-parameter search, specifically examining 'n-estimators': [100, 200, 300] and 'max-depth': [10, 20, 30]. The configuration yielding the best performance on the E-DAIC development set was applied to the test set for evaluation. A detailed overview of our methodology, from data collection to prediction, is provided in Figure 1.

Domain	Questionnaires	#Questions
Mental health	Agoraphobia, ADHD, Body Image, Borderline Personality Disorder, etc.	70
Personality	Agreeableness, Conscientiousness, Extraversion, Neuroticism, Openness	25
Therapeutic	Family History, Trauma History, Resilience	15
Direct	PHQ-8, PCL-C	25

Table 1: Questionnaire Domains and Constituents

Model	PHQ depression score		PTSD severity scale	
	Dev MSE	Test MSE	Dev MSE	Test MSE
LMIQ	23.87	20.42	144.17	192.93
LMIQ (Mixtral)	18.50	18.05	90.11	163.75
GPT-3.5 Impersonate Zero-Shot	27.62	24.16	154.66	195.55
GPT-3.5 Zero-Shot	35.62	33.42	251.70	336.49
GPT-3.5 Analyze & Embed	29.81	35.80	216.13	348.02
Direct Embedding with Ada-002	31.83	38.23	226.18	389.65
MentaLLama 7B Analyze & Embed	29.61	39.40	216.30	367.09
Guessing the mean value	32.06	42.80	234.52	407.2
TF-IDF Vectorization	31.47	43.36	227.61	392.63

Table 2: Summary of Model Performances in PHQ-Score and PTSD Severity Prediction Tasks

Method	MAE Dev/Test	RMSE Dev/Test
LMIQ (Ours)	3.87/3.46	4.78/4.30
Zhang and Guo (2024)	-/ 5.28	- / -
Niu et al. (2021)	3.73 / -	4.80 / -
Dai et al. (2021)	3.22/3.98	4.43/5.11
Lu et al. (2022)	4.48/-	5.37/-
Milintsevich et al. (2023)	5.51/ 5.03	-/ -

Table 3: Performance on the DAIC-WoZ Dataset

4 Experiments

We employ multiple relevant baselines and various ablated versions of our method to demonstrate the advantage of our approach in predicting psychiatric scores from psychological transcripts. Unless otherwise specified, all feature extraction methods are paired with the same Random Forest Regression used for LMIQ.

Direct Embedding with Ada-002 This baseline method uses text-embedding-ada-002 (Neelakantan et al., 2022), a model from the OpenAI API, to convert psychological transcripts into vector embeddings.

GPT-3.5 Analyze & Embed This baseline combines GPT-3.5's analysis of psychological transcripts, focusing on conditions like depression and PTSD, followed by embedding the response with text-embedding-ada-002. The prompt and a sample response are provided in Appendix A.2.

MentalLLAMA 7B Analyze & Embed Similar to GPT-3.5 Analyze & Embed, but employing the MentalLLama 7B model (Yang et al., 2023), a LLAMA (Touvron et al., 2023) model fine-tuned for depression detection using a large dataset of Reddit posts from mental health subreddits. Sample output is in Appendix A.3.

TF-IDF Vectorization This traditional approach vectorizes each transcript using the Term Frequency-Inverse Document Frequency (TF-IDF) method.

GPT-3.5 Analyze & Predict This baseline directs GPT-3.5 to perform a zero-shot prediction of psychiatric scores PHQ-8 and PCL-C, similar to (Galatzer-Levy et al., 2023).

GPT-3.5 Impersonate Zero-Shot This baseline instructs GPT-3.5 to answer the PHQ-8 and PCL-C questionnaires as if it were the subject, based on the psychological transcript. This method calculates scores directly as the sum of questionnaire responses, mimicking clinical evaluation. Unlike the main method, it bypasses the Random Forest model and operates as a zero-shot approach.

Naive Average Guess To assess score variability, the mean value of the training set is used as the prediction for all subjects, regardless of the transcript.

5 Results and Discussion

Table 2 shows that the LMIQ model yields lower Mean Squared Errors (MSEs) than all baselines for both depression and PTSD severity, indicating superior accuracy.

Additionally, LMIQ, using the Mixtral 7Bx8 model—which is believed to outperform GPT-3.5—shows improved results, highlighting the effectiveness of our approach and the potential of newer LLMs. The next best method predicts only clinical questionnaires, summing the results instead

of using a regression model. This simplified LMIQ approach, while still employing impersonation, performs only slightly worse than the full method. A full set of experiments examining the contribution of each questionnaire type is detailed in Appendix C. The Zero-Shot method, which directly predicts questionnaire scores, consistently underperforms.

The classic TF-IDF encoding method yields the highest MSEs in both tasks. GPT-3.5 Analyze & Embed and MentaLLama 7B Analyze & Embed show mixed results but consistently outperform the method that directly embeds the interview with Ada-002.

Evaluation on the DAIC-WoZ Dataset

The WOZ-DAIC dataset (Gratch et al., 2014), a predecessor of E-DAIC, lacks PTSD scores and features interviews with human interviewers. This earlier benchmark allows for comparison between LMIQ and published methods.

The comparison results are shown in Table 3, including both MAE and RMSE metrics, as per the literature. Our method outperforms text-based methods on the test set. HCAG (Niu et al., 2021) does not report test results and slightly outperforms us on the development set in one score. Dai et al. (2021) perform better on the development set but worse on the test set, suggesting potential overfitting. Lu et al. (2022) perform worse. Notably, the last two models include audio features, which our approach does not.

6 Conclusions

LMIQ leverages an LLM as an impersonator to complete both established and LLM-generated clinical questionnaires based on a subject's unstructured psychological dialogue, feeding these into a classifier for diagnosis. Although we do not validate the impersonated responses due to the lack of available answers, we demonstrate their effectiveness in predicting PHQ and PCL scores. Despite LLMs' limitations in numerical accuracy in some contexts, our results confirm the meaningfulness of the assigned values, both directly and through regression trees.

7 Limitations

Our method, while generic, has been validated solely against the E-DAIC and WOZ-DAIC benchmarks, highlighting a challenge in generalizability. To address this and promote clinical adoption, more data is needed, and benchmarking against variability among human experts is crucial. This step is vital for enhancing robustness and applicability in clinical settings, covering a wider range of psychological conditions and populations.

The LLM used is a generic one. Due to time constraints, experiments with Mixtral 7Bx8 and MentaLLama were limited. It would be valuable to evaluate models like Med-PaLM2 and others.

A key limitation of our work is LLMs' tendency to "hallucinate" or misinterpret context, a significant issue in the sensitive domain of psychological data, where inaccuracies can have serious consequences. Nonetheless, we aim to identify meaningful links between unstructured data and symptoms, acknowledging that perfectly reflecting individual experiences is unrealistic.

8 Ethics and Impact Statement

The use of LMIQ to transform unstructured text into psychiatric diagnoses raises ethical concerns, particularly the risk of assigning diagnostic labels without explicit consent. Additionally, the risks of misdiagnosis in psychiatry emphasize the need for caution when using machine learning tools, given their limited accuracy and potential biases, which could result in harmful outcomes. In compliance with the Health Insurance Portability and Accountability Act (HIPAA) of 1996, our method ensures the de-identification of patient-provider conversations before processing with LLMs, protecting patient privacy and legal standards. Thus, the ethical use of such technologies requires careful consideration to avoid unintended harm.

9 Acknowledgments

This work was supported by the Sagol Brain Institute, NIMH, the Israel Science Foundation (Grant No. 2923/20) within the Israel Precision Medicine Partnership program, and the Tel Aviv University Center for AI and Data Science (TAD). We are grateful for their support.

References

Tuka Al Hanai, Mohammad Ghassemi, and James Glass. 2018. Detecting Depression with Audio/Text Sequence Modeling of Interviews. In *Proc. Interspeech 2018*, pages 1716–1720.

Alaa Althubaiti. 2016. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare*, page 211.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, 5th edition. American Psychiatric Association, Washington, D.C.

Meysam Asgari, Izhak Shafran, and Lisa B. Sheeber. 2014. Inferring clinical depression from speech and spoken utterances. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE.

Sebastian Bauhoff. 2011. Systematic self-report bias in health data: impact on estimating cross-sectional and treatment effects. *Health Services and Outcomes Research Methodology*, 11(1–2):44–53.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Danilo Bzdok and Andreas Meyer-Lindenberg. 2018. Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(3):223–230.

Zhijun Dai, Heng Zhou, Qingfang Ba, Yang Zhou, Lifeng Wang, and Guochen Li. 2021. Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. *Journal of Affective Disorders*, 295:1040–1048.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068.

Dominic B. Dwyer, Peter Falkai, and Nikolaos Koutsouleris. 2018. Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, 14(1):91–118.

Isaac R. Galatzer-Levy, Daniel J. McDuff, Vivek Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. 2023. The capability of large language models to measure psychiatric functioning. *ArXiv*, abs/2308.01834.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

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. *arXiv preprint arXiv:1907.11692*.

Jiahao Lu, Bin Liu, Zheng Lian, Cong Cai, Jianhua Tao, and Ziping Zhao. 2022. Prediction of depression severity based on transformer encoder and cnn model. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE.

Matteo Malgaroli, Thomas D. Hull, James M. Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13(1).

Kirill Milintsevich, Kairit Sirts, and Gaël Dias. 2023. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, 10(1).

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and code embeddings by contrastive pre-training.

Meng Niu, Kai Chen, Qingcai Chen, and Lufeng Yang. 2021. Hcag: A hierarchical context-aware graph attention model for depression detection. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Bo Sun, Yinghui Zhang, Jun He, Lejun Yu, Qihua Xu, Dongliang Li, and Zhaoying Wang. 2017. A random forest regression method with selected-text feature for depression assessment. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, MM '17. ACM.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Lucy D. Vanes and Raymond J. Dolan. 2021. Transdiagnostic neuroimaging markers of psychiatric risk: A narrative review. *NeuroImage: Clinical*, 30:102634.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Sophia Ananiadou, and Jimin Huang. 2023. Mentallama: Interpretable mental health analysis on social media with large language models.

Jun Zhang and Yanrong Guo. 2024. Multilevel depression status detection based on fine-grained prompt learning. *Pattern Recognition Letters*, 178:167–173.

Appendix

A The full prompts used

A.1 System prompt - Question Impersonation

This is the prompt used in the main method, LMIQ.

"""Analyze a therapist-subject conversation transcript and related psychological questionnaire questions. Focus on understanding the subject's mental health by examining their dialogue for both explicit and implicit cues. Pay attention to signs of depression, PTSD, and other conditions, but also note the absence of these symptoms.

Your task is to provide answers to the questionnaire as if you were the subject, based on insights from the conversation. Ensure your responses are balanced, reflecting the subject's mental state as suggested by the transcript. Make informed deductions about the subject's emotional state, stress management, life satisfaction, social interactions, and professional aspirations, providing answers that reflect mental health issues or a neutral state as indicated by the transcript.

Transcript: {raw_psyc_transcript}
Answer the questions with an agreement level ranging from {agreement_range}.
Questions: {questions}"""

A.2 System Prompt - Analysis

This is the prompt used for the Analyze & Embed Baseline.

"""Analyze a therapist-subject conversation transcript. Focus on the clear and comprehensible parts of the subject's dialogue, as the therapist's words are omitted and there may be gaps due to recording issues. Your analysis should identify key statements or moments that reveal aspects of the subject's mental health, with particular attention to signs of depression and PTSD. Prioritize brevity and clarity in your analysis. Look for indicators of the subject's emotional state, stress management, life satisfaction, social interactions, and professional aspirations. Assess the intensity of these elements, capturing the range from mild to profound, based on the subject's expressions.

Ensure your analysis remains accurate and relevant to the transcript, avoiding unfounded assumptions or 'hallucinations'. Summarize your findings clearly and directly, using concise sentences that reflect the nuances of the subject's mental state. Disregard unclear or incomplete segments due to recording issues, and avoid structural phrases or introductions to maintain data integrity for subsequent embedding generation. The accuracy, clarity, and nuanced understanding of intensity in your summary are paramount.
Conversation transcript: {raw_psyc_transcript}"""

A sample response using GPT-3.5 is:

> The subject appears to have a positive outlook on life and their current location. They are interested in starting their own business and have a good relationship with their family, who have given them valuable advice. However, there are also signs of stress and irritability, as well as a recent episode of shutting down and cutting off a friend They also mention feeling lazy and less interested in activities they...Overall, the subject seems to have a mix of positive and negative emotions, and may benefit from further exploration and support for any underlying mental health issues.

A.3 Sample output by MentaLLama

A sample MentaLLama response is:

> The subject expresses a positive outlook on life, including enjoyment of reading books, cooking, and spending time with friends and family. However, they also mention experiencing congestion and having a break in their education. They express a desire to open up their own business but do not provide a specific reason. The subject mentions receiving advice from their parents and grandparents. Overall, the analysis suggests a balance between positive and negative experiences, indicating a potential need for further assessment and support."

B The full list of the questionnaires used.

For each topic from each domain, we provide the questions that were extracted in collaboration with ChatGPT 4.0.

B.1 Mental Health Domain

Agoraphobia

- Do you fear being in open or enclosed spaces?
- Do you avoid public transportation or being in crowds?
- Does the thought of leaving your home alone cause you anxiety?

– How do these fears limit your daily activities or lifestyle?
– Do you require a companion when going out due to these fears?

Attention-Deficit/Hyperactivity Disorder

– Do you find it difficult to organize tasks and activities?
– Do you often forget appointments or daily activities?
– Do you often make careless mistakes in work or other activities?
– Do you get easily sidetracked by extraneous stimuli?
– How much do these symptoms impact your performance in work or school?

Body Image

– How would you describe your level of satisfaction with your appearance?
– Do you often compare your body to others?
– How much does your perception of your body affect your daily life?
– Do you have any concerns about your eating habits or weight management?
– How confident do you feel in your abilities and decisions?

Borderline Personality Disorder

– Do you experience intense and unstable relationships with others?
– Do you often feel empty or bored?
– Do you experience mood swings that can last for a few hours to a few days?
– Do you have a fear of abandonment, either real or imagined?
– Do you engage in impulsive behaviors, like substance abuse or reckless driving?

Delusions

– Do you feel disconnected from reality at times?
– Do you have strong beliefs that others find unusual or unrealistic?
– Do you feel controlled or influenced by external forces or beings?
– Have you experienced changes in your perception or senses that others do not?
– Do these experiences cause you distress or impair your functioning?

Generalized Anxiety Disorder

– Do you find it hard to control your worrying?
– Does your anxiety interfere with your work, school, or family responsibilities?
– Do you experience physical symptoms of anxiety, like muscle tension or restlessness?
– Do you often feel irritable or on edge?
– Do you have trouble sleeping due to worry?

Hypomania/Mania

– Have you found yourself more talkative or speaking faster than usual?
– Do you often feel overly confident in your abilities or ideas?
– Have you engaged in risky behaviors, like excessive spending or reckless driving?
– Do you find your thoughts racing or jumping from topic to topic?
– Have others noticed a significant change in your mood or behavior?

Major Depressive Disorder

– Do you often feel hopeless or helpless?
– Have you noticed a change in your appetite or weight without trying to lose or gain weight?
– Do you struggle to concentrate on tasks or make decisions?
– Do you often feel worthless or excessively guilty about things?
– Have your sleep patterns changed, such as sleeping too much or too little?

Obsessive-Compulsive Disorder

– Do you check things repeatedly or have rituals that you feel compelled to perform?
– Do your thoughts or rituals cause you distress or interfere with your daily life?
– Do you spend more than an hour a day on these thoughts or rituals?
– Do you avoid certain situations or activities because of your fears or compulsions?
– Do you need to have things arranged in a specific order or manner?

Panic Attacks

– During panic attacks, do you feel like you're losing control or going crazy?
– Do you fear these attacks to the point of altering your daily routines?

– Have you visited the emergency room or sought medical help for these symptoms?
– Do you avoid places or situations for fear of triggering an attack?
– How do these attacks impact your daily life?

Persistent Depressive Disorder

– Have you experienced low mood more days than not for at least two years?
– Do you feel like you've been in a mild but constant state of depression?
– Do you find little pleasure in activities you once enjoyed?
– Do you struggle with feelings of inadequacy or low self-esteem?
– Have you experienced changes in your appetite or sleep patterns?

Post-Traumatic Stress Disorder

– Do you experience heightened vigilance or jumpiness?
– Are you engaging in self-destructive or risky behavior since the event?
– Do you feel numb or detached from people, activities, or surroundings?
– Do you find yourself being easily angered or having aggressive outbursts?
– Have you noticed any changes in your beliefs or feelings about yourself and others?

Social Phobia

– Do you fear being criticized or embarrassed in social situations?
– Does speaking to unfamiliar people cause you significant anxiety?
– Do you avoid social situations due to fear of being judged?
– Do physical symptoms like sweating or trembling accompany your fear in social settings?
– How does this fear impact your personal or professional life?

Substance Abuse

– Do you use any substances like drugs or alcohol regularly?
– How often do you find yourself using these substances?
– Have you noticed any negative impacts on your health, work, or relationships due to substance use?
– Do you feel a strong desire or compulsion to use these substances?
– Have you tried to cut down or stop using these substances in the past?

B.2 Personality

Agreeableness

– Do you often find yourself making compromises to maintain harmony in your relationships?
– Would you describe yourself as someone who is generally trusting of others?
– How often do you get into arguments with people?
– Do you tend to empathize easily with others?

Conscientiousness

– How often do you set and achieve long-term goals?
– Do you prefer having a set schedule or being spontaneous?
– How would you rate your ability to resist temptations or distractions?
– Do you take pride in the accuracy and detail of your work?
– How do you handle important deadlines?

Extraversion

– Do you enjoy being the center of attention in social gatherings?
– How often do you initiate conversations with people you don't know?
– Do you prefer group activities or solitary activities?
– Do you feel energized when interacting with a large group of people?
– How would you describe your level of assertiveness in social situations?

Neuroticism

– Do you often feel anxious or worried about various aspects of your life?
– How do you react to stressful situations?
– Do you frequently feel mood swings or emotional instability?
– Do you often have trouble sleeping due to worrying?
– How often do you experience feelings of sadness or depression?

Openness

- Do you enjoy trying new activities and visiting new places?
- How often do you engage in creative activities like writing, painting, or playing music?
- Do you enjoy discussing abstract concepts and ideas?
- How do you feel about change and variety in your life?
- Would you say you are open to new and diverse perspectives or opinions?

B.3 Therapeutic

Family History

- Do you feel you have a strong and positive relationship with your family members?
- Do you manage conflicts or disagreements within your family effectively?
- Is there a history of mental health issues or substance use in your family?
- Do you believe your family background has significantly influenced your current life choices and behaviors?
- Do you feel supported and understood by your family?

Trauma History

- Have you ever experienced a traumatic event such as physical, emotional, or sexual abuse?
- How do you feel this event has affected your life?
- Do you often think about or have flashbacks to this traumatic event?
- How do you typically cope with reminders of the trauma?
- Have you sought any professional help to deal with the aftermath of this traumatic experience?

Resilience

- How quickly do you recover from setbacks or disappointments?
- Do you often find positive aspects in negative situations?
- How do you usually cope with stress and pressure?
- Do you feel confident in your ability to handle new challenges?
- How often do you bounce back from hardships stronger than before?

B.4 Direct

PHQ-8

- How much are you experiencing little interest or pleasure in doing things?
- How likely are you to volunteer your time to help others?
- How much are you feeling down, depressed, or hopeless?
- How much are you having trouble with falling or staying asleep, or sleeping too much?
- How much are you feeling tired or having little energy?
- How much are you experiencing poor appetite or overeating?
- How much are you feeling bad about yourself, or that you are a failure or have let yourself or your family down?
- How much are you having trouble concentrating on things, like reading the newspaper or watching television?
- How much are you moving or speaking so slowly that other people might have noticed, or the opposite – being so fidgety or restless that you've been moving around a lot more than usual?

PCL-C

- How much are you re-experiencing disturbing memories, thoughts, or images of a stressful experience from the past?
- How much are you experiencing repeated, disturbing dreams of a stressful experience from the past?
- How much are you suddenly acting or feeling as if a stressful experience were happening again (as if you were reliving it)?
- How much are you feeling upset when something reminded you of a stressful experience from the past?
- How much are you having physical reactions (e.g., heart pounding, trouble breathing, sweating) when something reminded you of a stressful experience?
- How much are you avoiding thinking about or talking about a stressful experience from the past or avoiding having feelings related to it?
- How much are you avoiding activities or situations because they remind you of a stressful experience?

- How much are you having trouble remembering important parts of a stressful experience?
- How much are you losing interest in activities that you used to enjoy?
- How much are you feeling detached or estranged from others?
- How much are you feeling emotionally numb or being unable to have loving feelings for those close to you?
- How much are you feeling as if your future will somehow be cut short?
- How much are you having trouble falling or staying asleep?
- How much are you feeling irritable or having angry outbursts?
- How much are you having difficulty concentrating?
- How much are you being 'super alert' or watchful or on guard?
- How much are you feeling jumpy or easily startled?

C Ablating the questionnaires

LMIQ utilizes questionnaires from four domains: DSM-inspired clinical conditions, Five Factor Model personality traits, therapeutic aspects, and the original questions from the PHQ-8 and PCL-C assessments. By systematically omitting each domain, we analyze their individual and collective contributions to diagnostic accuracy using a RandomForest pipeline.

The results are presented in Table 4. Evidently, the direct questions are the most important for PTSD prediction, where the contribution of other domains is less significant, and they cannot replace the direct questionnaires when these are removed. In the case of Depression, the other three domains (combined) can lead to reasonable results even in the absence of the direct questionnaires and contribute more significantly to the test performance.

D Interpretability Analysis of LMIQ

A useful property of the Random Forest algorithm is the ability to identify the most important features. Applying this feature importance analysis to the models used to predict PHQ Scores and PCL Severity is provided in Tables 5 and 6, respectively, revealing the questions that the LMIQ models predominantly rely on for predictions.

The results show a strong alignment between questions, the associated symptoms, and their respective domains (depression or PTSD), highlighting the model's promise. The inclusion of less obvious questions among those deemed highly relevant, such as "Do you often find positive aspects in negative situations?" from the PHQ score influences, and "Do you experience mood swings that can last for a few hours to a few days?" from the PCL severity influences, underscores an opportunity to further explore the model's internal reasoning and its capacity to link everyday language with a wide range of symptoms and behaviors.

Questionnaire Domain				# Features	Performance on PHQ		Performance on PTSD	
M. Health	Personality	Therapuetic	Direct		Dev MSE	Test MSE	Dev MSE	Test MSE
✓	✗	✗	✓	95	22.62	22.05	134.31	177.15
✗	✗	✓	✓	40	24.60	21.00	142.25	168.62
✗	✓	✗	✓	50	26.08	20.26	143.87	166.62
✗	✗	✗	✓	25	26.32	22.74	136.96	188.26
✓	✓	✓	✓	135	23.87	20.42	144.17	192.93
✓	✓	✗	✗	95	26.17	24.95	164.11	195.17
✓	✗	✗	✗	70	24.88	25.44	167.94	195.81
✓	✓	✓	✗	110	25.85	23.76	173.30	230.40
✓	✗	✓	✗	85	25.34	23.45	177.68	229.44
✗	✓	✗	✗	25	32.13	29.47	204.92	252.67
✗	✗	✓	✗	15	32.68	26.87	238.84	255.28

Table 4: An ablation study regarding the contribution of the various domains.

Feature	Relative Importance
Do you often have trouble sleeping due to worrying	0.18
Have your sleep patterns changed, such as sleeping too much or too little?	0.16
Do these experiences cause you distress or impair your functioning?	0.03
Do you often find positive aspects in negative situations?	0.03
How often do you experience feelings of sadness or depression?	0.02

Table 5: Top 5 Questions Influencing the PHQ Score

Feature	Relative Importance
During panic attacks, do you feel like you're losing control or going crazy?	0.09
How do these attacks impact your daily life?	0.08
How do these fears limit your daily activities or lifestyle?	0.06
Do your thoughts or rituals cause you distress or interfere with your daily life?	0.04
Do you experience mood swings that can last for a few hours to a few days?	0.03

Table 6: Top 5 Questions Influencing the PCL Severity