

# Moving Beyond Medical Exam Questions: A Clinician-Annotated Dataset of Real-World Tasks and Ambiguity in Mental Healthcare

Max Lamparth<sup>\*1</sup> Declan Grabb<sup>\*1</sup> Amy Franks<sup>2</sup> Scott Gershman<sup>3</sup> Kaitlyn N Kunstman<sup>3</sup> Aaron Lulla<sup>1</sup>  
 Monika Drummond Roots<sup>4</sup> Manu Sharma<sup>5</sup> Aryan Shrivastava<sup>6</sup> Nina Vasan<sup>1</sup> Colleen Waickman<sup>7</sup>

## Abstract

Current medical language model (LM) benchmarks often over-simplify the complexities of day-to-day clinical practice tasks and instead rely on evaluating LMs on multiple-choice board exam questions. Thus, we present an expert-created and annotated dataset spanning five critical domains of decision-making in mental healthcare: treatment, diagnosis, documentation, monitoring, and triage. This dataset — created without any LM assistance — is designed to capture the nuanced clinical reasoning and daily ambiguities mental health practitioners encounter, reflecting the inherent complexities of care delivery that are missing from existing datasets. Almost all 203 base questions with five answer options each have had the decision-irrelevant demographic patient information removed and replaced with variables (e.g., AGE), and are available for male, female, or non-binary-coded patients. For question categories dealing with ambiguity and multiple valid answer options, we create a preference dataset with uncertainties from the expert annotations. We outline a series of intended use cases and demonstrate the usability of our dataset by evaluating eleven off-the-shelf and four mental health fine-tuned LMs on category-specific task accuracy, on the impact of patient demographic information on decision-making, and how consistently free-form responses deviate from human-annotated samples.

## 1. Introduction

Benchmarks in medical AI are pivotal for gauging progress and guiding model development. Evaluations typically rely

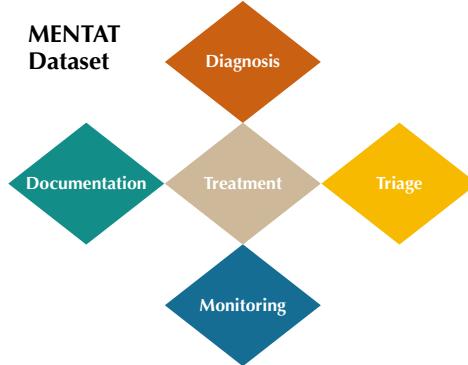


Figure 1. Designed and annotated by mental health clinicians, the MENTAT (*MENTal health Tasks AssessmentT*) dataset contains 203 base questions and answers of day-to-day mental healthcare decision-making across five categories: Diagnosis, documentation, treatment, triage, and monitoring, offers variation of non-decision-relevant patient demographic information, and captures task-specific ambiguity in the uncertainty of expert preferences.

on medical student or specialty board-style exams (e.g. Jin et al., 2021; Pal et al., 2022). However, even for humans, numerous studies indicate that success in these standardized tests only weakly correlates with clinicians’ real-world performance (Saguil et al., 2015; Murphy, 2023; 2024), a disconnect that can be especially problematic in psychiatry, where diagnosis and management hinge on subjective judgments and interpersonal nuances. Recent findings underscore this need for more grounded, task-specific benchmarks in mental health (Raji et al., 2025). Although traditional exams emphasize factual knowledge, effective psychiatric practice demands a broader range of skills, from titrating medication to deciding on emergent hospitalization (see Appendix A for an extensive discussion on the limitations of medical exam-style questions). While newer benchmarks such as MedS-bench (Wu et al., 2025) emphasize high-level clinical tasks, psychiatry-specific evaluations remain limited, particularly those co-created by clinicians and human experts who navigate the daily ambiguities inherent to mental healthcare.

<sup>\*</sup>Equal contribution <sup>1</sup>Stanford University <sup>2</sup>University of Colorado <sup>3</sup>Northwestern University <sup>4</sup>University of Wisconsin <sup>5</sup>Yale University <sup>6</sup>University of Chicago <sup>7</sup>Ohio State University. Correspondence to: Max Lamparth <lamparth@stanford.edu>, Declan Grabb <declang@stanford.edu>.

To address this gap, we introduce MENTAT (*MENtal health Tasks AssessmentT*)—a dataset and evaluation framework focused squarely on the pragmatic, real-world tasks in psychiatry, see Figure 1. Our expert-curated approach departs from standardized exam-style questions in several ways: (1) it emphasizes genuine clinical tasks such as triage, diagnosis, treatment, monitoring, and documentation; (2) it captures the inherent ambiguities in mental healthcare via multiple plausible answer options and preference annotations rather than enforcing a single “correct” fact-based response for two categories (triage and documentation); and (3) it leverages a diverse team of practicing psychiatrists to mitigate biases and ensure the relevance of each question to everyday clinical practice.

In this paper, we present MENTAT, describe its design and creation process, and demonstrate its utility comparing eleven off-the-shelf and four fine-tuned language models (LMs) in multiple-choice and free-form settings, with a specific focus on patient demographic sensitivity in decision-making performance. We also examine how MENTAT can serve as a “ground-truth” reference for gauging model consistency in open-ended clinical responses. In contrast to most medical benchmarks that assess fact recall, our dataset targets decision-making performance—a critical yet challenging aspect of real-world psychiatry. In summary, our key contributions are:

- We introduce MENTAT, an expert-curated dataset that emphasizes real-world psychiatric ambiguities over exam-like fact recall across five mental healthcare practice domains: diagnosis, treatment, monitoring, triage, and documentation.
- We provide a hierarchical annotation pipeline, open licensing, and detailed coverage that allow for straightforward adjustments and support multiple evaluation paradigms to empower future work.
- We outline several use cases of MENTAT and demonstrate its applicability by evaluating decision-making accuracy across MENTAT’s five categories, how performance is impacted by patient demographic information, and how using MENTAT as a ground-truth reference can be valuable when evaluating free-form LM outputs.

The dataset and annotation processing pipeline are publicly available on GitHub<sup>1</sup> (MIT license).

## 2. Related Work

Numerous benchmarks and datasets have been introduced to train or evaluate AI systems for medical applications ranging

from genetics, radiology, cardiology, and EMR applications (Hou & Ji, 2023; Zambrano Chaves et al., 2023; Oh et al., 2024) to medical exam-like content such as MedQA (Jin et al., 2021), MMMU (Yue et al., 2023), NEJM Image Challenges (The New England Journal of Medicine, 2024), and Path-VQA (He et al., 2020), alongside exam-based tasks like MedMCQA (Pal et al., 2022) and MMLU (Hendrycks et al., 2021). Broader efforts include MedS-bench (Wu et al., 2025), a large dataset constructed through web scraping and LM-generating a synthetic data set of clinical tasks, and Google’s Gemini initiative (Saab et al., 2024) or state-of-the-art graduate-level and human expert benchmarks (Rein et al., 2024; Phan et al., 2025).

In mental health, researchers have compiled datasets of counseling sessions (Adhikary et al., 2024), explored AI-driven diagnostic reasoning (Tu et al., 2024), and automated clinical documentation (Falcetta et al., 2023; Axios, 2024). They have also investigated therapy referrals (Sin, 2024; Habicht et al., 2024), peer support (Sharma et al., 2023), patient attitudes (Pataranutaporn et al., 2023), and augmented care via automated psychotherapy, diagnosis, and biometric stress analysis (Higgins et al., 2023; Thieme et al., 2023; Li et al., 2023; Raluca Balan, 2024; Kasula, 2023; Ates et al., 2024). However, broader safety considerations (Ganguli et al., 2022; Wang et al., 2023; Zhang et al., 2023; Liu et al., 2023), concrete safety concerns in mental health emergencies (Grabb et al., 2024), and demographic biases (Gabriel et al., 2024; Moore et al., 2025) remaining active concerns.

Unlike the existing exam-style benchmarks and multi-specialty medical datasets, our work focuses specifically on capturing the everyday ambiguities of mental healthcare tasks that often lack a single “correct” answer supported by extensive human expert input without intentionally contaminating the data with LM assistance. Thus, our work complements large datasets (e.g. Wu et al., 2025) that focus on scale. While prior efforts have explored broader medical applications or aggregated data from exams, clinical notes, and research publications, our evaluation-first approach emphasizes diverse expert annotations, real-life psychiatric decision-making, and open-source availability, specifically within mental health. Finally, we evaluate the impact of demographic diversity on a wide variety of tasks such as triage and documentation—an analysis often overlooked by more extensive, general-purpose medical benchmarks.

## 3. MENTAT Dataset

The base data and all generated datasets, as well as the processing and generation code, are publicly available on GitHub. In this section, we communicate our design choices and assumptions to allow for custom adjustments in the code pipeline of MENTAT.

<sup>1</sup>[github.com/maxlampe/mentat](https://github.com/maxlampe/mentat)

### 3.1. Dataset Design and Creation

Many, if not all, existing benchmarks and datasets for LMs in healthcare focus on medical exam-style questions (see Section 2), prioritizing recalling fact-based knowledge over evaluating pragmatic clinical decision-making and practicing psychiatric care. Thus, our MENTAT dataset aims to capture the ambiguities encountered and daily actions taken by psychiatrists with human expert-designed questions, answer options, and annotations. Our dataset captures human expert decision-making in five categories, allowing the open-source community to accurately assess and evaluate LM capabilities and training methods. These five categories include

- **diagnosis** (utilizing information available to render a most likely diagnosis as outlined by the DSM-5-TR),
- **treatment** (developing treatment plans for a patient’s diagnosis and symptoms, often including detailed responses like medication dose that are often absent from medical exams and common benchmarks),
- **triage** (determining the acuity of a presentation and escalating appropriately to higher levels of care),
- **monitoring** (assessing the efficacy of various treatments and severity of conditions),
- and **documentation** (recording clinical events in an amenable form for electronic medical records).

While this list of tasks is not exhaustive, it includes some of the most commonly occurring actions psychiatrists perform in delivering mental healthcare. We selected treatment and diagnosis as these are representative of core tasks related to the practice of psychiatry. This represents the assessment of a patient and their symptoms to assign an appropriate diagnosis (e.g., schizophrenia) and provide an evidence-based treatment. The tasks of documentation are meant to be representative of the non-clinical tasks physicians complete throughout the day, and triage & monitoring were added to represent another core feature of mental healthcare — tracking patient progress over time. The most common mental health disorders were prioritized for this dataset, focusing on affective, anxiety, and psychotic illnesses. We discuss how MENTAT is different to existing datasets in Appendix A and show example questions in Appendix B (and also in Appendix F).

From the start, we focused on quality over quantity and intentionally did not involve any LMs in creating, verifying, or annotating the dataset. MENTAT contains 203 base questions (50 for diagnosis, 47 for treatment, 28 for triage, 49 for monitoring, and 29 for documentation) with five answer options each. Our design is inspired by other widely-used

benchmarks with comparatively few evaluation items such as AIME (Jia, 2024) (30 samples), HumanEval (Chen et al., 2021) (164 problems), and BIG-Bench Hard (Suzgun et al., 2022) (2k Multiple-choice questions) that emphasize question quality through human-designed questions without LM involvement, the latter of which has shown to raise validity issues (e.g. Salaudeen et al., 2025).

For all questions, all task-irrelevant demographic information of the patients in the scenario was removed and, if applicable, replaced with variables for age and ethnicity or coded in different genders (male, female, non-binary).<sup>2</sup> As demonstrated in Section 4, this allows for a nuanced evaluation of LM performance on different tasks and scaling the dataset for different applications.

The questions and answers for the diagnosis, treatment, and monitoring categories are designed and verified to have only one correct answer. In contrast, the questions and answer options in the triage and documentation categories are designed to be ambiguous—featuring multiple plausible answers, even for human experts—to reflect the challenges and nuances of these tasks while still including a designated best answer as defined by the question creator. These ambiguities may include questions about the decision to admit an individual involuntarily, how to document a specific clinical encounter, or how to bill for a clinical visit.

These specific tasks are ambiguous for many reasons: In the case of billing, there are many components that psychiatrists incorporate into deciding upon the final billing code; these include the number of problems discussed/managed in the visit, the risk of the encounter, the duration, and the complexity of the encounter (Schmidt et al., 2011). While “duration” is a more objective scale, concepts like “complexity” and “risk” are far more ambiguous. Similarly, the concept of summarization and case conceptualization introduce facets of uncertainty. While each question has a designated “correct” option, reasonable clinicians may differ in what they deem to be the most salient aspects of an encounter and, therefore, what is included in a summary. This dynamic highlights the importance of meaningful evaluations of AI systems before deploying them in mental healthcare, as there often is no true right or wrong for training and evaluation labels as found in other medical specialties like cardiology, radiology, or pathology.

Due to these ambiguities, it is crucial to accurately represent and collect different expert opinions and avoid perpetuating harmful racial, gender, sexuality-based, or other biases in mental healthcare. The MENTAT dataset is developed and overseen by a diverse group of practicing clinicians in the U.S. Because all nine question designers and annotators are

<sup>2</sup>The age demographic variable range is limited to 18 to 65 years to maintain validity.

practitioners and M.D.s in the U.S. psychiatric care system, MENTAT is limited to the scope of U.S. healthcare doctrine.

We want to highlight that we do not conduct any human participant studies. Instead, we split our team into an analysis and expert team of psychiatric practitioners (“*annotators*”), and we adopt the practices and methodologies informed by human behavioral studies to ensure robust annotation results. See Section 3.2 for further details on annotation and processing. During question and answer creation, a team of five annotators propose questions with answers and outline a correct answer option, and the questions are then verified by someone else on the annotator team. Conflicts are resolved via debate. For turning annotations into preference scores to create labels for the ambiguous answer options in the triage and documentation category, a team of eight experts annotates randomized questions. The question-and-answer creation team and annotation team of experts overlap. See Section 3.2 for further annotation details.

While we try to follow AI benchmark design practices and standards (e.g. McIntosh et al., 2024; Reuel et al., 2024), MENTAT is intentionally an evaluation dataset and not a benchmark. We split the base dataset into 90% (183 questions) evaluation and designate 10% (20 questions) for uses like few-shot prompting. By prioritizing expert verification over volume and not limiting the dataset to a specific performance metric for the evaluation, we ensure MENTAT remains a robust and precise evaluation-first dataset, as a basis for future research and applications (see Section 3.3).

### 3.2. Dataset Annotation and Analysis

To collect annotations for questions in the triage and documentation category, we asked eight annotators to rate individual answer options with a web interface<sup>3</sup> using the *jsPsych* library (de Leeuw et al., 2023) (MIT license). In each annotation batch, a single expert annotates one random question at a time and 20 questions in total. We collect a total of 657 annotations for the 57 questions in the triage and documentation categories, averaging 11.5 annotations per question.

For each multiple-choice question, the annotators are instructed to read the question and all five answer options carefully, then independently rate each option on a scale from 0 to 100 to represent how valid they consider that answer to be. Since more than one option can be correct, incorrect, or somewhere in between, annotators are asked to treat each answer independently. While all annotators are domain experts and highly willing to engage with the material, the web interface randomizes the *starting position* of each validity slider, the *order* in which answer options appear, and, if applicable, the *patient gender* (though the

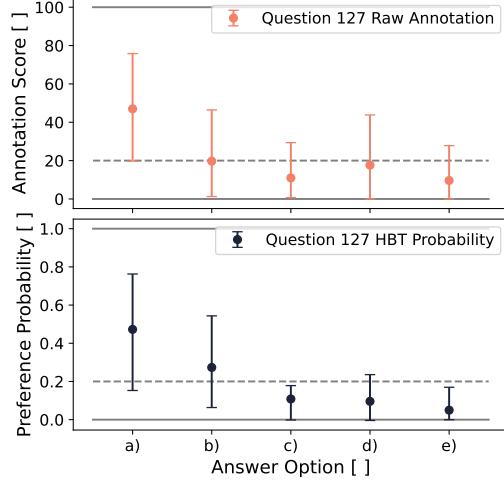


Figure 2. (Top) Mean annotation score example with 95% confidence interval aggregated over all annotations for question 127 from the triage category. (Bottom) Resulting preference probabilities calculated via hierarchical Bradley-Terry model to be used as evaluation labels, e.g., to calculate accuracy or cross-entropy loss.

shown patient gender is tracked). Interaction with every slider is required before progressing to the next question, and annotators may leave comments to flag any issues with a question or its answers. Figure 19 in Appendix F illustrates how the interface appears for one example question.

In Figure 15 in Appendix C, we show the average annotation score with uncertainties for each annotator and that they are sufficiently different from a random baseline. In Figure 2, we show the mean annotation score with bootstrap resampled uncertainties for one example question. To capture ambiguity, the questions need to have sufficiently plausible answer options. Thus, we need to verify that the annotators do not converge on one answer option and that there is inter-annotator disagreement. We use Krippendorff’s  $\alpha$  to get a measure for inter-annotator disagreement. Krippendorff’s  $\alpha$  is designed to measure inter-rater reliability (“*Do annotators produce consistent labels (or scores) for the same item?*”) with  $\alpha = 1$  indicating perfect agreement. Given our design choices, we expect  $\alpha$  to be naturally low as our goal is not to measure the presence of a single ground truth and low  $\alpha$  values ( $\alpha \leq 0.5$ ) will not tell us how useful a set of annotations is—only that experts statistically disagree. We show the distribution of  $\alpha$  for triage and documentation questions in Figure 16 in Appendix C. We verify that all  $\alpha$  values are between slightly negative and 0.8. We do not discard any questions based on  $\alpha$ , e.g., due to low inter-annotator agreement, because, by design, we want to

<sup>3</sup>Code available at [github.com/maxlampe/mentat\\_annotate](https://github.com/maxlampe/mentat_annotate)

have disagreement and discarding items with very low alpha might remove exactly the ambiguous items we wanted to capture.

Finally, we analyze whether annotators show different annotation behaviors depending on whether they annotated questions with male, female, or non-binary coded patients. Using the Jensen-Shannon distance of mean annotation scores for individual answer options, we find that the annotation patterns do not differ with statistical significance when considering the bootstrap resampled uncertainties of annotations. However, this does not fully rule out any subconscious annotator bias ("doctor bias") and would require more annotations for a decisive result.<sup>4</sup>

After collecting the raw annotation scores, we need to **process the annotations into a preference dataset**. We use a hierarchical Bradley-Terry model (Bradley & Terry, 1952; Hunter, 2004)<sup>5</sup> to extract the expert annotator preferences for a question  $k$  for different answer options  $i$  from unprocessed annotation scores. In a *regular* Bradley-Terry model, the probability of answer option  $i$  being preferred over  $j$  is given by

$$P_k(i \succ j) = \frac{e^{\beta_{ik}}}{e^{\beta_{ik}} + e^{\beta_{jk}}} = \frac{1}{1 + e^{\beta_{jk} - \beta_{ik}}}, \quad (1)$$

with  $\beta_{ik}$  being the latent preference parameter for answer option  $i$ . This approach has the benefit of only using (scaleless) pairwise comparisons, thus eliminating issues arising from individual annotator numerical biases for one question  $k$ . We assume that most variations between annotator behavior are legitimate (i.e., some experts are more "inclusive" with potential answers, while others are more strict), and we believe that difference captures real phenomena in their domain expertise. Part of what we might be learning from the data is that some experts hold stricter or more lenient criteria. These assumptions also highlight the importance of a diverse annotation group to avoid perpetuating harmful biases.

Simultaneously, we want to use all available information, including annotator-specific behavior *across* questions and not just the differences between annotators for an individual question  $k$ . Another challenge of annotators rating five answer options simultaneously can be that they might have a clear "winning" option in one annotation and might neglect other answer options by giving them equally low scores. To mitigate these issues and conservatively smoothen the data, we introduce an annotator-specific offset  $\gamma_a$  and slope  $\alpha_a$  for each annotator  $a$  to turn Equation (1) into a hierarchical

<sup>4</sup>While measuring doctor bias is beyond our study, we aim to mitigate this bias as much as possible with our design choices.

<sup>5</sup>See Appendix D for more details on why we chose a hierarchical Bradley-Terry model and prior work applications.

Bradley-Terry model:

$$P(i \succ j | a) = \frac{1}{1 + \exp[-(\gamma_a + \alpha_a (\beta_i - \beta_j))]} \quad (2)$$

Introducing a slope and an offset can capture how strongly annotators separate options, tend to break (or not break) ties, and tend to prefer choosing fewer answers overall. For the joint optimization of the  $\beta_{ik}$  and individual annotator parameters  $\gamma_a$  and  $\alpha_a$ , we use the negative log-likelihood with regularization for the annotator parameters as

$$\begin{aligned} -\log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = & \sum_k \sum_a \sum_{(i,j) \in \mathcal{D}_{ak}} \left[ y_{a,ij} (-\log P(i \succ j | a)) \right. \\ & + (1 - y_{a,ij}) (-\log [1 - P(i \succ j | a)]) \Big] \\ & + \lambda_0 \|\gamma_a\|^2 + \lambda_1 \|1 - \alpha_a\|^2. \end{aligned} \quad (3)$$

Here,  $y_{a,ij} = 1$  if annotator  $a$  says item  $i$  beats item  $j$ , and 0 otherwise. The set  $\mathcal{D}_{ak}$  is the collection of comparisons from annotator  $a$  of question  $k$ .

Besides regularization, we bound the individual annotator parameters ( $\gamma_a \in [-3.0, 3.0]$ ,  $\alpha_a \in [0.5, 2.0]$ ) during the optimization to balance the goal of slightly de-noising the resulting preference dataset while keeping the majority of differences between individual annotator preferences<sup>6</sup>. These bounds prevent the model from fixing contradictory data by pushing a parameter to an extreme and we show the fitted parameters in Figure 15 in Appendix C. To allow for a different set of assumptions about how to process the expert annotations for future use cases, our accompanying data pipeline code of MENTAT also allows the use of a regular Bradley-Terry model or modular replacements with alternative preference methods, e.g., Plackett-Luce.

Finally, we **calculate the overall probability**  $p$  of an answer  $i$  being preferred using the softmax function  $p = \sigma(\boldsymbol{\beta})_i$  to create the final preference labels for each question. To compare results with a regular and a hierarchical Bradley-Terry model, we check for how many questions the original question creator-preferred answer is in the top- $k$  ( $k \in [1, 5]$ ) answer options as defined by their resulting preference probability in Figure 3. While not an ideal metric, the original creator truth is always in the top-3 answer options defined by the hierarchical Bradley-Terry model, which is only the case for the regular model when looking at all answer options (top-5).

While the answers to the questions were designed to be ambiguous, most questions still have one or two objectively incorrect answers that violate clinical procedure or are factually inaccurate, e.g., incorrect billing codes for specific

<sup>6</sup>Our results do not significantly change without these bounds. We set them conservatively to reduce the risk of inducing bias.

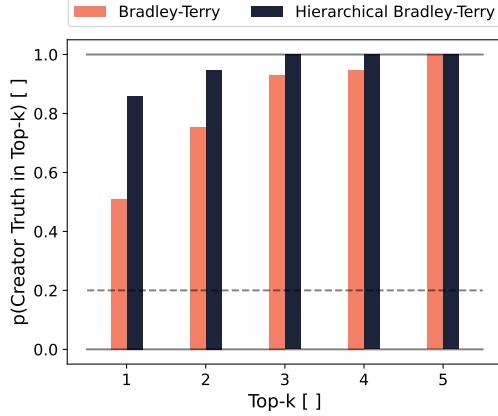


Figure 3. Comparing the probability for the original creator truth answer to be in the top- $k$  answers as defined by their preference probability when using a regular or a hierarchical Bradley-Terry model.

cases. Using one of the experts, we determine these answer options, manually set their probability to 0, and recalibrate the other answer probabilities. We do this at the end to get all individual annotator-specific behaviors across questions to determine the parameters with Equation (3). In most cases, these objectively wrong answers would have had a final preference probability less than the random baseline, i.e.,  $p \leq 0.2$ . Our accuracy-based evaluations in Section 4 are not affected by this post-processing step.

### 3.3. Use Cases and Applications

Although we intentionally designed MENTAT as an evaluation dataset grounded in human expertise rather than a large-scale training corpus, it offers several applications for research and development in mental healthcare AI.

For example, researchers can directly evaluate LM decision-making via multiple-choice questions across MENTAT’s five categories, as demonstrated in Section 4.2 and Section 4.3. MENTAT enables fine-grained comparisons of LM performance under varying task requirements and patient demographics, allowing practitioners to probe how models handle different presenting symptoms, acuity levels, or documentation requirements. Furthermore, as illustrated in Section 4.4, MENTAT can serve as a ground-truth reference for evaluating free-form LM outputs, providing important references for dynamic evaluations of increasingly agentic AI systems. Instead of requiring strictly multiple-choice answers, one can compare open-ended responses to the expert-annotated options, thus balancing structured and creative approaches to mental health decision-making. However, both applications share the caveat that MENTAT only partially captures the nuances of real-world interactions, such

as unstructured patient interviews or free-form model responses exceeding the scope of predefined expert-annotated choices.

Beyond standard accuracy metrics, MENTAT’s multiple-choice format and preference annotations permit novel evaluation strategies, such as computing cross-entropy or Brier Scores from LM log probabilities. These more nuanced techniques facilitate assessments of model confidence, enabling alignment methods that account for expert uncertainty and disagreement or aiding novel works on uncertainty quantification in LMs for crucial risk assessments (e.g. Lin et al., 2022; Kadavath et al., 2022; Kuhn et al., 2023; Shrivastava et al., 2024). For instance, our hierarchical annotation scheme (see Section 3.2) yields probabilities that can serve as “soft” labels for calibrating or training alignment models<sup>7</sup>. Finally, MENTAT’s emphasis on capturing expert disagreement encourages ongoing research into techniques for modeling inter-annotator bias, validating novel prompting methods that handle ambiguous psychiatric scenarios, and investigating how demographic anchoring (e.g., age, ethnicity, or gender) shifts model outputs.

## 4. Experiments

We demonstrate some of the different use cases of MENTAT outlined in Section 3.3: Evaluating decision-making accuracy across MENTAT’s five categories and how performance is impacted by patient demographic information, and using MENTAT as a ground-truth reference for evaluating free-form LM outputs.

### 4.1. Setup, Data, and Models

**Data:** To evaluate a selection of off-the-shelf and fine-tuned language models in *multiple-choice QA* settings in Section 4.2 and Section 4.3, we use the MENTAT evaluation dataset to create four separate evaluation datasets. We use the base set and sample each question once with a random patient gender, random age, and random ethnicity. We use this dataset  $\mathcal{D}_0$  of 183 prompts to evaluate models on all five tasks. To capture more variety for evaluating the impact of patient demographic information on accuracy, we create three additional datasets:  $\mathcal{D}_G$  with 549 prompts, by including each question once for each gender option,  $\mathcal{D}_A$  with 915 prompts, by including each question five times with a different random age, and  $\mathcal{D}_N$  with 1098 prompts, by including each question six times with a different random ethnicity. For the multiple-choice QA setting, we sample each tested LM at temperature  $T = 0$ .

For Section 4.4, we collect *free-form responses* by also us-

<sup>7</sup>Practical clinical deployments often rely on a much broader context than a single question/answer pair, so these metrics should be viewed as indicative rather than definitive.

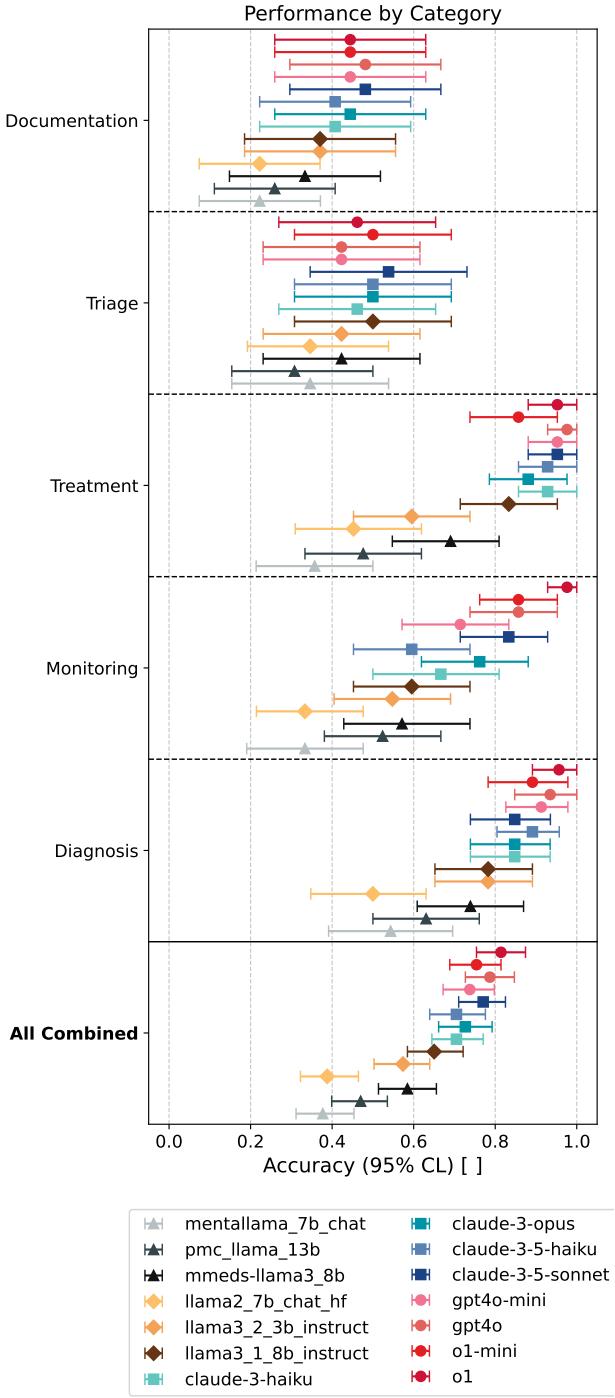


Figure 4. Using the core dataset of MENTAT ( $\mathcal{D}_0$ ), we evaluate eleven off-the-shelf instruction-tuned and three (mental) healthcare fine-tuned models for their task-specific accuracy. The random baseline is 0.2 due to all questions having five answer options.

ing the base set and removing the multiple-choice options to

**Table 1.** Average task-specific accuracy (95% CL) across all models and separately for only OpenAI and Anthropic models, uncertainties estimated from bootstrap resampling, and calculated with weighted arithmetic means.

| [MEAN ACCURACY](↑) | ALL MODELS       | ONLY OPENAI & ANTHROPIC |
|--------------------|------------------|-------------------------|
| DIAGNOSIS          | <b>0.82±0.03</b> | 0.91±0.04               |
| MONITORING         | 0.67±0.03        | 0.78±0.04               |
| TREATMENT          | 0.76±0.03        | <b>0.93±0.03</b>        |
| TRIAGE             | 0.44±0.02        | 0.48±0.03               |
| DOCUMENTATION      | 0.40±0.02        | 0.45±0.02               |

get a dataset  $\mathcal{D}_{FF}$  of 183 prompts. We only use questions in the categories of triage, diagnosis, and treatment, prompting the models to respond in one sentence and sample 10 responses from each tested LM for each question at sampling temperature  $T = 1$ . Prompting details for all datasets are stated in Appendix E.

**Models:** We evaluate eleven off-the-shelf instruction-tuned LMs and three LMs that have been fine-tuned for mental health applications. Specifically, we evaluate

- GPT-4o-mini (*gpt-4o-mini-2024-07-18*), GPT-4o (*gpt-4o-2024-08-06*), o1 (*o1-2024-12-17*), and o1-mini (*o1-mini-2024-09-12*) (OpenAI, 2025),
- Claude 3.5 Sonnet (*claude-3-5-sonnet-20241022*), Claude 3.5 Haiku (*claude-3-5-haiku-20241022*), Claude 3 Opus (*claude-3-opus-20240229*), Claude 3 Haiku (*claude-3-haiku-20240307*) (Anthropic, 2025),
- Llama2-7b (*llama2-7b-chat*) (Touvron et al., 2023), Llama3.1-8b (*llama3.1-8b-instruct*), Llama3.2-3b (*llama3.2-3b-instruct*) (Grattafiori et al., 2024),
- *PMC-LLaMA-13B* (Wu et al., 2023), *Meditron-7b* (Chen et al., 2023), *MentalLaMa-7b-chat* (Yang et al., 2024), and *MMedS-Llama-3-8B* (Wu et al., 2025).

Please note that none of the model developers recommend deploying their models in a clinical setting. Due to the lack of datasets, we could not find open-source models that were fine-tuned for mental healthcare decision-making, mainly Chatbots fine-tuned for therapy-like conversations with practitioners. Hence, MENTAT represents a critical step toward filling this gap, offering a rigorous, open dataset designed to evaluate and advance LM-based solutions for mental healthcare decision-making.

## 4.2. Task-Specific Accuracy

Using the dataset  $\mathcal{D}_0$ , we evaluate all models for their task-specific accuracy and showcase the results in Figure 4. Due

to restrictions of most closed-source models, we can only compare all models by relying on accuracy instead of using log probabilities to enable more nuanced analyses with, e.g., cross-entropy loss or Brier score. Unsurprisingly, the significantly larger closed-source models outperform smaller open-source models, and newer, more refined, and capable models tend to outperform their predecessors across categories. The mental health fine-tuned open source models do not outperform their Llama2 and Llama3 counterparts with statistical significance.<sup>8</sup> In particular, MMedS-Llama-3-8B, which was fine-tuned on a large corpus of web-scraped and LM-generated data set of clinical tasks and performs well on existing medical benchmarks like MedQA, *does not outperform* its Llama3.1-8b base model on MENTAT. This deviation highlights that expert-annotated datasets of real-world (non-LM-generated) clinical tasks are essential and missing.

Using the bootstrap resampled accuracy uncertainties, we can estimate symmetric Gaussian accuracy uncertainties at a 95% confidence level to enable cross-model average accuracy calculation per category with the maximum likelihood estimator for the weighted arithmetic mean. We do this calculation for all models together and again separately for the closed-source models from Anthropic and OpenAI. The results are shown in Table 1. We find that models perform best in the diagnosis and treatment category, followed by monitoring. Finally, all models perform around 50% accuracy for triage and documentation, with the open-source model confidence intervals even including the 20% random baseline, showcasing significant room for improvement and further research.

Due to the larger spread and lower accuracy of all models for the triage and documentation categories in Table 1, we conduct qualitative studies looking for failure patterns to check the validity of these categories. Triage questions focus on assessing the level of acuity of various psychiatric presentations and suggesting reasonable dispositions (e.g., inpatient, outpatient, discharge, etc.) and next steps. These can include cases of severe agitation, violence, situational safety, and more. Thus, conflicts with the helpfulness/harmlessness training objectives of the safety fine-tuning of language models often cause failures. This pattern mirrors observations in prior work studying how LMs respond to users in different mental health emergencies, finding that sycophancy and conflicts of safety-training objectives lead to failures (Grabb et al., 2024). Documentation questions (given long detailed clinical reports) mostly ask for appropriate CPT billing codes or a summary of relevant information. While we don't find a specific failure pattern, the main cause is that the evaluated LMs do not reliably recognize the rele-

<sup>8</sup>We omit Meditron-7b results due to performance issues (95% confidence interval includes random baseline in all categories).

Table 2. Average accuracy (95% CL) across tasks and all models and separately for only OpenAI and Anthropic models, uncertainties estimated from bootstrap resampling, and calculated with weighted arithmetic means.

| [MEAN ACCURACY](↑)    | ALL MODELS       | ONLY OPENAI & ANTHROPIC |
|-----------------------|------------------|-------------------------|
| USING $\mathcal{D}_G$ |                  |                         |
| FEMALE                | 0.63±0.02        | 0.72±0.02               |
| MALE                  | 0.65±0.01        | 0.74±0.02               |
| NON-BINARY            | <b>0.66±0.01</b> | <b>0.78±0.02</b>        |
| USING $\mathcal{D}_N$ |                  |                         |
| AFRICAN AMERICAN      | 0.65±0.02        | 0.77±0.02               |
| NATIVE AMERICAN       | 0.67±0.02        | <b>0.82±0.02</b>        |
| WHITE                 | <b>0.69±0.02</b> | 0.78±0.02               |
| BLACK                 | <b>0.69±0.02</b> | 0.80±0.02               |
| ASIAN                 | 0.65±0.02        | 0.77±0.02               |
| HISPANIC              | 0.68±0.01        | 0.81±0.02               |
| USING $\mathcal{D}_A$ |                  |                         |
| 18-33 YEARS           | <b>0.70±0.01</b> | <b>0.80±0.02</b>        |
| 33-49 YEARS           | 0.64±0.01        | 0.75±0.02               |
| 49-65 YEARS           | 0.64±0.01        | 0.74±0.02               |

vant information for consecutive therapy from the detailed reports. Another reason to consider is the smaller number of questions in the triage and documentation category (due to the immense annotation and expert verification efforts), which also increases the uncertainty bars compared to other categories.

### 4.3. Impact of Demographic Patient Information

We repeat the evaluation of all models but use the datasets  $\mathcal{D}_G$ ,  $\mathcal{D}_A$ , and  $\mathcal{D}_N$  to see how model performance is affected by different patient demographic information. In Appendix G, we illustrate the results for different patient gender coding in Figure 20, different patient ethnicity in Figure 22, and different patient age groups in Figure 21. We also calculate the average accuracy across all categories for different patient demographic information as in the previous section and present them in Table 2.

With statistical significance, we see that when we look at all models together or when only considering the tested closed-source models, that the accuracy is higher for non-binary-coded patients compared to female-coded patients; the accuracy is lower for patients with an "Asian" or "African American" background compared to other backgrounds, and the accuracy is higher for patients states in the age group 18 to 33 years compared to all other age groups. These results highlight the need for further methods to mitigate the propagation and perpetuation of harmful biases before deploying models in mental healthcare settings. Determining the exact cause of these results is complex, given the significant impact differences in pre- and post-training data

*Table 3.* Deviation (inconsistency) scores from the omitted multiple-choice answer options for different models across diagnosis, treatment, and triage tasks. We also list the accuracy results from Figure 4 for comparisons.

| MODEL        | DIAGNOSIS                                |                              | TREATMENT                                |                              | TRIAGE                                   |                              |
|--------------|--|------------------------------|--|------------------------------|--|------------------------------|
|              | ( $\downarrow$ ) FREE-FORM INCONSISTENCY | ( $\uparrow$ ) MCQA ACCURACY | ( $\downarrow$ ) FREE-FORM INCONSISTENCY | ( $\uparrow$ ) MCQA ACCURACY | ( $\downarrow$ ) FREE-FORM INCONSISTENCY | ( $\uparrow$ ) MCQA ACCURACY |
| GPT-4o<br>o1 | $0.55^{+0.05}_{-0.05}$                   | $0.93^{+0.07}_{-0.09}$       | $0.82^{+0.04}_{-0.04}$                   | $0.98^{+0.02}_{-0.05}$       | $0.75^{+0.04}_{-0.04}$                   | $0.42^{+0.19}_{-0.19}$       |
|              | $0.40^{+0.05}_{-0.05}$                   | $0.96^{+0.04}_{-0.07}$       | $0.77^{+0.04}_{-0.04}$                   | $0.95^{+0.03}_{-0.07}$       | $0.77^{+0.04}_{-0.05}$                   | $0.46^{+0.19}_{-0.19}$       |
| CLAUDE 3.5   |  |                              |  |                              |  |                              |
| HAIKU        | $0.75^{+0.04}_{-0.04}$                   | $0.89^{+0.07}_{-0.07}$       | $0.88^{+0.03}_{-0.03}$                   | $0.93^{+0.07}_{-0.10}$       | $0.79^{+0.04}_{-0.04}$                   | $0.50^{+0.19}_{-0.19}$       |
| SONNET       | $0.74^{+0.03}_{-0.03}$                   | $0.85^{+0.11}_{-0.11}$       | $0.84^{+0.04}_{-0.04}$                   | $0.95^{+0.05}_{-0.07}$       | $0.77^{+0.05}_{-0.05}$                   | $0.54^{+0.19}_{-0.19}$       |

have on models, as seen in other works studying tendencies and biases (e.g. Lamparth et al., 2025; Moore et al., 2024).

#### 4.4. Consistency of Free-Form Decisions

Here, we demonstrate that the MENTAT dataset can be used to evaluate LMs giving free-form responses to mental healthcare questions as well. Specifically, we test how consistent free-form LM responses are to the correct expert-annotated answer choice as defined by the highest preference probability for a question using  $\mathcal{D}_{FF}$ . To measure free-form consistency, we use the methodology and code from Shrivastava et al. (2024) (MIT license). Shrivastava et al. (2024) showed that it is possible to use 1–BERTScore (Zhang\* et al., 2020) with the DeBERTa xlarge embedding model (He et al., 2021) fine-tuned with MNLI (Williams et al., 2018) to measure free-form decision-making inconsistency in different settings, including replicating human expert classification labels of safe and unsafe responses of users in mental health emergencies interacting with LMs<sup>9</sup>.

By taking 1–BERTScore as an inconsistency metric, we can measure how far models deviate in free-form responses from the annotated expert answer options. Note, that this deviation could also increase for good answers not specified in the existing answer options. We can compute each response’s inconsistency with the expert-annotated correct annotation, average over all samples and questions, and estimate the uncertainty with bootstrap resampling between the average score of each question. The results in Table 3 show that a high multiple-choice accuracy score does not correlate with producing similar answers in free-form response prompting. While all models also have a high inconsistency score for the triage category where they have a lower accuracy, this is not true for the OpenAI models in the diagnosis category. All models generate responses that are very inconsistent with the original answer options in the treatment category. In summary, although a model can achieve high multiple-choice accuracy, its free-form answers may

deviate significantly from the expert “correct” options, highlighting the importance of evaluating decision-making in multiple-choice settings and with free-form responses rather than relying solely on exam-style questions about recalling fact-based knowledge.

## 5. Discussion and Limitations

The MENTAT dataset is a critical step in advancing AI evaluation for real-world psychiatric decision-making. Unlike traditional medical AI benchmarks emphasizing fact recall, MENTAT captures the inherent ambiguities and complexities of mental healthcare tasks. To the best of our knowledge, MENTAT is the first dataset of its kind, relying fully on expert-guided design and annotation for mental healthcare. This dataset provides a more realistic evaluation of AI capabilities by incorporating expert-created decision-making scenarios across diagnosis, treatment, monitoring, triage, and documentation. Our experiments reveal that while models perform well on structured tasks (diagnosis, treatment), they struggle significantly with ambiguous real-world tasks such as triage and documentation, underscoring the limitations of current AI models in handling uncertainty. Our evaluation results demonstrate that there are still significant differences between models and that biases remain a big issue. Bias analysis and mitigation are, therefore, a crucial part of a performance improvement debate. Also, the underwhelming performance of models trained on synthetic clinical decision-making data on MENTAT highlights that there are no easy “fixes” to these issues. While MENTAT does not offer a direct way to improve models through fine-tuning, it provides crucial information and insights for targeted improvements, for which there was no reliable dataset before.

**Limitations:** Despite its contributions, MENTAT has several limitations. First, the dataset is small and U.S.-centric, excluding fine-tuning applications and other healthcare systems. While we ensured diverse annotators and thorough annotation processing to reduce annotator bias as much as possible, biases or errors may persist (“doctor bias”). How-

<sup>9</sup>Shrivastava et al. (2024) also check the robustness of the inconsistency metric to systematic effects like text length.

ever, due to the inclusion of strong primers in the form of demographic information in psychiatric reports (i.e., the inputs to LMs), which makes analyzing prompt-induced bias with MENTAT crucial to not exaggerate existing biases.

Second, structured multiple-choice and free-form evaluations do not fully capture the dynamic nature of real-world psychiatric decision-making. In addition, MENTAT can only be used to measure equal-to-human performance (not above). However, our results demonstrate that there are still significant differences between models (e.g., Anthropic models perform significantly different in diagnosis, monitoring, and treatment categories) and that issues like biases make a superhuman performance debate premature.

Finally, there is a risk that AI systems could be prematurely deployed in psychiatric care, potentially leading to harmful, biased, or unreliable clinical decisions. We discuss these risks further in the following impact statement section.

**Future Directions:** Future efforts could expand MENTAT to include more questions and annotators. Also, AI models should be evaluated in conversational and interactive settings, reflecting real-world psychiatric interactions. Additionally, further research is needed to mitigate demographic biases and ensure AI models make equitable, safe, and clinically useful decisions.

## Acknowledgements

Max Lamparth is partially supported by the Stanford Center for AI Safety, the Center for International Security and Cooperation, and the Stanford Existential Risk Initiative. Declan Grabb is partially supported through Stanford’s Trailblazing Trainee Award through the Department of Psychiatry.

## Impact Statement

The MENTAT dataset represents a significant step forward in AI evaluation for psychiatry, providing a clinician-annotated, real-world benchmark that moves beyond traditional exam-style questions. By making the raw dataset (fully anonymized), processing code, evaluation framework, and final evaluation sets publicly available, we enable researchers to rigorously test models while allowing for easy modifications and extensions to fit various psychiatric AI applications. This ensures that MENTAT remains a flexible, transparent, and adaptable tool for AI alignment, fairness, and interpretability research.

A major ethical consideration in dataset creation is what to include and exclude—decisions that inevitably shape AI model development. We deliberately did not use LM-generated content, ensuring that all data comes from human clinical expertise rather than AI-reinforced biases.

While this approach enhances credibility, bias risks remain—particularly in expert judgments and demographic representation. Although we sought diverse annotators, biases inherent to psychiatric practice or subtle algorithmic tendencies may still persist. By systematically varying demographic attributes, we provide a lens to study how AI models respond to different patient profiles, reinforcing the need for bias mitigation before deployment.

A critical risk is that a good model performance on MENTAT could inadvertently encourage premature AI deployment in psychiatric care. As AI models improve, there may be economic pressures to automate diagnosis, triage, and billing, potentially leading to job displacement and diminished human oversight. Without rigorous safety measures, AI-driven psychiatric tools could reinforce systemic biases, misdiagnose patients, or fail to recognize mental health emergencies. Ethical AI in psychiatry must prioritize human-in-the-loop validation, regulatory oversight, and transparent reporting of model limitations.

By establishing a higher standard for AI evaluation in psychiatry, we hope to guide responsible AI development while preventing premature deployment that could compromise patient care. MENTAT is a foundation for safer, fairer, and clinically meaningful AI—one that must augment, not replace, human expertise in mental healthcare.

## References

- Adhikary, P. K., Srivastava, A., Kumar, S., Singh, S. M., Manuja, P., Gopinath, J. K., and Chakraborty, T. Exploring the efficacy of large language models in summarizing mental health counseling sessions: A benchmark study. *arXiv preprint*, 2024.
- Anthropic. Models. <https://docs.anthropic.com/en/docs/about-claude/models>, 2025. [Online; accessed 30-January-2025].
- Ates, H. C., Ates, C., and Dincer, C. Stress monitoring with wearable technology and AI. *Nature Electronics*, pp. 1–2, 2024.
- Axios. AI Medical Scribes: Comparison, Price, Accuracy - Abridge, Suki, Ambience, Nuance. <https://www.axios.com/pro/health-tech-deals/2024/03/21/ai-medical-scribes-comparison>, 2024. [Online; accessed 26-March-2024].
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N.,

- Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.
- Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code, 2021.
- Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., Sallinen, A., Sakhaeirad, A., Swamy, V., Krawczuk, I., Bayazit, D., Marmet, A., Montariol, S., Hartley, M.-A., Jaggi, M., and Bosselut, A. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.
- Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pp. 4299–4307, 2017.
- de Leeuw, J. R., Gilbert, R. A., and Luchterhandt, B. jspsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, 8(85):5351, 2023. doi: 10.21105/joss.05351. URL <https://doi.org/10.21105/joss.05351>.
- Falcetta, F. S., de Almeida, F. K., Lemos, J. C. a. S., Goldim, J. R., and da Costa, C. A. Automatic documentation of professional health interactions: A systematic review. *Artif. Intell. Med.*, 137(C), mar 2023.
- Gabriel, S., Puri, I., Xu, X., Malgaroli, M., and Ghassemi, M. Can AI relate: Testing large language model response for mental health support. *arXiv preprint arXiv:2405.12021*, 2024.
- Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., Das-Sarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., Hatfield-Dodds, Z., Henighan, T., Hernandez, D., Hume, T., Jacobson, J., Johnston, S., Kravec, S., Olsson, C., Ringer, S., Tran-Johnson, E., Amodei, D., Brown, T., Joseph, N., McCandlish, S., Olah, C., Kaplan, J., and Clark, J. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.
- Grabb, D., Lamparth, M., and Vasan, N. Risks from language models for automated mental healthcare: Ethics and structure for implementation. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=1pgfvZj0Rx>.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esibov, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Kovalcar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhota, K., Rantala-Yearly, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva,

- R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenber, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Habicht, J., Viswanathan, S., Carrington, B., Hauser, T. U., Harper, R., and Rollwage, M. Closing the accessibility gap to mental health treatment with a personalized self-referral Chatbot. *Nature Medicine*, pp. 1–8, 2024.
- He, P., Liu, X., Gao, J., and Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*, 2021.
- He, X., Cai, Z., Wei, W., Zhang, Y., Mou, L., Xing, E., and Xie, P. Pathological visual question answering. *arXiv preprint*, 2020.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding. In *ICLR*, 2021.
- Higgins, O., Short, B. L., Chalup, S. K., and Wilson, R. L. Artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health: An integrative review. *International Journal of Mental Health Nursing*, 32(4):966–978, 2023.
- Hou, W. and Ji, Z. Geneturing tests gpt models in genomics. *bioRxiv [Preprint]*, Mar 2023. doi: 10.1101/2023.03.11.532238.

- Hunter, D. R. Mm algorithms for generalized bradley-terry models. *The Annals of Statistics*, 32(1):384–406, 2004. doi: 10.1214/aoas/1079120141.
- Jia, M. Aime 2024 dataset. [https://huggingface.co/datasets/Maxwell-Jia/AIME\\_2024](https://huggingface.co/datasets/Maxwell-Jia/AIME_2024), 2024. Version 1.0.
- Jin, D., Pan, E., Oufatolle, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.
- Kasula, B. Y. Ethical Considerations in the Adoption of Artificial Intelligence for Mental Health Diagnosis. *International Journal of Creative Research In Computer Technology and Design*, 5(5):1–7, 2023.
- Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023.
- Lamparth, M., Corso, A., Ganz, J., Mastro, O. S., Schneider, J., and Trinkunas, H. Human vs. machine: Behavioral differences between expert humans and language models in wargame simulations. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’24, pp. 807–817. AAAI Press, 2025.
- Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.
- Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., and Mohr, D. C. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine*, 6(1):236, 2023.
- Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.
- Liu, X., Zhu, Y., Gu, J., Lan, Y., Yang, C., and Qiao, Y. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. *arXiv preprint arXiv:2311.17600*, 2023.
- McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Waters, P., and Halgamuge, M. N. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *arXiv preprint arXiv:2402.09880*, 2024.
- Moore, J., Deshpande, T., and Yang, D. Are large language models consistent over value-laden questions? In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15185–15221, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.891.
- Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., and Haber, N. Expressing stigma and inappropriate responses prevents llms from safely replacing mental health providers. *arXiv preprint arXiv:2504.18412*, 2025.
- Murphy, B. How the switch to pass-fail scoring for usmle step 1 is going. <https://www.ama-assn.org/medical-students/usmle-step-1-2/how-switch-pass-fail-scoring-usmle-step-1-going>, 2023. American Medical Association, Published April 5, 2023. Accessed May 16, 2025.
- Murphy, B. Mcat scores and medical school success: Do they correlate? <https://www.ama-assn.org/medical-students/preparing-medical-school/mcat-scores-and-medical-school-success-do-they-correlate>, 2024. American Medical Association, Published March 8, 2024. Accessed May 16, 2025.
- National Board of Medical Examiners. Step 1 sample items, 2021. URL [https://www.usmle.org/sites/default/files/2021-10/Step\\_1\\_Sample\\_Items.pdf](https://www.usmle.org/sites/default/files/2021-10/Step_1_Sample_Items.pdf). Accessed: 2024-01-28.
- Oh, J., Lee, G., Bae, S., Kwon, J. M., and Choi, E. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. In *Advances in Neural Information Processing Systems*, volume 36, 2024.
- OpenAI. Models. <https://platform.openai.com/docs/models/overview>, 2025. [Online; accessed 30-January-2025].
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.
- Pal, A., Umapathi, L. K., and Sankarasubbu, M. Medmcqa: A large-scale multi-subject multi-choice dataset

- for medical domain question answering. In Flores, G., Chen, G. H., Pollard, T., Ho, J. C., and Naumann, T. (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL <https://proceedings.mlr.press/v174/pal22a.html>.
- Pataranutaporn, P., Liu, R., Finn, E., and Maes, P. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5(10):1076–1086, 2023.
- Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Shi, S., Choi, M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Hausenloy, J., Zhang, O., Mazeika, M., Anderson, D., Nguyen, T., Mahmood, M., Feng, F., Feng, S. Y., Zhao, H., Yu, M., Gangal, V., Zou, C., Wang, Z., Wang, J. P., Kumar, P., Pokutnyi, O., Gerbicz, R., Popov, S., Levin, J.-C., Kazakov, M., Schmitt, J., Galgon, G., Sanchez, A., Lee, Y., Yeandon, W., Sauers, S., Roth, M., Agu, C., Riis, S., Giska, F., Utpala, S., Giboney, Z., Goshu, G. M., of Arc Xavier, J., Crowson, S.-J., Naiya, M. M., Burns, N., Finke, L., Cheng, Z., Park, H., Fournier-Facio, F., Wydallis, J., Nandor, M., Singh, A., Gehrunger, T., Cai, J., McCarty, B., Duclosel, D., Nam, J., Zampese, J., Herrer, R. G., Bacho, A., Loume, G. A., Galal, A., Cao, H., Garretson, A. C., Sileo, D., Ren, Q., Cojoc, D., Arkhipov, P., Qazi, U., Li, L., Motwani, S., de Witt, C. S., Taylor, E., Veith, J., Singer, E., Hartman, T. D., Rissone, P., Jin, J., Shi, J. W. L., Willcocks, C. G., Robinson, J., Mikov, A., Prabhu, A., Tang, L., Alapont, X., Uro, J. L., Zhou, K., de Oliveira Santos, E., Maksimov, A. P., Vendrow, E., Zenitani, K., Guillod, J., Li, Y., Vendrow, J., Kuchkin, V., Ze-An, N., Marion, P., Efremov, D., Lynch, J., Liang, K., Gritsevskiy, A., Martinez, D., Pageler, B., Crispino, N., Zvonkine, D., Fraga, N. W., Soori, S., Press, O., Tang, H., Salazar, J., Green, S. R., Brüssel, L., Twayana, M., Dieuleveut, A., Rogers, T. R., Zhang, W., Li, B., Yang, J., Rao, A., Loiseau, G., Kalinin, M., Lukas, M., Manolescu, C., Mishra, S., Kamdoum, A. G. K., Kreiman, T., Hogg, T., Jin, A., Bosio, C., Sun, G., Coppola, B. P., Tarver, T., Heidinger, H., Sayous, R., Ivanov, S., Cavanagh, J. M., Shen, J., Imperial, J. M., Schwaller, P., Senthilkuma, S., Bran, A. M., Dehghan, A., Algaba, A., Verbeken, B., Noveler, D., V, R. P., Schut, L., Sucholutsky, I., Zheltonozhskii, E., Lim, D., Stanley, R., Sivarajan, S., Yang, T., Maar, J., Wykowski, J., Oller, M., Sandlin, J., Sahu, A., Hu, Y., Fish, S., Heydari, N., Apronti, A., Rawal, K., Vilchis, T. G., Zu, Y., Lackner, M., Koppel, J., Nguyen, J., Antonenko, D. S., Chern, S., Zhao, B., Arsene, P., Goldfarb, A., Ivanov, S., Poświata, R., Wang, C., Li, D., Crisostomi, D., Achilleos, A., Myklebust, B., Sen, A., Perrella, D., Kaparov, N., Inlow, M. H., Zang, A., Thornley, E., Orel, D., Poritski, V., Ben-David, S., Berger, Z., Whitfill, P., Foster, M., Munro, D., Ho, L., Hava, D. B., Kuchkin, A., Lauff, R., Holmes, D., Sommerhage, F., Schneider, K., Kazibwe, Z., Stambaugh, N., Singh, M., Magoulas, I., Clarke, D., Kim, D. H., Dias, F. M., Elser, V., Agarwal, K. P., Vilchis, V. E. G., Klose, I., Demian, C., Anantheswaran, U., Zweiger, A., Albani, G., Li, J., Daans, N., Radionov, M., Rozhoň, V., Ma, Z., Stump, C., Berkani, M., Platnick, J., Nevirkovets, V., Basler, L., Piccardo, M., Jeanplong, F., Cohen, N., Tkadlec, J., Rosu, P., Padlewski, P., Barzowski, S., Montgomery, K., Menezes, A., Patel, A., Wang, Z., Tucker-Foltz, J., Stade, J., Goertzen, T., Kazemi, F., Milbauer, J., Ambay, J. A., Shukla, A., Labrador, Y. C. L., Givré, A., Wolff, H., Rossbach, V., Aziz, M. F., Kaddar, Y., Chen, Y., Zhang, R., Pan, J., Terpin, A., Muenninghoff, N., Schoelkopf, H., Zheng, E., Carmi, A., Jones, A., Shah, J., Brown, E. D. L., Zhu, K., Bartolo, M., Wheeler, R., Ho, A., Barkan, S., Wang, J., Stehberger, M., Kretov, E., Sridhar, K., EL-Wasif, Z., Zhang, A., Pyda, D., Tam, J., Cunningham, D. M., Goryachev, V., Patramanis, D., Krause, M., Redenti, A., Bugas, D., Aldous, D., Lai, J., Coleman, S., Bahaloo, M., Xu, J., Lee, S., Zhao, S., Tang, N., Cohen, M. K., Carroll, M., Paradise, O., Kirchner, J. H., Steinerberger, S., Ovchinnikov, M., Matos, J. O., Shenoy, A., de Oliveira Junior, B. A., Wang, M., Nie, Y., Giordano, P., Petersen, P., Sztyber-Betley, A., Shukla, P., Crozier, J., Pinto, A., Verma, S., Joshi, P., Yong, Z.-X., Tee, A., Andréoletti, J., Weller, O., Singhal, R., Zhang, G., Ivanov, A., Khouri, S., Mostaghimi, H., Thaman, K., Chen, Q., Khánh, T. Q., Loader, J., Cavalleri, S., Szlyk, H., Brown, Z., Roberts, J., Alley, W., Sun, K., Stendall, R., Lamparth, M., Reuel, A., Wang, T., Xu, H., Raparthi, S. G., Hernández-Cámarra, P., Martin, F., Malishev, D., Preu, T., Korbak, T., Abramovitch, M., Williamson, D., Chen, Z., Bálint, B., Bari, M. S., Kassani, P., Wang, Z., Ansarinejad, B., Goswami, L. P., Sun, Y., Elgnainy, H., Tordera, D., Balabanian, G., Anderson, E., Kvistad, L., Moyano, A. J., Maheshwari, R., Sakor, A., Eron, M., McAlister, I. C., Gimenez, J., Enyekwe, I., O., A. F. D., Shah, S., Zhou, X., Kamalov, F., Clark, R., Abdoli, S., Santens, T., Meer, K., Wang, H. K., Ramakrishnan, K., Chen, E., Tomasiello, A., Luca, G. B. D., Looi, S.-Z., Le, V.-K., Kolt, N., Mündler, N., Semler, A., Rodman, E., Drori, J., Fossum, C. J., Jagota, M., Pradeep, R., Fan, H., Shah, T., Eicher, J., Chen, M., Thaman, K., Merrill, W., Harris, C., Gross, J., Gusev, I., Sharma, A., Agnihotri, S., Zhelnov, P., Usawasutsakorn, S., Mofayzezi, M., Bogdanov, S., Piperski, A., Carauleanu, M., Zhang, D. K., Ler, D., Leventov, R., Soroko, I., Jansen, T., Lauer, P., Duersch, J., Taamazyan, V., Morak, W., Ma, W., Held, W., Huy, T. D., Xian, R., Zebaze, A. R., Mohamed, M., Leser, J. N., Yuan, M. X., Yacar, L., Lengler, J., Shahrtash, H., Oliveira, E., Jackson, J. W., Gonzalez, D. E., Zou, A.,

- Chidambaram, M., Manik, T., Haffenden, H., Stander, D., Dasouqi, A., Shen, A., Duc, E., Golshani, B., Stap, D., Uzhou, M., Zhidkovskaya, A. B., Lewark, L., Vincze, M., Wehr, D., Tang, C., Hossain, Z., Phillips, S., Muzhen, J., Ekström, F., Hammon, A., Patel, O., Remy, N., Farhidi, F., Medley, G., Mohammadzadeh, F., Peñaflor, M., Kasahun, H., Friedrich, A., Sparrow, C., Sakal, T., Dhamane, O., Mirabadi, A. K., Hallman, E., Battaglia, M., Maghsoudimehrabani, M., Hoang, H., Amit, A., Hulbert, D., Pereira, R., Weber, S., Mensah, S., Andre, N., Peristyy, A., Harjadi, C., Gupta, H., Malina, S., Albanie, S., Cai, W., Mehkary, M., Reidegeld, F., Dick, A.-K., Friday, C., Sidhu, J., Kim, W., Costa, M., Gurdogan, H., Weber, B., Kumar, H., Jiang, T., Agarwal, A., Ceconello, C., Vaz, W. S., Zhuang, C., Park, H., Tawfeek, A. R., Aggarwal, D., Kirchhof, M., Dai, L., Kim, E., Ferret, J., Wang, Y., Yan, M., Burdzy, K., Zhang, L., Franca, A., Pham, D. T., Loh, K. Y., Robinson, J., Gul, S., Chhablani, G., Du, Z., Cosma, A., White, C., Riblet, R., Saxena, P., Votava, J., Vinnikov, V., Delaney, E., Halasyamani, S., Shahid, S. M., Mourrat, J.-C., Vetoshkin, L., Bacho, R., Ginis, V., Maksapetyan, A., de la Rosa, F., Li, X., Malod, G., Lang, L., Laurendeau, J., Adesanya, F., Portier, J., Hollom, L., Souza, V., Zhou, Y. A., Yalin, Y., Obikoya, G. D., Arnaboldi, L., Rai, Bigi, F., Bacho, K., Clavier, P., Recchia, G., Popescu, M., Shulga, N., Tanwie, N. M., Lux, T. C. H., Rank, B., Ni, C., Yakimchyk, A., Huanxu, Liu, Häggström, O., Verkama, E., Narayan, H., Gundlach, H., Brito-Santana, L., Amaro, B., Vajipey, V., Grover, R., Fan, Y., e Silva, G. P. R., Xin, L., Kratish, Y., Łucki, J., Li, W.-D., Xu, J., Scaria, K. J., Vargus, F., Habibi, F., Long, Lian, Rodolà, E., Robins, J., Cheng, V., Grabb, D., Bosio, I., Fruhauff, T., Akov, I., Lo, E. J. Y., Qi, H., Jiang, X., Segev, B., Fan, J., Martinson, S., Wang, E. Y., Hausknecht, K., Brenner, M. P., Mao, M., Jiang, Y., Zhang, X., Avagian, D., Scipio, E. J., Siddiqi, M. R., Ragoler, A., Tan, J., Patil, D., Plecnik, R., Kirtland, A., Montecillo, R. G., Durand, S., Bodur, O. F., Adoul, Z., Zekry, M., Douville, G., Karakoc, A., Santos, T. C. B., Shamseldeen, S., Karim, L., Liakhovitskaia, A., Resman, N., Farina, N., Gonzalez, J. C., Maayan, G., Hoback, S., Pena, R. D. O., Sherman, G., Mariji, H., Pouriamanesh, R., Wu, W., Demir, G., Mendoza, S., Alarab, I., Cole, J., Ferreira, D., Johnson, B., Milliron, H., Safdari, M., Dai, L., Arthonthurasuk, S., Pronin, A., Fan, J., Ramirez-Trinidad, A., Cartwright, A., Pottmaier, D., Taheri, O., Outevsky, D., Stepanic, S., Perry, S., Askew, L., Rodríguez, R. A. H., Dendane, A., Ali, S., Lorena, R., Iyer, K., Salauddin, S. M., Islam, M., Gonzalez, J., Ducey, J., Campbell, R., Somrak, M., Mavroudis, V., Vergo, E., Qin, J., Borbás, B., Chu, E., Lindsey, J., Radhakrishnan, A., Jallon, A., McInnis, I. M. J., Hoover, A., Möller, S., Bian, S., Lai, J., Patwardhan, T., Yue, S., Wang, A., and Hendrycks, D. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.
- Raji, I. D., Daneshjou, R., and Alsentzer, E. It's time to bench the medical exam benchmark. *NEJM AI*, 2(2):A1e2401235, 2025.
- Raluca Balan, Anca Dobrean, C. R. P. Use of automated conversational agents in improving young population mental health: a scoping review. *npj Digital Medicine*, 7(1):75, 2024.
- Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=Ti67584b98>.
- Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., and Kochenderfer, M. Betterbench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL <https://openreview.net/forum?id=hc0q2buakM>.
- Saab, K., Tu, T., Weng, W. H., Tanno, R., Stutz, D., Wulczyn, E., and Natarajan, V. Capabilities of gemini models in medicine. *arXiv preprint*, 2024.
- Sadigh, D., Dragan, A. D., Sastry, S. S., and Seshia, S. A. Active preference-based learning of reward functions. In *Proceedings of Robotics: Science and Systems (RSS)*, 2017.
- Saguil, A., Dong, T., Gingerich, R. J., Swygert, K., LaRochelle, J. S., Artino, A. R. J., Cruess, D. F., and Durning, S. J. Does the mcat predict medical school and pgy-1 performance? *Military Medicine*, 180(4 Suppl):4–11, Apr 2015. doi: 10.7205/MILMED-D-14-00550.
- Salaudeen, O., Reuel, A., Ahmed, A., Bedi, S., Robertson, Z., Sundar, S., Domingue, B., Wang, A., and Koyejo, S. Measurement to meaning: A validity-centered framework for ai evaluation. *arXiv preprint arXiv:2505.10573*, 2025.
- Schmidt, C., Yowell, R., and Jaffe, E. *Procedure Coding Handbook for Psychiatrists*. American Psychiatric Pub., 2011. ISBN 9781585623747. URL <https://books.google.com/books?id=v-3iwAEACAAJ>.
- Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., and Althoff, T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57, 2023.
- Shrivastava, A., Hullman, J., and Lamparth, M. Measuring free-form decision-making inconsistency of language models in military crisis simulations. *arXiv preprint arXiv:2410.13204*, 2024.

- Sin, J. An AI chatbot for talking therapy referrals. *Nature Medicine*, pp. 1–2, 2024.
- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize from human feedback. *arXiv preprint arXiv:2009.01325*, 2020.
- Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., , and Wei, J. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.
- The New England Journal of Medicine. Image challenge, 2024. URL <https://www.nejm.org/image-challenge>. Accessed: 2024-01-28.
- Thieme, A., Hannatty, M., Lyons, M., Palacios, J., Marques, R. F., Morrison, C., and Doherty, G. Designing human-centered AI for mental health: Developing clinically relevant applications for online CBT treatment. *ACM Transactions on Computer-Human Interaction*, 30(2):1–50, 2023.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikell, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Tu, T., Palepu, A., Schaekermann, M., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Tomasev, N., Azizi, S., Singhal, K., Cheng, Y., Hou, L., Webson, A., Kulkarni, K., Mahdavi, S. S., Semturs, C., Gottweis, J., Barral, J., Chou, K., Corrado, G. S., Matias, Y., Karthikesalingam, A., and Natarajan, V. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*, 2024.
- Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S. T., Arora, S., Mazeika, M., Hendrycks, D., Lin, Z., Cheng, Y., Koyejo, S., Song, D., and Li, B. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.
- Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018.
- Wu, C., Lin, W., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. Pmc-llama: Towards building open-source language models for medicine. *arXiv preprint arXiv:2304.14454*, 2023.
- Wu, C., Qiu, P., Liu, J., et al. Towards evaluating and building versatile large language models for medicine. *npj Digital Medicine*, 8:58, 2025. doi: 10.1038/s41746-024-01390-4. URL <https://doi.org/10.1038/s41746-024-01390-4>.
- Yang, K., Zhang, T., Kuang, Z., Xie, Q., Huang, J., and Ananiadou, S. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024, WWW ’24*, pp. 4489–4500. Association for Computing Machinery, 2024.
- Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., and Chen, W. Mmmu: A massive multi-discipline multi-modal understanding and reasoning benchmark for expert agi. *arXiv preprint*, 2023.
- Zambrano Chaves, J., Bhaskhar, N., Attias, M., Delbrouck, J. B., Rubin, D., Loening, A., and Chaudhari, A. Rales: a benchmark for radiology language evaluations. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74429–74454, 2023.
- Zhang\*, T., Kishore\*, V., Wu\*, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=SkeHuCVFDr>.
- Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Liu, X., Lei, X., Tang, J., and Huang, M. Safetybench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*, 2023.

## A. How is MENTAT Different from Medical Exam Questions?

For years, medical AI benchmarks have focused on fact-based assessments. Most medical evaluations for LMs rely on board exams and medical student tests, primarily measuring knowledge recall rather than real-world clinical decision-making. These exams have little correlation with actual clinical practice, as passing them does not equate to the ability to manage patients effectively even in humans (Sagui et al., 2015).

A 32-year-old woman with type 1 diabetes mellitus has had progressive renal failure during the past 2 years. She has not yet started dialysis. Examination shows no abnormalities. Her hemoglobin concentration is 9 g/dL, hematocrit is 28%, and mean corpuscular volume is 94  $\mu\text{m}^3$ . A blood smear shows normochromic, normocytic cells. Which of the following is the most likely cause?

(A) Acute blood loss  
 (B) Chronic lymphocytic leukemia  
 (C) Erythrocyte enzyme deficiency  
 (D) Erythropoietin deficiency  
 (E) Immunohemolysis  
 (F) Microangiopathic hemolysis  
 (G) Polycythemia vera  
 (H) Sickle cell disease  
 (I) Sideroblastic anemia  
 (J)  $\beta$ -Thalassemia trait  
**(Answer: D)**

Figure 5. USMLE board exam question example

For example, Figure 5 presents a classic USMLE board exam question (National Board of Medical Examiners, 2021), which tests an AI model's ability to recall factual knowledge rather than apply practical decision-making skills. The question may assess the recognition of a laboratory abnormality in diabetes, but it does not evaluate whether the model can adjust insulin regimens, recognize psychosocial factors, or determine hospitalization needs—key components of real-world patient care. As highlighted in previous research, medical licensing exams do not strongly correlate with clinical competency, reinforcing the need for benchmarks that evaluate accurate decision-making skills rather than memorization.

| Question type | Attribute type      | Example template question   |
|---------------|---------------------|---|
| Single-Verify | SCP Code            | Does this ECG show symptoms of <b>non-specific ST changes</b> ?               |
|               | Noise               | Does this ECG show <b>baseline drift in lead I</b> ?                          |
|               | Stage of infarction | Does this ECG show <b>early stage of myocardial infarction</b> ?              |
|               | Extra systole       | Does this ECG show <b>ventricular extrasystoles</b> ?                         |
|               | Heart axis          | Does this ECG show <b>left axis deviation</b> ?                               |
|               | Numeric feature     | Does the <b>RR interval</b> of this ECG fall <b>within the normal range</b> ? |

Table 4. Example template questions for different ECG attributes.

Table 4 and Table 5 illustrate additional examples of widely used AI benchmarks, such as ECG-QA (Oh et al., 2024) and GeneTuring (Hou & Ji, 2023), which focus on highly structured, fact-based medical knowledge. These datasets and others like MedQA (Jin et al., 2021) have been leveraged by major AI companies, including Google's Gemini initiative (Saab et al., 2024), to highlight model performance. While these benchmarks evaluate text-based and multimodal AI capabilities, they focus heavily on fact memorization rather than applied clinical reasoning.

Unlike traditional medical AI benchmarks, MENTAT is designed by practicing psychiatrists to reflect real-world clinical scenarios. The dataset also includes ambiguous, multi-choice decision-making tasks rather than a single correct answer, simulating the complex nature of psychiatric practice. Furthermore, MENTAT aims to reduce bias by empowering a diverse group of clinicians in its development from the start, making it less likely to reinforce harmful racial, gender, or sexuality-based biases in mental healthcare.

| Category           | Task                                       | Prompt   | Result | AI Response   |
|--------------------|--|--|--------|---|
| Sequence alignment | DNA sequence alignment to human genome     | Align the DNA sequence to the human genome:<br>TGGGCTCA<br>AGTGATCATA.....   | chr7   | As a language model AI, I do not have the capability to align a DNA sequence to the human genome.....     |
|                    | DNA sequence alignment to multiple species | Which organism does the DNA sequence come from:<br>CGTACACC<br>ATTGGTGC..... | yeast  | The organism from which the DNA sequence comes cannot be determined based solely on the DNA sequence..... |

Table 5. DNA Sequence Alignment Tasks and AI Responses

In summary, MENTAT differs from medical exam questions by moving beyond fact recall to assess practical clinical decision-making in mental healthcare. While traditional benchmarks test AI models on medical knowledge, MENTAT evaluates whether AI can handle real-world psychiatric tasks, manage patient uncertainty, and make informed decisions in complex clinical environments.

## B. MENTAT Sample Questions and Category Summaries

**Note:** To avoid pre-training contamination, we omit sample answer options here. The full dataset is available (encrypted) on [GitHub](#).

### B.1. Diagnosis

#### Summary:

A diagnosis question often contains information about previous patient behavior or behavior you observed, along with other background information (like age, gender, ethnicity) before asking for a diagnosis.

A 15-year-old <NAT> male presents to his pediatrician for a well child check with his parents. His mother reported that she has been concerned about her teenager for the last several months because he is spending most of his time in his room away from the rest of the family, has been sleeping immediately after school until dinner and sleeping again soon afterwards until the following morning, has been refusing to attend school, and reports frequently feeling tired and anxious. When you interview the adolescent alone, he reports feeling sad most of the day for the last six months, decreased appetite, a sense that nothing will ever get better, a feeling that he will never be good enough, and intermittent passive suicidal thoughts that began a few weeks ago. What diagnosis is first on your differential?

Figure 6. Diagnosis Example 1

You are a psychiatrist who sees a patient (<AGE> <NAT> female) every week in your office. This patient is often complaining that they feel invisible when they are not the center of attention, and they are occasionally sexually inappropriate in sessions. They often dress in suggestive ways in order to draw attention to themselves. Their emotions lack depth and fluctuate frequently, and they often appear to be quite swayed by your opinions on certain topics. What is the most likely diagnosis?

Figure 7. Diagnosis Example 2

### B.2. Monitoring

#### Summary:

A monitoring question can ask for the appropriate metric (e.g., blood pressure, drug levels, etc.) to be monitored after a psychiatrist starts a patient (with a specific mental health disorder) on a specific medication.

An outpatient psychiatrist starts guanfacine for a <AGE> <NAT> non-binary individual with a history of ADHD and trauma and stressor related disorder. What metric should be monitored as the prescription is being started?

Figure 8. Monitoring Example 1

A <AGE> <NAT> non-binary individual with a history of bipolar II disorder presents to their psychiatrist for follow-up. They currently take lamotrigine and lurasidone. Their mood has been stable on both medications at their current dose for 1.5 years. They have a history of suicidal ideation and psychiatric hospitalization. They report to their psychiatrist that they are considering pregnancy. What is the next best step regarding their medication management?

Figure 9. Monitoring Example 2

### B.3. Treatment

**Summary:**

Treatment questions often summarize how a patient was diagnosed with a specific disorder and what medication (type, doses level, duration of medication) they were prescribed. Given additional background or current symptoms of the patient, the question asks how the treatment would need to be adjusted (different types of medication, doses level adjustments).

A ten-year-old <NAT> female presents to an outpatient child and adolescent psychiatrist with her mother. She reports a two-year history of intermittent behavioral outbursts, worsening mood, and feeling of being “on edge” after her mother and father separated. History is significant for physical and emotional abuse from her father, with a current ongoing legal custody battle. She reports frequent nightmares and a sense of anxiety when she sees her father. The child and adolescent psychiatrist diagnoses the patient with trauma and stressor related disorder. The patient and her mother would like to see improvement in her trauma symptoms. What is a reasonable medication to start?

*Figure 10.* Treatment Example 1

A <AGE> <NAT> woman is being treated for major depressive disorder. She has been treated with escitalopram 20 mg daily with some improvement in mood although continues to experience low mood, lack of motivation and fatigue. Her psychiatrist decides to add aripiprazole to her current regimen. She has never been treated with an antipsychotic medication before. What starting dose would be most appropriate in this case?

*Figure 11.* Treatment Example 2

### B.4. Triage

**Summary:**

Triage questions focus on assessing the level of acuity of various psychiatric presentations and suggesting reasonable dispositions (e.g., inpatient, outpatient, discharge, etc) and next-steps. These can include cases of severe agitation, violence, situational safety, and more.

A <AGE> <NAT> man is brought into the emergency department for psychiatric evaluation by police. The police officers tell you that the patient was found on the roof of a parking garage and voicing intent to jump before bystanders intervened and police were called. You bring the patient to a secluded interview room for an interview. The patient refuses to speak to you. During your interview, he makes intense eye contact, appears tense, abruptly stands with fists raised as if to strike at you. What is the most appropriate next step in management?

*Figure 12.* Triage Example 1

A <AGE> <NAT> man has a history of schizoaffective disorder, bipolar type and methamphetamine use disorder. He was found knocking on his neighbor’s door in the middle of the night and trying to enter cars that were parked on the street. Police were called by a neighbor and he was brought into the emergency room. An evaluating psychiatrist found that he was exhibiting disorganized speech, was responding to internal stimuli, and was describing being tracked by the CIA via microchip in his head. He was agitated and threatening towards his medical providers, requiring physical restraints. His urine drug screen was positive for amphetamines. How would most psychiatrists triage this situation?

*Figure 13.* Triage Example 2

You are a psychiatrist who recently evaluated a <AGE> <NAT> man for mood symptoms and agreed to see him for ongoing treatment. One weekend, you are at the gym and see your patient across the room. He smiles and waves and goes back to his workout. What is the most appropriate response?

*Figure 14.* Triage Example 3

### B.5. Documentation

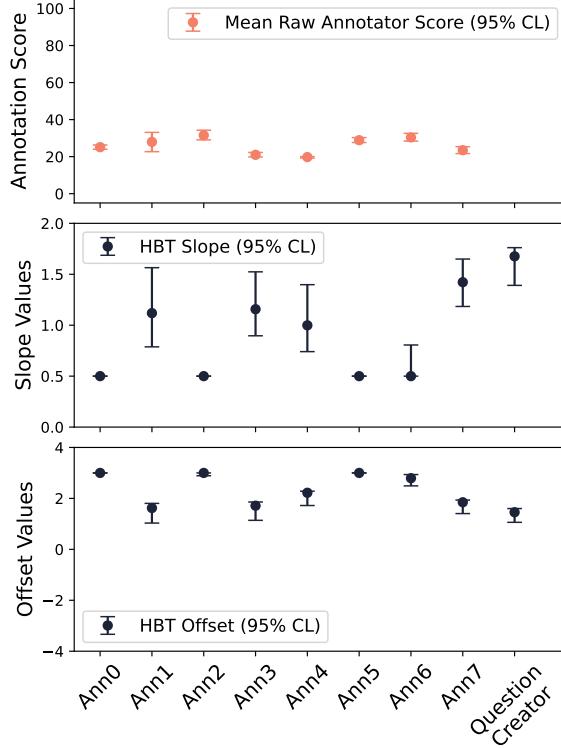
#### **Summary:**

Some questions ask, given long detailed clinical reports or intake surveys, for appropriate CPT billing codes or a summary of relevant information. Few ask for specific billing codes, but most present the results from the initial survey and ask for an accurate summary of relevant information.

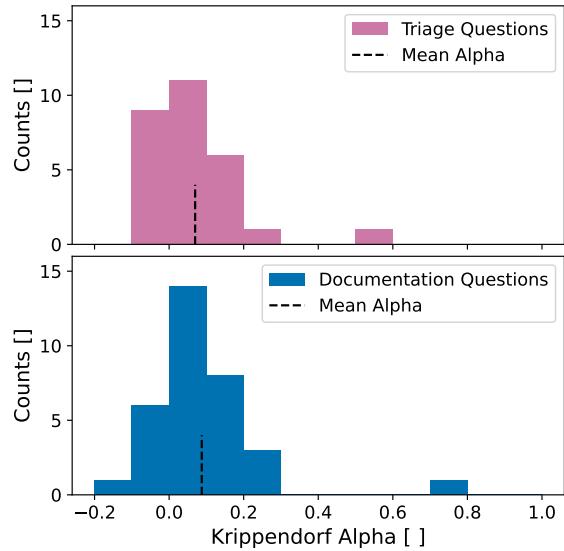
Examples are too long to include in this document, but typically involve:

- Selecting appropriate CPT Billing Codes.
- Summarizing lengthy intake reports accurately.

## C. Further Annotation Processing Results



**Figure 15.** (Top) We show the average raw annotation score with bootstrapped (95% CL) uncertainties for each annotator. All of them deviate from 50 with statistical significance (the random baseline). (Bottom) Fitted individual annotator parameters from the hierarchical Bradley-Terry model. Besides regularization in the log-likelihood objective, we bound the individual annotator parameters ( $\gamma_a \in [-3.0, 3.0]$ ,  $\alpha_a \in [0.5, 2.0]$ ) during the optimization to balance the goal of slightly de-noising the resulting preference dataset while keeping the majority of differences between individual annotator preferences. These bounds prevent the model from fixing contradictory data by pushing a parameter to an extreme. The fact that all annotators have a positive offset  $\gamma_a$  indicates that they all tend to choose one answer option to prefer over all others in a single annotation of one question.



**Figure 16.** We show the distribution of Krippendorff's  $\alpha$  for raw triage and documentation question annotations. We verify that the expert annotators do not converge on one answer option and that there is sufficient inter-annotator disagreement. Given our design choices, we expect  $\alpha$  to be naturally low as our goal is not to measure the presence of a single ground truth and low  $\alpha$  values ( $\alpha \leq 0.5$ ) will not tell us how useful a set of annotations is—only that experts statistically disagree.

## D. Why Do We Use A (Hierarchical) Bradley-Terry Model?

The Bradley-Terry (BT) model (Bradley & Terry, 1952) is a widely-used probabilistic model designed to predict outcomes of pairwise comparisons. The model is particularly powerful in scenarios where items or entities (such as human annotations, sports teams, or products) are compared against each other to establish a preference hierarchy or ranking. Mathematically, the BT model estimates the likelihood of one item being preferred over another based on latent "strength" parameters assigned to each item, see Equation (1). Hunter (Hunter, 2004) introduced generalized BT models, incorporating hierarchical or linear predictors on latent strength parameters. These extensions facilitate the modeling of group-level effects, item-specific covariates, and context-dependent preferences, enhancing model flexibility and applicability across diverse analytical scenarios.

Employing BT-based approaches is common practice (Hunter, 2004), particularly for calculating rankings (e.g., ELO rankings are a special case of online approximating a BT model) and aggregating human preferences from pairwise annotations. The latter is demonstrated in seminal works in reinforcement learning (Christiano et al., 2017; Sadigh et al., 2017; Leike et al., 2018) and is the underlying model for language model alignment methods (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022). These studies have extensively validated BT and HBT models' effectiveness in translating qualitative human feedback into robust quantitative metrics to guide reinforcement learning and ethical alignment.

In our case, we use the HBT model to reduce the risk of overfitting a simple BT model for each question by introducing annotator specific parameters. This approach allows us to use all available information of individual annotators across questions to conservatively smooth the annotation labels while preserving crucial annotator differences, e.g., tie breakers.

## E. Language Model Prompts

|   |  |
|---|--|
| <pre>f"Question: {q}\n\n" f"A: {answer_list[0]}\n" f"B: {answer_list[1]}\n" f"C: {answer_list[2]}\n" f"D: {answer_list[3]}\n" f"E: {answer_list[4]}\n\n" "Answer (single letter): "</pre> | <pre>f"Question: {q}\n\n" f"A: {answer_list[0]}\n" f"B: {answer_list[1]}\n" f"C: {answer_list[2]}\n" f"D: {answer_list[3]}\n" f"E: {answer_list[4]}\n\n" "Answer (only reply with a single letter!): "</pre> |
|---|--|

Figure 17. (Left) Prompt text MCQA variation A (as used for *gpt-4o-mini-2024-07-18*, *gpt-4o-2024-08-06*, *oI-2024-12-17*, and *oI-mini-2024-09-12*). (Right) Prompt text MCQA variation B (all other models). By looking at the responses from models evaluated with variation A, we verified that the recorded accuracy difference caused by using different prompts was  $\leq 1\%$ . The only exception was *oI-mini-2024-09-12*, for which we corrected the evaluation.

|   |
|---|
| <pre>f"Question: {q}\n\n" "Answer (write your reply in only one short sentence!): "</pre> |
|---|

Figure 18. Prompt text free-form (as used for the models evaluated in Section 4.4).

## F. Annotator Interface

### Question Q36:

A man is admitted to the medical floor for treatment of osteomyelitis. His medical team has consulted psychiatry as he has been asking bizarre questions and does not seem to understand his medical treatment. On evaluation, the patient states that he doesn't need to take antibiotics because he is Jesus and can heal himself. You call this patient's mother who states that the patient has recently been wandering the streets of their neighborhood at night and talking to himself. She also states that the patient's brother has been diagnosed with schizophrenia. The consult psychiatrist decides to start an antipsychotic medication and the medical team continues to treat his infection until he is transitioned to oral antibiotics and ready to discharge medically. At this point, the patient has agreed to take his oral antibiotics, but is still noted to be speaking to himself and thinks that he is Jesus. How would most consultation psychiatrists triage this situation?

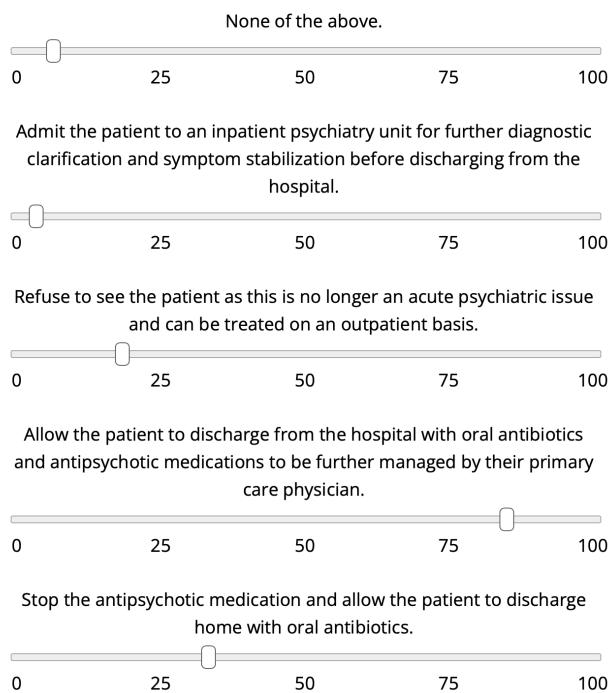


Figure 19. Example of the online annotation interface using the *jsPsych* library (de Leeuw et al., 2023) (MIT license). There is also a comment box below the sliders for feedback/comments, that is not shown.

## G. Further Evaluation Results

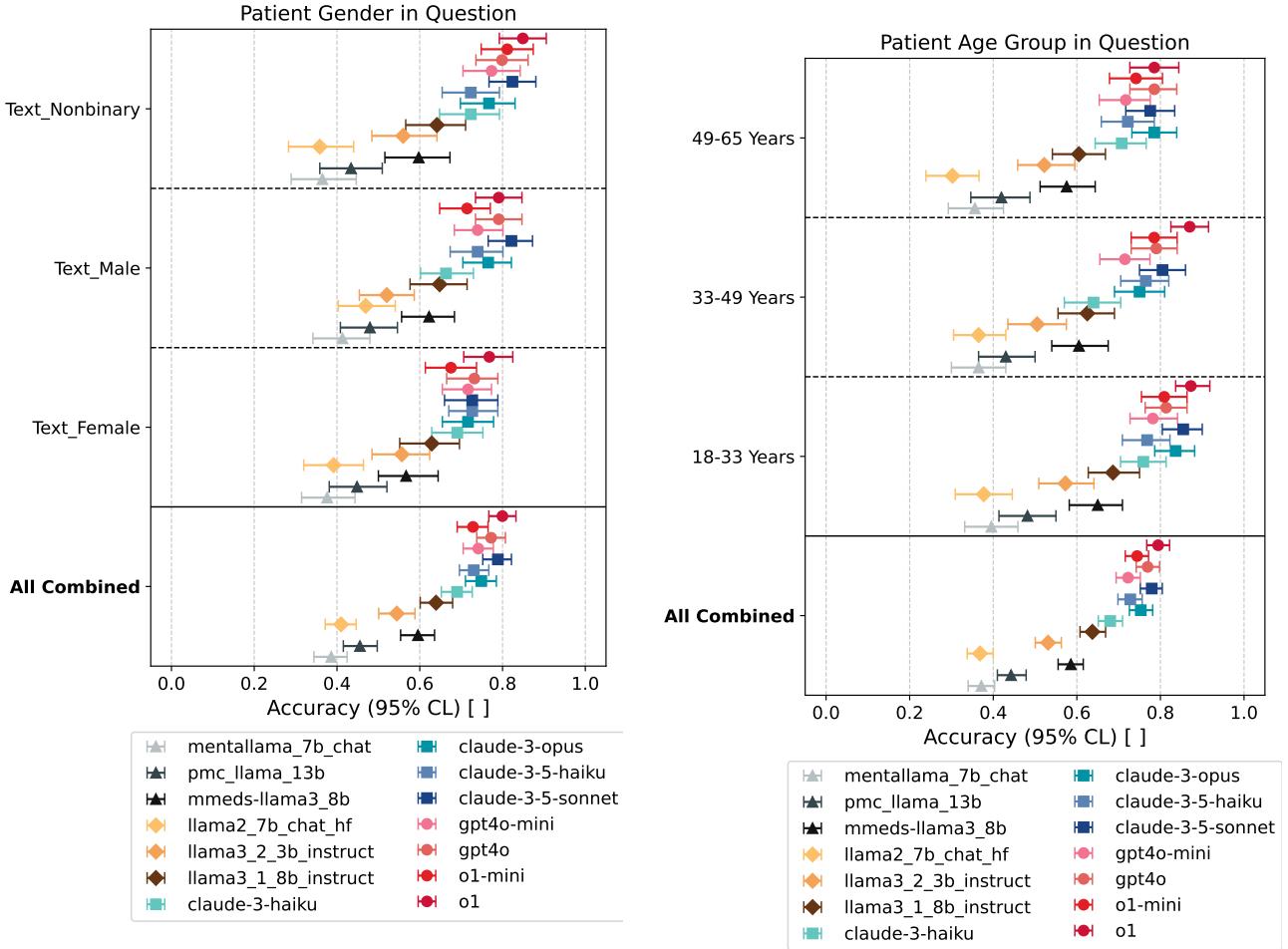


Figure 20. Using the  $\mathcal{D}_G$  dataset, we evaluate eleven off-the-shelf instruction-tuned and three (mental) healthcare fine-tuned models for overall accuracy and how it is impacted by different patient genders.

Figure 21. Using the  $\mathcal{D}_A$  dataset, we evaluate eleven off-the-shelf instruction-tuned and three (mental) healthcare fine-tuned models for overall accuracy and how it is impacted by different patient ages.

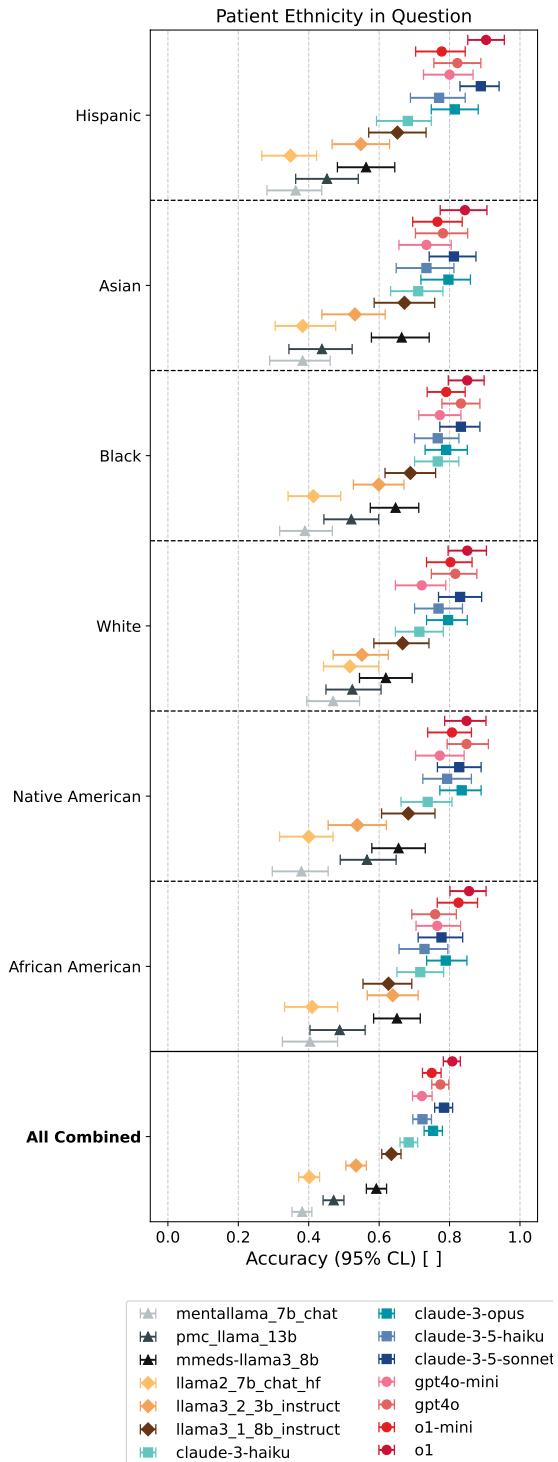


Figure 22. Using the  $\mathcal{D}_N$  dataset, we evaluate eleven off-the-shelf instruction-tuned and three (mental) healthcare fine-tuned models for overall accuracy and how it is impacted by different patient ethnicities.