MAQuA: Multi-outcome Adaptive Question-Asking for Mental Health using Item Response Theory

Vasudha Varadarajan[*♠], **Hui Xu**[*♡], **Rebecca Astrid Boehme**[†],
Mariam Marlan Mirström[‡], **Sverker Sikström**[‡] **& H. Andrew Schwartz**[♡,♢]
[♠]Language Technologies Institute, Carnegie Mellon University.
[♡]Stony Brook University, [†]Aarhus University, [‡]Department of Psychology, Lund University.
[♢]College of Connected Computing, Vanderbilt University.
vvaradar@andrew.cmu.edu, hansen.schwartz@vanderbilt.edu

Abstract

Recent advances in LLMs offer new opportunities for scalable, interactive mental health assessment, but excessive querying burdens users and is inefficient for real-world assessments across transdiagnostic symptom profiles. We introduce MAQuA, a multi-outcome modeling and adaptive question-asking framework for simultaneous, multidimensional mental health assessment. Combining multi-outcome modeling on language responses with item response theory (IRT) and factor analysis, MAQuA selects the questions with most informative responses across *multiple* dimensions at each turn to optimize diagnostic information, improving accuracy, and potentially reducing response burden. Empirical results in a novel dataset reveal that MAQuA reduces the number of assessment questions required for score stabilization by 50 to 87% compared to random ordering (e.g., achieving stable depression scores with 71% fewer questions and eating disorder scores with 85% fewer questions). MAQuA demonstrates robust performance in both internalizing (depression, anxiety) and externalizing (substance use, eating disorder) domains, with early stopping strategies that could further potentially reduce patient time and burden. These findings position MAQuA as a powerful and efficient tool for scalable, nuanced, and interactive mental health assessments, advancing the integration of LLM-based agents into real-world clinical workflows.

1 Introduction

Recent progress in large language models (LLMs) has enabled the automated inference of mental health scores from patient-generated natural language. However, comprehensive evaluations indicate that such LLM-based assessments are inconsistent (Ji et al., 2022) and, in many cases, less accurate than dedicated condition-specific models with established psychometric validity (Harrigian et al., 2020). These limitations present critical barriers to the clinical adoption and trustworthiness of LLMs in the mental health domain.

Traditional NLP approaches in mental health have often relied on annotations of specific conditions derived from social media, focusing primarily on single-task models, such as encoder-based classifiers for specific mental health and well-being dimensions such as depression (Coppersmith et al., 2015; Eichstaedt et al., 2018), suicidal ideation or risk (Shen et al., 2017; Varadarajan et al., 2024a), resilience (Mahwish et al., 2026), anxiety (Owen et al., 2020; Juhng et al., 2023). The narrow scope of modeling a single-condition mental health score typically does not capture the comorbidities or the complex, multidimensional nature of the symptoms observed by real-world clinicians (Shani and Stade, 2025; Soni et al., 2025; Varadarajan et al., 2025). More importantly, they do not address the interactive paradigm in which LLM agents engage with users in *prompted* settings: language generated in response to structured diagnostic interviews, as would be typical in real-world clinical settings.

Further, in actual clinical practice, clinicians dynamically adapt their lines of questioning based on prior information received, avoiding redundancy, clarifying ambiguous responses, and addressing emergent concerns (James et al., 2010; Welch et al., 2025). While LLMs excel at modeling linguistic patterns, there are mixed signals on their ability to dynamically ground inferred states in underlying mental health constructs (Singh et al., 2025; Ganesan et al., 2024), especially given the multi-objective challenge of simultaneously selecting next-best questions and evaluating mental health status (Li et al., 2025; Sener and Koltun, 2018). Furthermore, maximizing screening accuracy within the constraints of limited clinician-patient interaction time remains a key priority, as excessive probing, such as via LLM-based dialogue agents can be mentally taxing and lead to decision fatigue or disengagement (Jin et al., 2025), highlighting the need for adaptive systems that select only the most informative and relevant follow-up questions.

To address these challenges, we propose **MAQuA**, an adaptive language-based assessment framework that supports multidimensional mental health modeling. This framework offers two large advantages over previous methods for language-based mental health assessments: (1) it captures *multiple* underlying condition scores simultaneously, and (2) it *adaptively* selects the most informative follow-up questions. Our framework helps guide interactions efficiently toward richer, multi-

[*]equal contribution

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6659–6677

faceted insight about mental health, while also making it suitable to operate alongside LLM agents. Building on item response theory (IRT)-based adaptive assessments introduced by Varadarajan et al. (2024b), our results demonstrate that optimizing information gain across multiple conditions simultaneously can be even more effective than single-condition models or LLM inferences.

In this work, we first explore modeling capabilities for multiple conditions, empirically benchmarking single-task and multitask models on a multidimensional mental health dataset. Then we assess the effectiveness of adaptive question selection in mental health estimation. Our research investigates the following research questions: (1) whether cross-condition information sharing improves per-condition predictive performance; (2) if automatically inferred mental health dimensions are more validated compared to standard questionnaires; and (3) if multidimensional IRT further sustains validity of mental health measures in subsequent adaptive assessment turns. Our main contributions include: (a) a systematic comparison of techniques for improving multi-condition modeling, (b) an adaptive assessment framework for optimizing information gain across outcomes, (c) empirical results demonstrating robust gains in multidimensional adaptive assessment, and (d) the release of code along with novel questionnaire-driven dataset to support future research[1].

2 Background

With a growing need for scalable and nuanced approaches to mental health assessments, especially given that traditional clinical interviews and fixed-item scale assessments are limited by patient burden, clinician time, and difficulties in manually capturing multiple overlapping mental health conditions, Item Response Theory has emerged as an alternative measurement paradigm that enables adaptive assessments instead of traditional questionnaires and interviews. In this section, we provide background on adaptive testing from psychometrics that we employ in the `MAQuA` framework.

Algorithm 1 Adaptive Language-Based Assessment (ALBA)

Initialize $\theta \leftarrow$ initial estimate of trait level
Initialize item prompt pool and empty response list `responses` $\leftarrow \emptyset$
while `responses`, and stopping rule not met on θ **do**
    Select next item p to maximize information at current θ
    Present prompt p and capture free-text response t
    Compute discrete response score s = NLP_score(t)
    Append (p, s) to `responses`
    Update θ using IRT scoring method on `responses`
end while
Output final estimate θ

[1] https://github.com/humanlab/MAQuA-IRT-framework, dataset available upon request

Item Response Theory (IRT) is a probabilistic, data-driven measurement framework that models the relationship between an individual's latent trait score (such as depression severity) and their probability of specific item responses on a questionnaire (Embretson and Reise, 2000; Hambleton et al., 1991).

In this framework, an *item* refers to a single question in the questionnaire or a battery of questionnaires. In single-dimensional IRT, this relationship is modeled with respect to one latent trait (usually denoted θ), accounting for item-specific parameters such as difficulty and discrimination. This means that the model assumes that there is a single factor or "ability" of the respondent being measured by the set of all questions. Fitting a 2-parameter IRT model on survey participants' responses estimates the difficulty of question i (denoted by β_i for single dimensional, b_i for multi-dimensional IRT; more difficult the question, harder it is to answer affirmatively) as well as the discrimination (denoted by α_i for single dimensional, a_i for multi-dimensional IRT; more discriminating the question, more the response of a question informs high ability versus low ability). IRT enables precise ordering and calibration of items based on their informativeness in measuring the latent trait, supporting adaptive and individualized assessment. Adaptive language-based assessments (Varadarajan et al., 2024b) were first introduced using single-dimensional IRT, summarized in Algorithm 1.

Exploratory Factor analysis (EFA) is a statistical technique used to uncover the underlying structure of a set of observed variables by identifying clusters of variables that co-vary together, known as factors or latent constructs (Cudeck and MacCallum, 2000). Although it is very similar to Principal Component Analysis (PCA), it captures the multidimensional, often overlapping nature of latent variables by allowing factors to be correlated, unlike PCA which assumes uncorrelated components. This approach captures meaningful psychological constructs like depression and anxiety, supporting the development of sensitive, multidimensional assessments for accurate mental health assessments. EFA allows us to distill large and complex data such as responses to multiple questionnaires assessing various psychological conditions. Each factor represents a distinct underlying construct that accounts for shared variance among the observed variables, providing insight into how different mental health symptoms or traits may be related at a deeper level.

Multidimensional item response theory (MIRT) therefore relies on factor analysis (FA) to identify and model multiple latent traits underlying assessment items. FA reveals how items relate to different, but correlated psychological dimensions such as depression or anxiety. These factor structures guide the MIRT model by linking items to specific traits, ensuring that the model accurately captures the multidimensional nature of the data. This allows MIRT to estimate individual' scores across overlapping traits efficiently and precisely for

multi-outcome assessment. [2]

This theory has been applied to discrete response patterns such as multiple choice questions, Likert scale etc., and single dimensional IRT modeling. In this paper, we provide a framework to translate effectively into multidimensional language-based assessments.

3 Dataset

To explore our questions regarding multiple conditions (multi-dimensional) modeling as well as adaptive question-asking, we selected a unique set of material including language questions, diagnoses, and validated clinical scales. Our dataset contains a battery of standard Likert-scale-based, psychological questionnaires as well as a large set of questions with open-ended language-based responses from real users – such a dataset is the first of its kind, in our knowledge.

The data was collected in two phases: first, we prescreened the participants to establish a diverse sampling pool including participants who stated to were diagnosed by a mental health professional. We focused on several common mental disorders, such as mood disorders (i.e., major depression, generalized anxiety disorder, bipolar disorder), autism spectrum disorder, attention-deficit/hyperactivity disorder, eating disorders, obsessive-compulsive disorder, post-traumatic stress disorder, and substance use disorders (i.e., alcohol and/or drug abuse). To enhance ecological validity, we intentionally included participants with comorbidities or co-occurring mental health illnesses, as they are the rule rather than the exception in clinical practice. In a second phase, we collected mental health data from these screened individuals using standardized rating scales, free-text narratives, and open-response questions targeting nine common mental disorders. We recruited a small

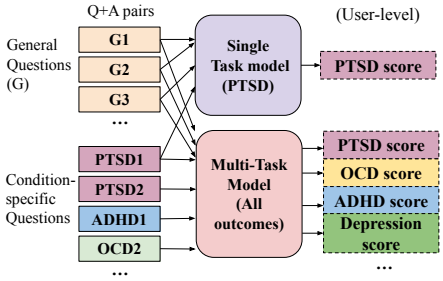

Figure 1: Single-task models are set up to predict a single mental health condition score based on language responses to the general questions (**GenQ**; questions in figure prefixed by **G**) as well as the condition-specific questions (**CondQ**; questions in the figure prefixed with **PTSD, ADHD, OCD** etc.). Multi-task models, on the other hand, are set up to take in all the language responses and predict all of the mental health scores simultaneously.

set of 515 participants (297 female, 182 male, 35 non-binary, 1 who preferred not to categorize their gender identity) who were diagnosed with any of nine mental health conditions (Anxiety Disorders (**AD**), Bipolar Disorder (**BD**), Depression (**D**), Attention Deficit Hyperactivity Disorder (**ADHD**), Post-Traumatic Stress Disorder (**PTSD**), Obsessive-Compulsive Disorder (**OCD**), Eating Disorders (**ED**), Addiction and Substance Abuse (**A**), Autism (**AU**)). The age of the sample ranged from 18 to 78 (*mean = 38.9*, *SD = 12.3*) We screened participants for having one (or more) ongoing mental disorders (∼50 of each disorder). The distribution of all the **diagnoses** across the participants is shown in Figure A.1. The participants first took a screening questionnaire for diagnoses and treatment for the ten mental health conditions to qualify for eligibility to participate in the survey. Here, 186 individuals reported receiving no treatment at all, while 329 reported receiving at least one form of treatment (i.e., medication and/ or psychotherapy).

All participants then took ten rating scale questionnaires along with **language-based questions** related to all the mental health conditions considered. A total of 48 language-response questions were developed based on DSM-5 criteria to capture key symptoms, frequency, and onset, and these items were reviewed by clinical psychologists to ensure clarity; 42 questions required descriptive-word responses and 5 essay responses. The dataset has two kinds of language questions posed to the participants: General (**GenQ**) and Condition-Specific (**CondQ**). There are 10 General Questions related to mental health in general, and all the other 38 questions are Condition-specific, mapping to the one of the questionnaire scales collected that we describe below.

Ten **validated clinical scales or mental health scores** were also administered: the PHQ-9 (Kroenke et al. (2001); 9 items, 4-point Likert) for depression, GAD-7 (Spitzer et al. (2006); 7 items, 4-point Likert) for anxiety, MDQ (Miller et al. (2004); 14 binary items plus one 4-point Likert item) for bipolar disorder, RAADS-14 (Eriksson et al. (2013); 14 items, 4-point Likert) for autism, ASRS Part A (Adler et al. (2006); 6 items, 5-point Likert) for ADHD, NSESSS-PTSD (LeBeau et al. (2014); 9 items, 5-point Likert plus an open-text trauma description) for PTSD, BOCS (Goodman et al. (1989); 15 items, 3-point Likert plus an open-response categorization) for obsessive-compulsive symptoms, EDE-QS (Fairburn and Beglin (2008); 12 items, 4-point Likert) for eating disorders, and two substance use instruments: the AUDIT (Allen et al. (1997); 8 items, 5-point Likert plus 2 items, 3-point Likert) for alcohol misuse and the DUDIT (Berman et al. (2007); 9 items, 5-point Likert plus 2 items, 3-point Likert) for drug abuse. Post-screening, we invited the selected participants to participate in the main study, with recruitment conducted via Prolific. The language questions, some eliciting descriptive words and some open-ended essay-like responses, are listed in Table A.2. They were asked in random order to eliminate any systematic priming effects.

[2]Detailed background on multidimensional IRT is provided in the Table A.1.

4 Methods

We begin by exploring robustness of multidimensional models to estimate mental health trait scores from language responses. Next, we detail our application of factor analysis to uncover the underlying trait structure and map items to their respective dimensions. Finally, we describe the adaptive testing approach which leverages this factor structure within a multidimensional IRT framework to guide item selection, response scoring, and iterative trait estimation for efficient and precise multidimensional mental health assessment (See Algorithm 2).

4.1 Multi-outcome Modeling

Given the multitude of psychological dimensions, associated language data, and the occurrence of comorbid diagnoses, linguistic expressions intended to capture one dimension may also provide valuable information about others. To investigate this, we frame the language-based modeling of mental health scores in two configurations: single-task and multi-task. Using a stratified sampling approach based on depression outcomes (PHQ scores) over the users, we generate 9 folds for evaluation, 5 as train set, 3 as development set and 1 test set. Each fold contains the responses of around 58 users each. Language representations are derived from all text responses utilizing an encoder model. We experimented with popular models as shown in Table 4, and for the rest of the experiments, we used the `nomic-embed-text-v1.5` model (Nussbaum et al., 2024), which are subsequently reduced to 16 dimensions through Matryoshka embeddings (Kusupati et al., 2022). We then train linear regression models to predict standard mental health questionnaire scores across ten dimensions (including PHQ for depression, GAD for anxiety etc. as explained in §3). The single-task models predict each mental health dimension independently, whereas the multi-task models simultaneously predict all ten scores (see Figure 1).

For model optimization, we performed hyperparameter tuning over ranges of learning rates, weight decay values, and used the Adam optimizer. We also compared output scaling methods for the regression, finding that min-max scaling consistently outperformed z-score normalization across all settings. Experiments were conducted using a single NVIDIA A6000 GPU. Further, we also prompt `llama3.2-1B` model (Grattafiori et al., 2024) to compare the capacilities of zero-shot models in estimating the mental health condition scores (A.3). We also explore several model variants to analyze different aspects of the language data.

Encoder-based Multi-outcome Modeling We compare various encoder models for multi-outcome modeling, the results are shown in Table 4. We find that Matryoshka reduction on the number of dimensions helps with modeling the less represented conditions, across all the conditions other than Depression and Anxiety.

Algorithm 2 MAQuA: Multi-Adaptive Question-Asking

Multi-outcome Modeling: Multi-outcome regression models to capture mental health scores from language
Apply threshold-based discretization to transform continuous or modeled scores into discrete item-level responses suitable for factor analysis
Factor Analysis: Using discretized response data $\mathbf{X}$, estimate factor loading matrix $\mathbf{\Lambda}$ and latent factor scores $\mathbf{f}_i$ for all individuals i
Determine number of factors m, factor correlations, and item-to-factor structure from $\mathbf{\Lambda}$
Multidimensional IRT-based Adaptive Question Asking:
  Initialize MIRT parameters $\{a_j, b_j, \mathbf{w}_j\}_{j=1}^{p}$ for each item j based on $\mathbf{\Lambda}$ and factor structure
  Set initial latent trait estimates $\boldsymbol{\theta}_i^{(0)}$ for each individual i
Initialize item prompt pool and `responses` $\leftarrow \emptyset$
while stopping criteria not met on $\boldsymbol{\theta}_i^{(t)}$ and pool not empty **do**
  For each candidate item p, compute Fisher information matrix $\mathcal{I}_p(\boldsymbol{\theta}_i^{(t)})$
  Select next item p^* maximizing $\det(\mathcal{I}_{p^*}(\boldsymbol{\theta}_i^{(t)}))$ over all candidates (D-optimality)
  Present prompt p to individual and capture text response t
  Compute multi-outcome discrete response score $\mathbf{s} =$ NLP_discretize_score(t) aligned with MIRT model item format
  Append $(p, \mathbf{s})$ to `responses`
  Update $\boldsymbol{\theta}_i^{(t+1)}$ using maximum likelihood on `responses`
  $t \leftarrow t + 1$
end while
Output: Final multidimensional trait estimates $\boldsymbol{\theta}_i^{(t)}$ for individual i

1. Aggregation type To aggregate multiple language responses from each participant, we consider two main approaches. The first, **input aggregation**, involves averaging the embeddings of all input responses for a user and then using this combined representation to predict an overall mental health score at the user level. The second approach, **output aggregation**, treats each language response separately: a model predicts a mental health score for each response, and then these scores are combined by averaging or another method for all questions related to the same condition to produce a final predicted score for that condition. This allows us to compare whether it is more effective to aggregate at the language level or the level of model predictions.

2. Question Information We explore the role of the language of question wording on the modeling of mental health outcomes. Unlike conventional language-based assessments that rely on ecological data such as social media posts, our setting is distinct because the language analyzed is generated as direct responses to specific prompts. To understand whether the phrasing of the questions themselves affects the models, we conduct an ablation study that incorporates the question ID as

Model	Aggr. type	Depression	Anxiety	Bipolar	Autism	Drug use	OCD	ADHD	PTSD	Eating	Alcohol use	Avg.
		Pearson Correlation with Validated Clinical Scales										
LLaMA-3.2-1B (zero-shot)		.179	.209	-.023	.015	.013	-.026	.019	.049	.007	.087	.052
LLaMa-3.1-8B (zero-shot)		.703	.590	.242	.368	.179	.512	.384	.452	.514	.495	.444
Qwen3-30B (zero-shot)		.754	.678	.503	**.545**	-	.427	-	.244	**.570**	-	.372
Single Task	Input	.763	.699	.398	.399	.351	.499	.424	.493	.409	.412	.485
Single Task	Output	.775	.703	.372	.425	.304	.569	.497	.468	.330	.355	.479
Multi Task	Input	**.784**	**.722**	**.446**	**.449**	**.419**	**.570**	**.560**	**.532**	**.468**	**.478**	**.543**
Multi Task	Output	.433	.443	.401	.307	.394	.408	.380	.366	.387	.411	.389
		Pointwise Biserial Correlation with Self-Reported Diagnoses										
		Depression	Anxiety	Bipolar	Autism	Substance use	OCD	ADHD	PTSD	Eating	-	Avg.
Clinical Scales (upper-bound)		.404	.423	.440	.454	.097	.133	.073	.172	.080	-	.253
LLaMA-3.2-1B (zero-shot)		.032	.104	.036	034	.065	-.005	.029	.071	.069	-	.048
LLaMa-3.1-8B (zero-shot)		**.423**	**.429**	.210	.237	.439	.183	.258	**.306**	.336	-	.313
Qwen3-30B (zero-shot)		.409	.466	**.451**	**.308**	-	**.342**	-	.178	**.367**	-	.279
Single Task	Input	.346	.333	.220	.036	.165	.136	.081	.160	.078	.	.173
Single Task	Output	**.389**	.393	.135	.170	.123	.146	.179	**.269**	.037	.	.205
Multi Task	Input	**.388**	**.428**	**.244**	**.218**	**.269**	**.173**	**.195**	.249	.108	-	**.252**
Multi Task	Output	.379	.415	.228	.139	.190	.127	.182	.239	**.149**	-	.227

Table 1: Comparison of aggregation strategies and task formulations for predicting multiple psychological scores from language. We report Pearson correlation with validated clinical scales (upper) and point-biserial correlation with a looser criteria - self-reported diagnostic labels (lower). **Bold** indicates statistically significant improvement ($p < 0.05$) over the second-best model in the same column, based on a paired t-test.

	with Q (Lang)				Ablation	
	GenQ	CondQ	CondQ + GenQ	All Q	With Q (ID)	No Q
Depression	.782	.785	.785	.784	.795	.792
Anxiety	.721	.724	.723	.722	.725	.724
Bipolar	.440	.440	.432	**.446**	.394	.390
Autism	.450	.451	.444	.449	.423	.429
Drug use	.423	**.431**	.423	.419	.324	.272
OCD	.566	**.574**	.570	.570	.561	.568
ADHD	.558	**.560**	.554	**.560**	.532	.532
PTSD	.534	.538	.536	.532	.490	.493
Eating	.458	.463	.459	**.468**	.416	.335
Alcohol use	.469	.474	.465	**.478**	.374	.359
Average	.540	**.544**	.539	**.543**	.503	.489

Table 2: Ablation study examining the effect of question inclusion and representation in multi-outcome modeling. We compare models trained with different subsets of questions and evaluate the impact of removing question embeddings or replacing them with question IDs. **Bold** indicates statistical significance as defined in Table 1.

Mental Health Condition	Factor 1 Loading	Factor 2 Loading	Dominant Factor
Depression	.908	.210	1
Anxiety	.953	.198	1
Bipolar	.779	.546	1,2
Autism	.861	.063	1
Substance use	.305	.870	2
OCD	.945	.240	1
ADHD	.918	.274	1
PTSD	.716	.430	1,2
Eating disorder	.091	.928	2
Alcohol use	.672	.418	1,2

Table 3: Exploratory factor analysis results for mental health conditions. The dominant factor indicates which factor has the highest loading for each condition. Bipolar, PTSD, and alcohol use disorder are modeled as cross-loadings.

Model	Number of dim	Depression	Anxiety	Bipolar	Autism	Substance use	OCD	ADHD	PTSD	Eating	Alcohol use	Avg.
nomic-embed-text-v1	16	**.784**	**.722**	.446	.449	**.419**	**.570**	**.560**	**.532**	**.468**	**.478**	**.543**
nomic-embed-text-v1	768	.763	.688	.442	**.487**	.239	.426	.403	.265	.270	.232	.421
mxbai-embed-large-v1	1024	.758	.668	**.480**	.484	.246	.434	.406	.240	.298	.150	.417
roberta-base	768	.733	.658	.394	.428	.148	.441	.298	.234	.252	.231	.382
roberta-large	1024	.780	.675	.385	.449	.210	.429	.418	.269	.274	.225	.411

Table 4: Comparison of popular embeddings (Nussbaum et al., 2024; Lee et al., 2024; Liu et al., 2019) derived from encoder models, for multitask, input-aggregated setting. While any of these models could be used with our framework, we find Matryoshka reduction and sentence-embedding formulation to be particularly helpful in modeling in low-resource settings.

an input feature. This allows us to disentangle whether it is the unique identity of the question, rather than its linguistic content, that primarily drives the modeling performance.

We report the Pearson correlation of the predicted scores against the validated clinical scales they were originally trained on. Further, we also report the Pointwise Biserial correaltion of the predicted scores against each of the nine diagnoses collected (binary-valued). The results are shown in Tables 1, 2. Together, Tables 1, 2 demonstrate that multi-outcome modeling substantially improves prediction in data-scarce settings, while ablation results confirm that gains are driven primarily by shared linguistic information rather than question-specific identifiers. The best model determined was then used to train across 9 folds with 4 folds as the regression task train set, 4 folds as MIRT train set and 1 fold for MIRT test set, and MAQuA was run on a 9-fold cross validation, reporting the aggregated scores across all the test sets. This design choice was made to ensure enough training data for both the multi-outcome modeling as well as the MIRT modeling.

4.2 Factor Analysis

Multidimensional adaptive question-asking algorithm requires setting up a factor model, which is defined through factor analysis to determine the factor structure, that takes in all the questions. We run exploratory factor analysis to determine the optimal factor structure for this dataset for the predicted user-level scores. The factor loadings are reported in Table 3 and Figure A.2. Then we run aggregate the multi-outcome model predictions at a question level, across all users, to analyze how each question loads on to the factors by applying the factor model at the question-level. This yields a question-level loading which is used to define the MIRT model (Table 5). We find that two significant factors emerged

Factor	Condition / Symptom
F1	Depression (OMD), Anxiety (A), Bipolar Disorder (BD), Autism Spectrum Disorder (ASD), Obsessive-Compulsive Disorder (OCD), Attention Deficit Hyperactivity Disorder (ADHD), Post-Traumatic Stress Disorder (PTSD), Eating Disorders (ED)
F2	Bipolar Disorder (BD), Substance Use (SUB), Obsessive-Compulsive Disorder (OCD), Eating Disorders (ED)

Table 5: Question-level mapping to the two significant factors that was further used to specify the MIRT model (see Table B for legend).

based on parallel analysis (Horn, 1965). The first factor is characterized by strong loadings from measures of depression, anxiety, PTSD, ADHD, and autism, suggesting that it reflects a broad *internalizing* or emotional distress dimension. The second factor is dominated by high loadings for drug and alcohol use, pointing toward a substance use or *externalizing* dimension. Conditions such as bipolar disorder, PTSD, and alcohol use disorders exhibit notable loadings on both factors (within 20% of each other), indicating that they share features with both internalizing and externalizing constructs.

4.3 Adaptive Testing

We use the `mirtCAT`[3], a computerized adaptive testing framework based on `mirt` to implement adaptive testing. The adaptive testing is done by choosing the most informative question at each turn, determined using D-optimality, a heuristic determined to be optimal for multidimensional IRT in (Chalmers, 2016). It maximises the determinant of the Fisher information matrix at each turn of the questions, thus maximizing information gain across all the underlying dimensions.

Discretization Since IRT models operate on discrete, ordered responses, continuous scores obtained from regression cannot be directly used as inputs. Instead, these continuous predictions must be discretized into k ordinal levels to allow the model to estimate category response functions. This discretization enables the IRT model to capture how each level probabilistically relates to underlying latent dimensions, following standard practices in multidimensional IRT modeling (Reckase, 2006). We select the maximum number of discrete levels that ensures sufficient representation across the training data, following the heuristics in prior work on discretizing language-based scores for IRT (Varadarajan et al., 2024b). We discretize scores into four ordinal levels based on the quartile thresholds of the model outputs. Using greater granularity leads to data sparsity, as certain levels would be underrepresented, reducing model stability and interpretability.

We compare random question ordering to adaptive question-asking using MAQuA, and further benchmark these against GPT-4, one of the most popular LLMs previously studied for its mental health estimation capabilities and ability to handle long contexts (Moëll, 2024; OpenAI, 2023; Ganesan et al., 2024). At each step, we report the Pearson correlations between estimated scores and validated clinical scales. The prompt can be found in Table A.4.

Stabilization points We compare the ordering and estimation capabilities of `gpt4` to showcase the efficacy of IRT in adaptive question-asking. To do this, we formalize a metric that summarizes the stability of estimation at each turn of question-asking: the rolling standard deviation of the estimates at each turn of question-asking. We set the threshold at 0.01 to determine the point at which the estimated mental health scores from MAQuA for each dimension mostly stabilizes and approaches convergence. This metric can also help determine early stopping during deployment MAQuA (Figure 2).

5 Results

Multi-outcome Modeling The results in Table 1 compare the predictive performance of single-task and multi-task models, as well as different aggregation strategies, across ten psychological dimensions. Multi-task models generally achieved higher correlations than single-task models when using input-level aggregation, with a strong performance across all the ten dimensions averaging at a correlation of 0.543 against the validated clinical scales and 0.252 against diagnoses.

[3] https://CRAN.R-project.org/package=mirtCAT

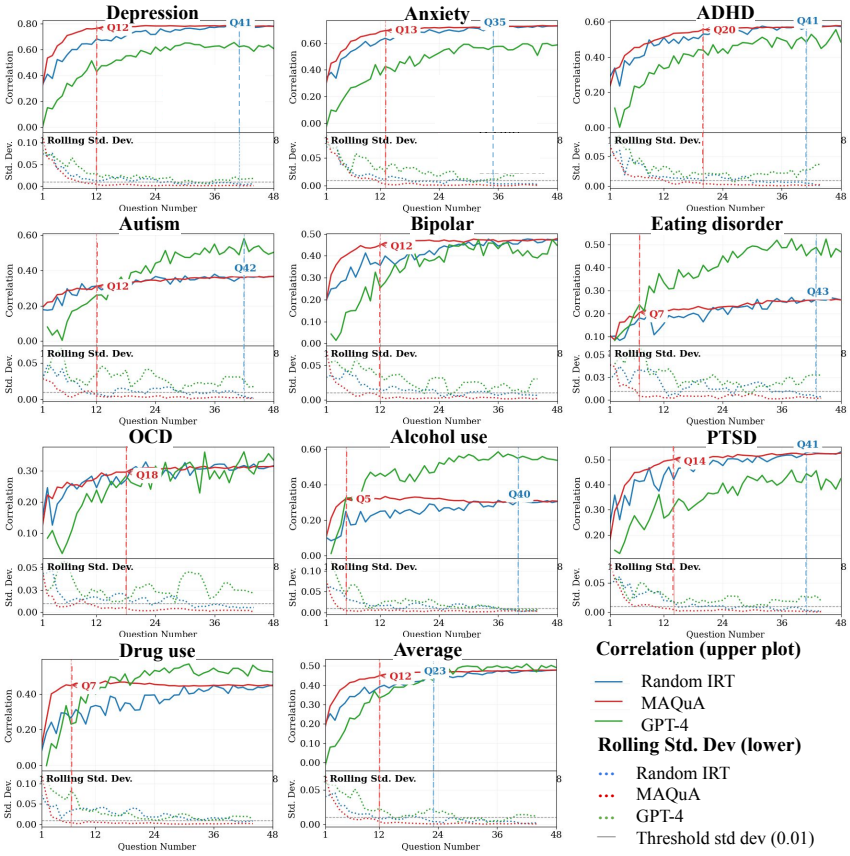

Figure 2: Pearson correlations of MAQUA-estimated scores over the number of questions asked along with their rolling standard deviation of the correlations. The vertical line shows the stability of the estimation based on a threshold for the standard deviation. Our adaptive method consistently stabilizes in at most 50% the number of questions as random. While GPT-4 shows promise in its estimation capabilities of factor 2 (externalizing disorders) - substance use, alcohol use, eating disorder.

In the case of single-task models, output-level aggregation performed better over input-level aggregation, presumably due to missing data points due to skipped responses for the less common conditions. Surprisingly, multi-task model with output aggregation performs the worst across all the dimensions; this could be because of all the individual questions not being relevant to most of the dimensions at the same time, forcing spurious correlations to be meaningful signals. In general, all models performed higher on internalizing (depression, anxiety, OCD) factors as opposed to externalizing (substance use). Most interestingly, we found that with respect to self-reported diagnoses, the models performed better than some of the validated clinical scales.

The utility of encoder-based models for mental health score estimation is further supported by the performance of LLM inferences: the inferred scores from the language responses of the participants are almost always outperformed by encoder-based multi-outcome models, when compared to the validated clinical scales. However, Qwen3-30B (Yang et al., 2025) in particular has high associations with the validated scales, indicating that larger LLMs could have stronger mental health inference capabilities. On the other hand, the associations with self-reported diagnoses, especially for LLaMa-3.1-8B (Grattafiori et al., 2024) and Qwen3-30B models consistently match or surpass associations with the well-established, validated clinical scales. These LLMs are also prone to mistakes: they miss inferences on some conditions, or their outputs are ill-formatted and difficult to parse, and the reported correlations are calculaed on valild outputs only. However, this result on surpassing clinical scales is especially notable as it potentially indicates that LLM inferences could complement existing mental health inference pipelines.

Table 2 shows that including all the questions, General (GenQ) as well as Condition-specific (CondQ), alongside non-Condition-specific questions (i.e., the condition-specific questions that are related to other conditions and not the considered condition) performs as well as using just Condition-specific questions. In particular, drug use, OCD and ADHD are best captured with Condition-specific questions alone. Further, on performing ablation with the language of the question, we find that the questions, including just the question ID, indeed add context to modeling multiple outcomes at once.

Condition	1	8	16	24	32	48	Ques. to stabilize
Depression							
GPT-4	-.001	.375	.518	.618	.637	.608	> 48
Random	.327	.642	.686	.759	.753	.778	42
MAQUA	.324	**.743**	**.772**	**.782**	**.783**	.781	12 (71%↓)
Anxiety							
GPT-4	-.029	.293	.461	.527	.565	.587	> 48
Random	.332	.612	.647	.690	.694	.730	34
MAQUA	.301	**.650**	**.705**	**.716**	**.723**	.727	13 (62%↓)
ADHD							
GPT-4	-	.248	.419	.447	.513	.486	> 48
Random	.282	.490	.497	.568	.569	.579	41
MAQUA	.235	.482	**.540**	.561	.563	.576	17 (56%↓)
Autism							
GPT-4	-	.180	.324	.441	.496	.502	> 48
Random	.184	.320	.298	.361	.354	.360	43
MAQUA	.194	.295	.329	.347	.360	.366	6 (86%↓)
Bipolar							
GPT-4	-	.195	.323	.398	.474	.448	> 48
Random	.202	.420	.390	.452	.454	.473	> 48
MAQUA	.201	**.440**	**.467**	**.470**	**.472**	.473	12 (75%↓)
Eating Disorder							
GPT-4	-	.213	.358	.451	.479	.468	> 48
Random	.096	.197	.199	.200	.223	.263	> 48
MAQUA	.116	**.218**	**.221**	**.229**	.249	.261	7 (85%↓)
OCD							
GPT-4	-	.163	.246	316	.307	.336	> 48
Random	.143	.253	.261	.278	.303	.318	39
MAQUA	.120	**.253**	**.300**	**.305**	**.312**	.318	19 (51%↓)
Alcohol Use							
GPT-4	-	.343	.425	.525	.535	.537	> 48
Random	.115	.191	.249	.269	.285	.309	40
MAQUA	.105	**.320**	**.327**	**.311**	**.301**	.307	5 (87%↓)
PTSD							
GPT-4	-	.220	.330	.407	.438	.424	> 48
Random	.236	.474	.468	.502	.499	.526	35
MAQUA	.176	.464	**.511**	**.513**	.517	.527	13 (63%↓)
Drug Use							
GPT-4	-	.240	.489	.530	.516	.521	> 48
Random	.078	.287	.309	.343	.430	.451	> 48
MAQUA	**.114**	**.458**	**.464**	**.447**	**.445**	.447	6 (87%↓)

Table 6: Random and MAQUA Scores by condition across the 48 questions, each averaged across 20 runs. The last column shows the number of questions it takes for stabilization.

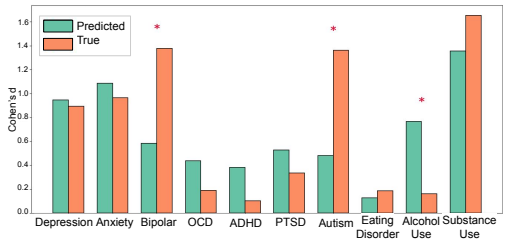

Figure 3: Cohen's d against the reported diagnoses for our best multitask model against the validated clinical scores (considered *ground truth* in the modeling). * indicates one being significantly better correlated to diagnosis than the other.

Multi-outcome Model Prediction Structure We show the difference between how the "*ground truth*" or validated clinical scores are related to diagnoses as compared to the predicted scores from the best performing multi-outcome model in Figure 3. Among ten conditions, only three conditions show significant differences – Bipolar, Autism and Alcohol use. While Bipolar and Autism are better captured by the original validated clinical scores, the predicted scores for Alcohol Use actually outperform the clinically validated scores (AUDIT), showing that alcohol use might be better captured when the other conditions are taken into account as well. Moreover, diagnoses could be preliminary and *wrong* (Mendel et al., 2011), since it could be a proxy for some other mental health condition. This could lead to spurious correlations that are better disambiguated with signals from a multitask model.

Adaptive Question-Asking Figure 2 shows the when the estimated scores from MAQUA are significantly better than their random counterpart, as well as GPT-4. After the first question, MAQUA seems to consistently jump to improve estimates across all dimensions, whereas in the case of random, the jumps are inconsistent. Based on the correlation curves presented in Figure 2, all the three methods can lead to a plateau in score estimation across most of the dimensions, suggesting that modeling multiple conditions can lead to the model learning the shared semantics across the conditions, regardless of the question-asking strategy used. MAQUA doesn't always converge better: in the case of ADHD, Autism and OCD, random question-asking is just as good as MAQUA in terms of when convergence (small change in estimated scores in subsewuent turns) or maximum performance is reached. Despite that, MAQUA outperforms random ordered question-asking, especially for conditions like bipolar, alcohol and drug use, depression, anxiety and eating disorder, especially in the first few turns (2-12). This suggests that both the factors, *internalizing* and *externalizing*, are being prioritized almost equally when optimizing for information gain over multiple turns of question-asking, which indicates that the chosen model is quite effective in adaptive question-asking while optimizing across ten distinct conditions.

In contrast, GPT-4's estimation and question-asking behavior differ. Despite being prompted to assess all ten conditions simultaneously (rather than focusing on two factor scores), GPT-4 notably excels over the IRT-based methods in handling externalizing factors: specifically eating disorder, substance use, and alcohol use conditions. This advantage may stem from limitations inherent in the item response theory model, particularly regarding its representation of comorbidity effects within each data subset. Evidence supporting this includes the regression model's performance, which approximates GPT-4's correlation levels (Table 1). These findings suggest that a hybrid modeling approach could more effectively capture both internalizing and externalizing aspects.

To determine early stopping criteria and when a question-asking session could be potentially shortened, we also report the stabilization points in Figure 2 by marking the question numbers. Stabilization does not necessarily indicate peak performance, it indicates

slower ascent after that point. The vertical lines indicate the point (n^{th} question) where the rolling standard deviation drops below a threshold. As reported in Table 6, employing the early stopping rule could lead to $50-85\%$ reduction in the number of questions across all the mental health conditions being evaluated simultaneously. Figure 2 shows stabilization points for individual and averaged correlation across all the ten conditions. We find that on an average, MAQUA takes about 12 questions to reach the stabilization point (threshold$= 0.01$) whereas random question-asking takes 24 and GPT-4 doesn't converge even at 48 questions. This offers significant potential to save time for both LLM-patient and clinician-patient interactions while reducing overall burden, and indicates that out-of-the-box, zero-shot LLMs, even if promising, might need to be complemented with principled, theoretically-derived methods for complex, sensitive mental health workflows.

6 Related Work

While large language models (LLMs) have shown promising results in zero-shot and few-shot prediction tasks (Ganesan et al., 2024; Hur et al., 2024), finetuned or instruction-tuned models remain generally more reliable and better validated across a range of mental health outcomes (Xu et al., 2024). Recent advances demonstrate that open-ended language responses to standardized questions can predict mental health scores with high accuracies: sometimes with correlations exceeding 0.8 with established clinical rating scales (Kjell et al., 2022; Varadarajan et al., 2024b; Sikström et al., 2023). These include pre-trained language models tailored for mental healthcare applications, such as ClinicalBERT and MentalBERT (Alsentzer et al., 2019; Ji et al., 2022).

Large language models (LLMs) commonly address referential and vague queries by employing targeted, selective prompts, which has been shown to improve answer accuracy and reduce errors (Zhang et al., 2025; Kuhn et al., 2023). However, despite these improvements, LLMs still fall short of human conversational subtlety and adaptiveness when it comes to clarification and follow-up questions. In terms of LLMs for mental health support, Rosenman et al. (2024) demonstrate that LLMs can effectively transform unstructured psychological interviews into structured questionnaires, enabling automated, multidimensional psychiatric evaluation, though reliability and consistency still require further improvement for clinical deployment. Complementing this, Nguyen et al. (2025) explore LLMs' ability to engage in mental health counseling, showing that, though LLMs can generate contextually relevant follow-up questions, they often lag behind human clinicians in empathy, specificity, and diagnostic nuance, and in crafting clarifying or probing questions that are crucial for effective counseling. Similarly, Yang et al. (2023) assess LLM performance across a spectrum of mental health tasks, finding that LLMs frequently overlook emotional cues or oversimplify questions, limiting their utility for nuanced clinical interpretation. This underscores the need for methods like ours that explicitly model multiple mental health factors to guide strategic question selection. Our framework enhances both the efficiency and accuracy of mental health assessments by optimizing inquiry and serving as a comprehensive diagnostic tool.

While adaptive testing has gained significant traction in educational settings, its multidimensional applications remain relatively underexplored, particularly when leveraging open-ended language responses. Most existing approaches in NLP rely on unidimensional item response theory (IRT) focusing on single outcomes (Lalor et al., 2016; Varadarajan et al., 2024b). To date, no prior work has effectively bridged this gap by integrating adaptive item selection with multitask learning for language-based, multidimensional mental health evaluation.

To our knowledge, this work is the first to tackle this challenge. MAQUA combines multi-outcome modeling with multidimensional IRT (MIRT) to adaptively select open-ended questions for assessing multiple mental health conditions at once. Our system uniquely models ten overlapping mental health constructs from targeted language data, learning to identify the most diagnostically informative questions while explicitly capturing their latent comorbid relationships.

7 Conclusion

This work presents MAQUA, a novel adaptive, language-based framework that enables efficient, simultaneous assessment of multiple mental health dimensions by leveraging the strengths of modern language modeling along with multidimensional item response theory. Our empirical findings demonstrate that multi-task modeling with both shared and unique linguistic features significantly improves predictive accuracy across ten distinct mental health outcomes compared to single-task baselines. Moreover, by integrating adaptive question selection optimized for information gain across multiple dimensions, MAQUA substantially reduces the number of questions required to achieve stable diagnostic estimates, cutting patient burden by up to 85% without compromising accuracy.

These results highlight the potential in LLM mental health agents for combining advanced language understanding with adaptive testing to overcome limitations of prior approaches that face challenges related to the inconsistency of LLM inferences, and are not optimized to reduce patient burden. MAQUA's effectiveness in modeling transdiagnostic symptom profiles marks an important step toward scalable, interactive, and clinically valid mental health assessment tools. Future research should extend and test this framework on real-time conversational settings, improving generalizability in diverse clinical populations.

Limitations

This work has several important limitations. First, all participants provided responses in English and were primarily recruited from the UK, Sweden, and the US, which may restrict the applicability of our findings to other languages and cultural settings. Additionally, ADHD is underrepresented in the dataset, limiting the reliability of conclusions related to this condition.

We use a fixed set of questions to maintain control over the assessment and prevent large language models from hallucinating or generating harmful, misleading questions. Although this limits open-ended question selection, it is an important safeguard for participant safety and assessment reliability. Future work will explore flexible approaches with robust hallucination mitigation strategies. While the multitask capabilities of large language models (LLMs) are critical, they were not explicitly explored in this study. Specifically, the effects of reinforcement learning from human feedback (RLHF) or direct preference optimization (DPO) on question sequencing and preferences are unexplored and warrant further investigation.

Although our multi-outcome models for mental health assessment are not fully accurate for clinical diagnosis, we proceed with modeling downstream question-asking since it more closely mirrors how such models would be deployed in real-life settings. However, these models have not been tested in actual clinical environments and should not be used for diagnosis. They are instead intended as screening tools that may complement therapists and clinicians within their processes.

Ethical Considerations

We acknowledge that we used AI assistants for paraphrasing and grammar checks, but the conception of ideas, experimentation and writing was done by the authors. All the generated text was carefully reviewed by the authors.

Participants were compensated hourly, based on the time spent on the questionnaire, leading to a payment ranging from £6.29 to £9.84, ensuring fair recognition of their contribution and respecting the principle of voluntary and informed participation. All responses were anonymized to protect participant confidentiality, and the data were stored securely on protected servers to safeguard privacy and comply with data protection standards. The study received approval from the Lund University Institutional Review Board (IRB), ensuring that appropriate ethical oversight was in place throughout the research process.

MAQuA is designed to reduce the burden of mental health assessments and address diagnostic inconsistencies, thereby promoting ethical principles of minimizing participant fatigue and improving assessment accuracy. By combining multi-outcome modeling with multidimensional item response theory (MIRT), MAQuA selects the most informative questions adaptively, reducing assessment length by up to 85% while maintaining psychometric validity, which supports participant well-being and respects their time.

MAQuA represents a methodological integration of advanced natural language processing and psychometric measurement theory, grounded in ethical commitments to scientific integrity, transparency, and reproducibility. The accompanying dataset is intended to encourage further validation and responsible innovation, ensuring that AI tools in mental health are developed with accountability and sustained patient trust, and should not be used to develop models to diagnose without thorough clinical supervision.

Acknowledgements

This work was supported in part by a grant from the NIH-NIAAA (R01 AA028032) awarded to H. Andrew Schwartz at Stony Brook University and a grant from the CDC/NIOSH (U01 OH012476). The conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, NIH, NIOSH, CDC, any other government organization, or the U.S. Government.

References

Lenard A Adler, Thomas Spencer, Stephen V Faraone, Ronald C Kessler, Mary J Howes, Joseph Biederman, and Kristina Secnik. 2006. Validity of pilot adult adhd self-report scale (asrs) to rate adult adhd symptoms. *Annals of Clinical Psychiatry*, 18(3):145–148.

John P Allen, Raye Z Litten, Joanne B Fertig, and Thomas Babor. 1997. A review of research on the alcohol use disorders identification test (audit). *Alcoholism: clinical and experimental research*, 21(4):613–619.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Anne H Berman, Hans Bergman, Tom Palmstierna, and Frans Schlyter. 2007. Dudit. *The Drug Use Disorders Identification Test–E. MANUAL. Karolinska institutet, Stockholm*.

R Philip Chalmers. 2012. mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, 48:1–29.

R Philip Chalmers. 2016. Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, 71:1–38.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology:*

From Linguistic Signal to Clinical Reality, pages 31–39. Association for Computational Linguistics.

Robert Cudeck and Robert C MacCallum. 2000. *Exploratory Factor Analysis*. Lawrence Erlbaum Associates, Mahwah, NJ.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Susan E Embretson and Steven P Reise. 2000. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Mahwah, NJ.

Jonna M Eriksson, Lisa MJ Andersen, and Susanne Bejerot. 2013. Raads-14 screen: validity of a screening tool for autism spectrum disorder in an adult psychiatric population. *Molecular Autism*, 4:1–11.

Christopher G Fairburn and Sarah J Beglin. 2008. Eating disorder examination questionnaire. *Cognitive behavior therapy and eating disorders*, 309:313.

Adithya V Ganesan, Vasudha Varadarajan, Yash Kumar Lal, Veerle C Eijsbroek, Katarina Kjell, Oscar NE Kjell, Tanuja Dhanasekaran, Elizabeth C Stade, Johannes C Eichstaedt, Ryan L Boyd, and 1 others. 2024. Explaining gpt-4's schema of depression using machine behavior analysis. *arXiv preprint arXiv:2411.13800*.

Wayne K Goodman, Lawrence H Price, Steven A Rasmussen, Carolyn Mazure, Roberta L Fleischmann, Candy L Hill, George R Heninger, and Dennis S Charney. 1989. The yale-brown obsessive compulsive scale: I. development, use, and reliability. *Archives of general psychiatry*, 46(11):1006–1011.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. 1991. *Fundamentals of Item Response Theory*. Sage Publications, Newbury Park, CA.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3774–3788.

John L Horn. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.

Jihyun K Hur, Joseph Heffner, Gloria W Feng, Jutta Joormann, and Robb B Rutledge. 2024. Language sentiment predicts changes in depressive symptoms. *Proceedings of the National Academy of Sciences*, 121(39):e2321321121.

Ian Andrew James, Rachel Morse, and Alan Howarth. 2010. The science and art of asking questions in cognitive therapy. *Behavioural and Cognitive Psychotherapy*, 38(1):83–93.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 7184–7190.

Seungwan Jin, Bogoan Kim, and Kyungsik Han. 2025. "I Don't Know Why I Should Use This App": Holistic Analysis on User Engagement Challenges in Mobile Mental Health. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Swanie Juhng, Matthew Matero, Vasudha Varadarajan, Johannes Eichstaedt, Adithya V Ganesan, and H Andrew Schwartz. 2023. Discourse-level representations can improve prediction of degree of anxiety. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1500–1511.

Oscar NE Kjell, Salvatore Giorgi, H Andrew Schwartz, Lyle H Ungar, David B Yaden, Margaret L Kern, and Johannes C Eichstaedt. 2022. Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific Reports*, 12(1):3918.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Clam: Selective clarification for ambiguous questions with generative language models. In *Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML)*. ArXiv preprint arXiv:2212.07769.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30233–30249.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.

Richard LeBeau, Emily Mischel, Heidi Resnick, Dean Kilpatrick, Matthew Friedman, and Michelle Craske. 2014. Dimensional assessment of posttraumatic stress disorder in dsm-5. *Psychiatry research*, 218(1-2):143–147.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Jonathan S Ilgen, Yulia Tsvetkov, and Maarten Sap. 2025. Aligning llms to ask good questions a case study in clinical reasoning. *arXiv preprint arXiv:2502.14860*.

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. *Preprint*, arXiv:1907.11692.

Syeda Mahwish, Ryan L. Boyd, Vasudha Varadarajan, Roman Kotov, Benjamin J. Luft, H. Andrew Schwartz, and Sean A. P. Clouston. 2026. Measuring resilience using language modeling: A computational approach to observing resilience. *Journal of Traumatic Stress*, pages 1–14.

Rosmarie Mendel, Eva Traut-Mattausch, Eva Jonas, Stefan Leucht, John M Kane, Katja Maino, Werner Kissling, and Johannes Hamann. 2011. Confirmation bias: why psychiatrists stick to wrong preliminary diagnoses. *Psychological medicine*, 41(12):2651–2659.

Christopher J Miller, Joshua Klugman, Debra A Berv, Kristen J Rosenquist, and S Nassir Ghaemi. 2004. The sensitivity and specificity of the Mood Disorder Questionnaire for detecting bipolar disorder. *Journal of Affective Disorders*, 81(2):167–171.

Birger Moëll. 2024. Comparing the efficacy of gpt-4 and chat-gpt in mental health care: A blind assessment of large language models for psychological support. *arXiv preprint arXiv:2405.09300*.

Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellayi Gopalakrishnan, Ekta Raj, Zihang Li, Heather J. Soled, Michael L. Birnbaum, Srijan Kumar, and Munmun De Choudhury. 2025. Do large language models align with core mental health counseling competencies? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7488–7511, Albuquerque, New Mexico. Association for Computational Linguistics.

Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

OpenAI. 2023. GPT-4 Technical Report. `https://openai.com/research/gpt-4`.

David Owen, Jose Camacho-Collados, and Luis Espinosa Anke. 2020. Towards preemptive detection of depression and anxiety in Twitter. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 82–89. Association for Computational Linguistics.

Mark D Reckase. 2006. 18 multidimensional item response theory. *Handbook of statistics*, 26:607–642.

Gony Rosenman, Talma Hendler, and Lior Wolf. 2024. LLM questionnaire completion for automatic psychiatric assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 403–415, Miami, Florida, USA. Association for Computational Linguistics.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31.

Chen Shani and Elizabeth Stade. 2025. Measuring mental health variables in computational research: Toward validated, dimensional, and transdiagnostic approaches. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 69–78, Albuquerque, New Mexico. Association for Computational Linguistics.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3838–3844.

Sverker Sikström, Oscar NE Kjell, Katarina Kjell, and Johannes Lundberg. 2023. Precise language models can estimate psychological states and abilities from text. *Scientific Reports*, 13(1):11679.

Khushboo Singh, Vasudha Varadarajan, Adithya V Ganesan, August Håkan Nilsson, Nikita Soni, Syeda Mahwish, Pranav Chitale, Ryan L. Boyd, Lyle Ungar, Richard N Rosenthal, and H. Schwartz. 2025. Systematic evaluation of auto-encoding and large language model representations for capturing author states and traits. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18955–18973, Vienna, Austria. Association for Computational Linguistics.

Nikita Soni, August Håkan Nilsson, Syeda Mahwish, Vasudha Varadarajan, H. Andrew Schwartz, and Ryan L. Boyd. 2025. Who we are, where we are: Mental health at the intersection of person, situation, and large language models. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 300–313, Albuquerque, New Mexico. Association for Computational Linguistics.

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097.

Vasudha Varadarajan, Allison Lahnala, Adithya V Ganesan, Gourab Dey, Siddharth Mangalik, Ana-Maria Bucur, Nikita Soni, Rajath Rao, Kevin Lanning, Isabella Vallejo, Lucie Flek, H. Andrew Schwartz, Charles

Welch, and Ryan Boyd. 2024a. 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, St. Julians, Malta. Association for Computational Linguistics.

Vasudha Varadarajan, Allison Lahnala, Sujeeth Vankudari, Akshay Raghavan, Scott Feltman, Syeda Mahwish, Camilo Ruggero, Roman Kotov, and H. Andrew Schwartz. 2025. Linking language-based distortion detection to mental health outcomes. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 62–68, Albuquerque, New Mexico. Association for Computational Linguistics.

Vasudha Varadarajan, Sverker Sikström, Oscar Kjell, and H. Andrew Schwartz. 2024b. ALBA: Adaptive Language-Based Assessments for Mental Health. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2466–2478, Mexico City, Mexico. Association for Computational Linguistics.

Charles Welch, Allison Lahnala, Vasudha Varadarajan, Lucie Flek, Rada Mihalcea, J Lomax Boyd, and João Sedoc. 2025. Isca: A framework for interview-style conversational agents. *arXiv preprint arXiv:2508.14344*.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Michael J.Q. Zhang, W. Bradley Knox, and Eunsol Choi. 2025. Modeling future conversation turns to teach llms to ask clarifying questions. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

A Details on Multidimensional IRT

A detailed description of the modeling of MIRT is described in Table A.1. Most of the terms and explanations were derived from Chalmers (2012).

B Dataset Details

Figure A.1 shows the distribution of diagnoses among all participants in the dataset, and Table A.2 lists the language-response questions. Because AUDIT and DUDIT assess overlapping symptoms, our language-based questions grouped alcohol and drug use under a broader substance abuse category, allowing participants to discuss their addictions more generally rather than focusing on alcohol or drugs in particular. Additionally, although ADHD is common in the general population, it is underrepresented in our dataset; many participants with an ADHD diagnosis dropped out before completion due to the survey's overall length (over 100 questions).

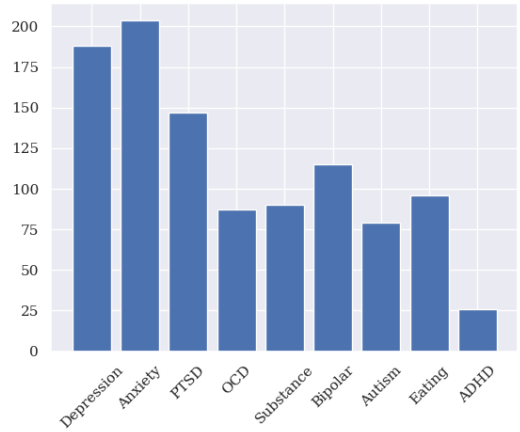

Figure A.1: The number of participants in the dataset that reported diagnosis for each of the conditions.

Figure A.2: Question texts loading on to the two factors.

C Question-level factors loading

After training the multi-outcome models, we found the multitask input aggregation model to perform the best. This model was then fed question-answer pair representations for all the training set as input, and the model inferred scores for each of the question-answer pair across all the users. These scores were then aggregated at a *question-level* for applying the factor analysis model derived on user-aggregated scores, to understand how much each question contributes to understanding the

Multidimensional IRT details	
Description	Multidimensional Item Response Theory (MIRT) extends classical unidimensional item response theory (IRT) to better capture complex psychological constructs by modeling multiple underlying dimensions of symptoms, rather than a single overall trait. This is especially important for mental health assessments, since symptoms often span affective, cognitive, and physical domains that interact and overlap, making a nuanced, multidimensional model necessary for accurately representing mental health scores.
Latent Trait Vector	In multidimensional item response theory (MIRT), the latent trait vector is defined as: $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$, where each component θ_k represents the individual's standing on the k-th latent dimension. This vector characterizes the respondent's abilities or traits across multiple correlated or independent dimensions. The latent traits $\boldsymbol{\theta}$ are typically assumed to follow a multivariate normal distribution: $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the mean vector (often 0) and $\boldsymbol{\Sigma}$ is the covariance matrix capturing correlations among the latent traits. This multivariate representation enables modeling the probability of a particular response to an item as a function of these multiple latent traits and item parameters in a probabilistic framework.
Item Parameters	Each item j in a multidimensional item response theory (MIRT) model is characterized by a discrimination vector: $\boldsymbol{a}_j = (a_{j1}, a_{j2}, \ldots, a_{jm})$, which specifies the sensitivity of item j to each of the m latent traits. In other words, each component a_{jk} represents how strongly item j relates to latent dimension k. For a single item, we set a_{jk} to be the same across all the thresholds of a polytomous (graded response or rating scale) model. Each item also has threshold or difficulty parameters, denoted as b_{jk}, which indicate the location along the latent dimension(s) where the item optimally differentiates between respondents with different trait levels. Since we use polytomous models, multiple thresholds b_{jk} are used to correspond to different response categories.
Model	The probability that an individual i with latent trait vector $\boldsymbol{\theta}_i$ responds correctly (or supports) item j is modeled by a multidimensional logistic function: $P(u_{ij} = 1 \mid \boldsymbol{\theta}_i, \boldsymbol{a}_j, d_j) = \frac{1}{1+\exp[(\boldsymbol{a}_j' \boldsymbol{\theta}_i + d_j)]}$, where $\boldsymbol{a}_j$ is the discrimination vector for item j, and d_j is the difficulty (threshold) parameter. For polytomous (ordinal) responses, MIRT generalizes the unidimensional graded response model by estimating the probability of responding in each category as the difference between adjacent category response functions. For item j with response categories $k = 1, \ldots, K$, the probability of responding in category k given latent traits $\boldsymbol{\theta}$ is modeled as: $P(Y_j = k \mid \boldsymbol{\theta}) = P(Y_j \geq k \mid \boldsymbol{\theta}) - P(Y_j \geq k+1 \mid \boldsymbol{\theta})$, where each $P(Y_j \geq k \mid \boldsymbol{\theta})$ is computed using a multidimensional logistic function involving the discrimination vector $\boldsymbol{a}_j$, latent trait vector $\boldsymbol{\theta}$, and category threshold parameters b_{jk}. This approach captures the ordered nature of responses while simultaneously considering multiple latent dimensions.
Estimation Algorithm	The learning of the item parameters in IRT is typically enabled through expectation-maximization algorithm. However, QMCEM is better than traditional EM for multidimensional IRT because it uses quasi-random (evenly sampled) sequences to approximate high-dimensional integrals, rather than relying on random sampling or standard numerical methods. This approach provides more even coverage of the multidimensional latent trait space, which reduces variance in the integral estimates needed for parameter updates. QMCEM typically converges faster and yields more accurate and stable parameter estimates in multidimensional settings. Standard EM used in single-factor IRT could be slower, less precise, and prone to instability in high dimensions due to the inefficiency and unevenness of random samples used for the needed integrations.

Table A.1: Details on MIRT model that was used for the experiments.

ID	Question Text	Response Type
A1	Describe your worries and their strength, in the past few weeks. Write 5 descriptive words. If this statement does not resonate with you, please type 'not relevant' in the first text box.	words
A3	Describe how your mood has influenced your behavior in the past few weeks. Write at least 3 descriptive words.	words
A4	Describe places or activities you have avoided due to anxiety. Write at least 3 descriptive words. If this statement does not resonate with you, please type 'not relevant' in the first text box.	words
ADHD1	Describe your attention during tasks or assignments. Think about your workplace or school. Write at least 3 descriptive words.	words
ADHD2	Describe activities of restlessness, impulsivity, and decisions you made without thinking it through. Write at least 3 descriptive words. If this statement does not resonate with you, please type 'not relevant' in the first text box.	words
ASD2	Describe your typical social interaction and the typical way of communication. Write at least 2 descriptive words.	words
ASD3	Describe situations where you are intensely focused on specific topics or activities, to the exclusion of others. Write at least 3 descriptive words. If this statement does not resonate with you, please type 'not relevant' in the first text box.	words
ASD4	Describe situations where your senses were particularly overwhelmed, or distressed. Write at least 3 descriptive words. If this statement does not resonate with you, or you do not commonly experience such situations, please type 'not relevant' in the first text box.	words
ASD5	Describe your daily routine in general terms, and feelings when this routine is changed. Write at least 3 descriptive words.	words
ASD6	Describe how you navigate, experience, and maintain social relationships. Write at least 3 descriptive words.	words
BD2	You experienced recurring cycle of mood swings, moving from highs to lows and back again. If so, can you share a timeline of when you experienced episodes of elevated mood followed by depressive episodes? If this statement does not resonate with you, please type 'not relevant' in the text box.	essay
BD3	Describe impulsive or risky behaviors you have been engaged in lately. Write at least 2 descriptive words. If this statement does not resonate with you, please type 'not relevant' in the first text box.	words
ED1	Describe your eating habits that differ from other people. Consider your last week. Write at least 3 words. If this statement does not resonate with you, please type 'not relevant' in the first text box.	words
ED2	Describe your thoughts about food. Write at least 2 words.	words
ED3	Describe your thoughts about your weight, shape, or appearance. Write at least 2 words.	words
ED4	Describe the control over your eating behavior and related feelings. Write at least 1 word.	words
ED5	Describe behaviors and emotions you relate to food. Write at least 1 word.	words
ED6	Describe the impact your eating behaviors have on your daily life and relationships. Write at least 1 word.	words

ID	Question Text	Response Type
G1	Describe your mental health in a paragraph. Write at least 300 words.	essay
G10	When did you first notice difficulties in relation to your mental health? (open response)	essay
G12	Describe how your emotions and social relations have been influenced by your mental health. Write at least 3 descriptive words. If this statement does not resonate with you, please type 'not relevant' in the first text box.	words
G2	Describe your mental health. Write 5 descriptive words.	words
G3	Describe how your mental health has influenced your behavior in the past few weeks. Write at least 2 descriptive words.	words
G4	Describe how your mental health has influenced your work performance in the past few weeks. Write at least 2 descriptive words.	words
G5	Describe how your body felt in the past few weeks. Think about physical symptoms that have relevance for you. Write at least 3 descriptive words.	words
G6	Describe things you have been unable to do, concentrate on, make decisions on, or carry out due to your mental health. Write at least 3 descriptive words. If this statement does not resonate with you, please type 'not relevant'.	words
G7	Describe how your mood has influenced your daily life, in the past few weeks. Write 3 descriptive words.	words
G8	Consider your main mental health symptoms, how long have you been experiencing them? (open response)	essay
G9	Describe how your attention and activity level have influenced your social relationships. Write at least 3 descriptive words.	words
G91	Describe how your attention and activity level have influenced your work. Write at least 3 descriptive words.	words
OCD1	Describe recurring thoughts you experienced, and their content, in the past few weeks. Write at least 3 descriptive words. If this statement does not resonate with you, please type 'not relevant' in the first text box.	words
OCD2	Describe actions or rituals that you felt compelled to perform repeatedly, in the past few weeks. Write at least 3 descriptive words. If this statement does not resonate with you, please type 'not relevant' in the first text box.	words
OCD3	Describe obsessive thoughts or compulsions that you attempted to resist. Write at least 3 descriptive words. If this statement does not resonate with you, please type 'not relevant' in the first text box.	words
OMD1	Describe your changes, if any, in your mood or emotions in the past few weeks. Write at least 2 descriptive words.	words
OMD2	Describe a persistent mood or emotions you experienced in the past few weeks. Write 5 descriptive words.	words

ID	Question Text	Response Type
OMD3	Describe your ability to enjoy things in the past few weeks. Write at least 2 descriptive words.	words
OMD4	Describe how your appetite has been lately. Write at least 1 descriptive word.	words
OMD5	Describe how your sleep has been lately. Write at least 1 descriptive word.	words
OMD6	Describe how your motivation and/or energy level has been lately. Write at least 2 descriptive words.	words
PTSD1	Describe impactful events you experienced and that are still influencing your life. Write a paragraph with at least 300 words.	essay
PTSD2	Describe impactful events you experienced and that are still influencing your life. Write 5 descriptive words.	words
PTSD3	Describe thoughts, memories, or dreams related to impactful events that are influencing your life. Write 5 descriptive words. If this statement does not resonate with you, please type 'not relevant' in the first text box.	words
PTSD4	What was the traumatic event? (open response)	essay
SUB1	List drugs or substances that you have used. Include alcohol in this list, if relevant. (open response)	essay
SUB2	Describe the circumstances under which you use substances. Write at least 2 words.	words
SUB3	Describe your thoughts, behavior, and feelings when you are not using substances that you typically use. Write at least 1 word.	words
SUB4	Describe social, educational, or occupational consequences you experienced due to your usage of substances. Write at least 1 word.	words
SUB5	Describe risky behavior that you engage in during your usage of substances. Write at least 3 descriptive words. If this statement does not resonate with you, please type 'not relevant' in the text box.	words
SUB6	Describe your tolerance level towards substances. Write at least 1 word.	words

Table A.2: The language response question that the participants responded to, along with the type of response and question code. The relevance of the question codes to specific factors is shown in Table 5.

two factors found to be significant, which is shown in Table 5, and visually shown as a wordcloud in Figure A.2. This was used to define the MIRT model for training and adaptive testing.

Carefully read the entire transcript, considering both explicit answers and patterns, emotional tone, and recurring themes throughout the conversation.
Rely exclusively on the transcript; do not infer information not present or speculate beyond what is written.
For each mental health dimension below, provide only the mental health scores requested, nothing more:

* A single numerical score, adhering to the questionnaire's scale (see below for ranges).
* If insufficient information is present to make an informed rating for a dimension, give a score of 0.

Mental health dimensions and corresponding standardized rating scales:

* Depression score: PHQ-9 (range: 0–27)
* Anxiety score: GAD-7 (range: 0–21)
* Mood_Disorder score: MDQ (range: 0–15)
* Obsessive_Compulsive Disorder score: Y-BOCS (range: 0–40)
* Autism score: RAADS (range: 0–42)
* Eating_Disorder score: EDE-QS (range: 0–72)
* PTSD score: NSESSS (range: 0–36)
* Alcohol_Use score: AUDIT (range: 0–40)
* Substance_Use score: DUDIT (range: 0–44)
* ADHD score: ASRS Part A (range: 0–24)

Table A.3: Prompt for score estimation with Llama3.2 1B model.

You are an expert conversational agent conducting a mental health assessment.
Your goal is to select, one at a time, the most informative question from a provided list to accurately score ten standardized mental health dimensions.
Use prior conversation context to guide your choices. After each patient response, reassess and select the next optimal question.
If no further questions will improve scoring or none remain, output final scores for each dimension.
Never answer the questions yourself—only select a question to ask or output the final scores.
<interview_history>

This section contains the interview history or conversational context so far.
Use this information to inform your next question selection.
If this is empty, you are at the start of the interview.
...
</interview_history>

<candidate_questions>
This section contains the list of questions left to be asked at the current turn.
Each question is labeled with a unique code and text.
Select the most informative question from this list to ask next, unless you determine that the assessment should stop.
...
</candidate_questions>

Output format is as below. Only output in this JSON format with the mentioned keys, don't output anything else:
{
question_code: a single question code from candidate_questions to ask next, or -1 if stopping,
depression_score: X/100,
anxiety_score: X/100,
mood_disorder_score: X/100,
ocd_score: X/100,
autism_score: X/100,
eating_disorder_score: X/100,
ptsd_score: X/100,
alcohol_use_score: X/100,
substance_use_score: X/100,
adhd_score: X/100
}

Table A.4: Adaptive Question-asking prompt for GPT-4