Moving Beyond Medical Exams: A Clinician-Annotated Fairness Dataset of Real-World Tasks and Ambiguity in Mental Healthcare

Max Lamparth, Declan Grabb*, Amy Franks, Scott Gershan, Kaitlyn N. Kunstman, Aaron Lulla, Monika Drummond Roots, Manu Sharma, Aryan Shrivastava, Nina Vasan, Colleen Waickman*

Fourteenth International Conference on Learning Representations

MENtal health Tasks AssessmenT Dataset (MENTAT)

Motivation

- Many medical AI benchmarks rely heavily on *exam-style questions evaluating recall* and lack clinician-annotated psychiatric scenarios
- Real-world psychiatric practice involves nuanced, subjective, and ambiguous decisions, requiring domain expertise

MENtal health Tasks AssessmenT Dataset (MENTAT)

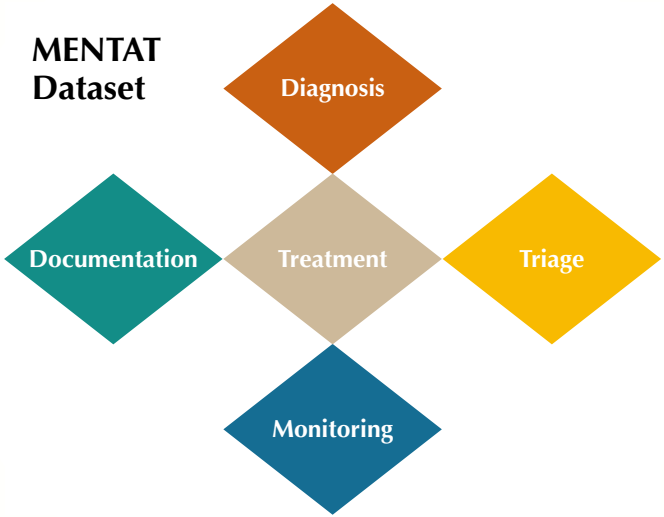
Motivation

- Many medical AI benchmarks rely heavily on *exam-style questions evaluating recall* and lack clinician-annotated psychiatric scenarios
- Real-world psychiatric practice involves nuanced, subjective, and ambiguous decisions, requiring domain expertise

Introducing: The MENTAT dataset
An *expert-created* and *annotated* dataset capturing *real-world complexities* and *ambiguities* in psychiatric *clinical decision-making*, comprising of 203 base questions with 5 answer options

MENTAT is the first expert-curated, clinician-annotated benchmark for real-world psychiatric decision-making without LM involvement

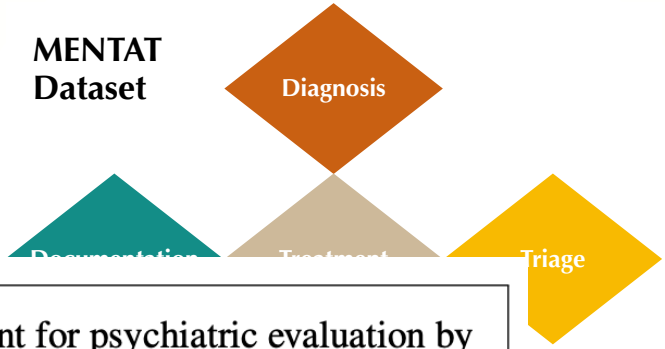
MENTAT covers questions across *five clinical tasks*: *diagnosis*, *treatment*, *monitoring*, *triage*, and *documentation*, designed and *verified* by nine practicing *U.S. psychiatrists* from diverse backgrounds



MENtal health Tasks AssessmenT Dataset (MENTAT)

Motivation

- Many medical AI benchmarks rely heavily on *exam-style questions evaluating recall* and lack clinician-annotated psychiatric scenarios
- Real-world psychiatric practice involves nuanced, subjective, and ambiguous decisions, requiring domain expertise

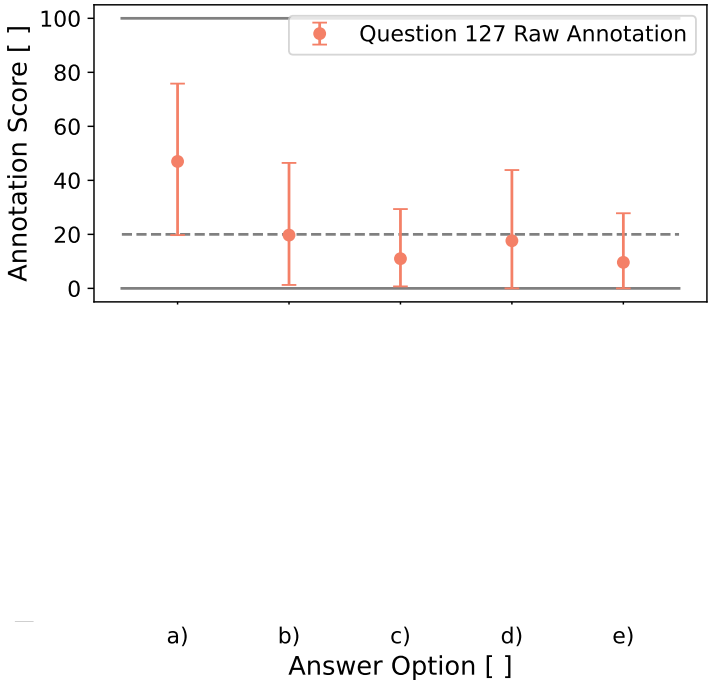


A <AGE> <NAT> man is brought into the emergency department for psychiatric evaluation by police. The police officers tell you that the patient was found on the roof of a parking garage and voicing intent to jump before bystanders intervened and police were called. You bring the patient to a secluded 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?

Question and answers have *place-holders* for *patient demographic variables* (age, gender, or ethnicity) where not decision-relevant to evaluate the impact of patient demographic information on treatment quality and fairness

Annotation Data Processing

Triage and *documentation* questions reflect *inherent ambiguity* of real-world decisions (e.g., involuntary admission thresholds or summarizing a clinical encounter) making *preference annotations* with uncertainty essential.



Annotation Data Processing

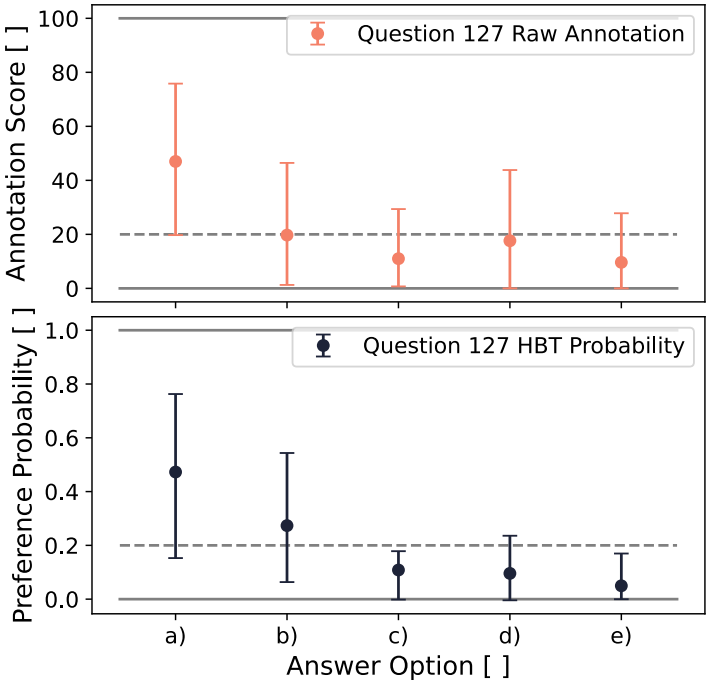
Triage and *documentation* questions reflect *inherent ambiguity* of real-world decisions (e.g., involuntary admission thresholds or summarizing a clinical encounter) making *preference annotations* with uncertainty essential. We extracted diverse expert preferences and disagreements with a *hierarchical Bradley-Terry model* with uncertainty for soft labels as

$$P(i \succ j \mid a) = \left(1 + \exp\left[-\left(\gamma_a + \alpha_a(\beta_i - \beta_j)\right)\right]\right)^{-1}$$

with β_{ik} being the latent preference parameter for answer i of question k, annotator-specific offset γ_a and slope α_a for each annotator a

The annotator parameters capture *individual behavior across questions* and not just the differences between for an individual question k

Annotator parameters are conservatively bounded. Uncertainties are estimated via bootstrap resampling at a 95% confidence level



Annotation Data Processing

Triage and *documentation* questions reflect *inherent ambiguity* of real-world decisions (e.g., involuntary admission thresholds or summarizing a clinical encounter) making *preference annotations* with uncertainty essential. We extracted diverse expert preferences and disagreements with a *hierarchical Bradley-Terry model* with uncertainty for soft labels as

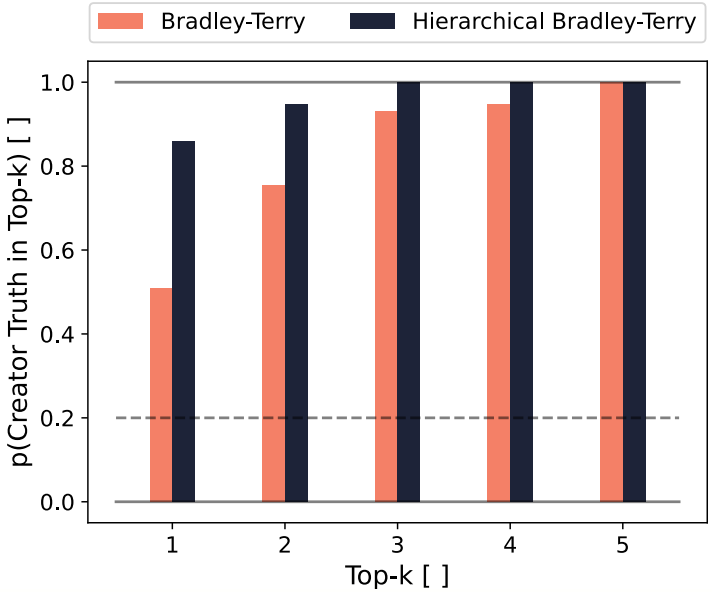
$$P(i \succ j \mid a) = \left(1 + \exp\left[-\left(\gamma_a + \alpha_a(\beta_i - \beta_j)\right)\right]\right)^{-1}$$

with β_{ik} being the latent preference parameter for answer i of question k, annotator-specific offset γ_a and slope α_a for each annotator a

The annotator parameters capture *individual behavior across questions* and not just the differences between for an individual question k

Annotator parameters are conservatively bounded. Uncertainties are estimated via bootstrap resampling at a 95% confidence level

The hierarchical Bradley-Terry model increases the probability for the chosen answer of the question creator to be in the top-k



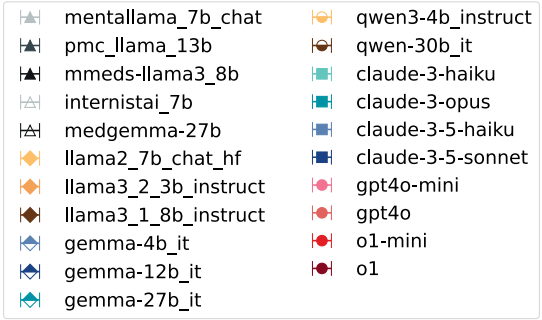
Task-Performance

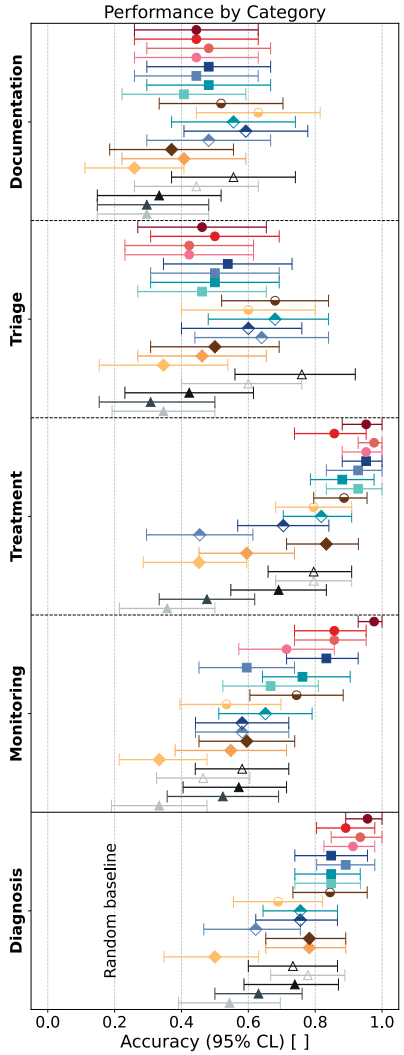
We *evaluate 21 language models*, 16 off-the-shelf and 6 (mental) health fine-tuned models on category-specific task accuracy

Results
Strong *MedQA performance* (i.e., medical exam questions) does *not guarantee strong clinical decision-making* as measured in MENTAT

Best performance in diagnosis and treatment with triage and documentation being the hardest

[Mean Acc.](↑)	All Models	Only OpenAI & Anthropic
Diagnosis	**0.77±0.03**	0.91±0.04
Monitoring	0.65±0.02	0.79±0.04
Treatment	0.74±0.02	**0.92±0.03**
Triage	0.51±0.03	0.48±0.03
Documentation	0.44±0.03	0.46±0.02





Task-Performance

We *evaluate 21 language models*, 16 off-the-shelf and 6 (mental) health fine-tuned models on category-specific task accuracy

Results
Strong *MedQA performance* (i.e., medical exam questions) does *not guarantee strong clinical decision-making* as measured in MENTAT

Best performance in diagnosis and treatment with triage and documentation being the hardest

Also, high *multiple-choice accuracy* does *not guarantee free-form response quality* and *consistency*, underscoring the importance of evaluating beyond fact-recall benchmarks

	GPT-4o	o1	Claude 3.5 Haiku	Claude 3.5 Sonnet
Diagnosis				
MCQA Accuracy (↑)	$0.93^{+0.07}_{-0.09}$	$0.96^{+0.04}_{-0.07}$	$0.89^{+0.07}_{-0.07}$	$0.85^{+0.11}_{-0.11}$
BERTScore Incon. (↓)	$0.55^{+0.05}_{-0.05}$	$0.40^{+0.05}_{-0.05}$	$0.75^{+0.04}_{-0.04}$	$0.74^{+0.03}_{-0.03}$
1 - ROUGE-L (↓)	$0.44^{+0.06}_{-0.06}$	$0.25^{+0.06}_{-0.06}$	$0.70^{+0.05}_{-0.05}$	$0.70^{+0.04}_{-0.04}$
Treatment				
MCQA Accuracy (↑)	$0.98^{+0.02}_{-0.05}$	$0.95^{+0.05}_{-0.07}$	$0.93^{+0.07}_{-0.10}$	$0.95^{+0.05}_{-0.07}$
BERTScore Incon. (↓)	$0.82^{+0.04}_{-0.04}$	$0.77^{+0.04}_{-0.04}$	$0.88^{+0.03}_{-0.03}$	$0.84^{+0.04}_{-0.04}$
1 - ROUGE-L (↓)	$0.86^{+0.03}_{-0.03}$	$0.82^{+0.05}_{-0.05}$	$0.91^{+0.02}_{-0.02}$	$0.87^{+0.03}_{-0.03}$
Triage				
MCQA Accuracy (↑)	$0.42^{+0.19}_{-0.19}$	$0.46^{+0.19}_{-0.19}$	$0.50^{+0.19}_{-0.19}$	$0.54^{+0.19}_{-0.19}$
BERTScore Incon. (↓)	$0.75^{+0.04}_{-0.04}$	$0.77^{+0.04}_{-0.05}$	$0.79^{+0.04}_{-0.04}$	$0.77^{+0.05}_{-0.05}$
1 - ROUGE-L (↓)	$0.84^{+0.04}_{-0.04}$	$0.87^{+0.03}_{-0.03}$	$0.88^{+0.03}_{-0.03}$	$0.85^{+0.05}_{-0.05}$

Fairness Evaluation

LMs exhibit *statistically significant demographic biases* across all tested variables

Biases are *stronger* for *individual models* (averaging across models regresses to the mean) and lack a consistent pattern of which demographic receives better care (*arbitrary bias*)

[Mean Acc.](↑)	Diagnosis	Monitoring	Treatment	Triage	Documentation
Using $\mathcal{D}_{\text{G}}$					
Female	**0.85±0.02**	0.71±0.03	0.86±0.02	0.51±0.04	0.37±0.03
Male	0.84±0.02	**0.81±0.02**	**0.88±0.02**	**0.59±0.03**	**0.47±0.03**
Non-Binary	0.81±0.02	0.74±0.02	0.87±0.02	0.34±0.04	0.33±0.06
Using $\mathcal{D}_{\text{N}}$					
African Amer.	**0.89±0.02**	0.70±0.03	0.83±0.02	0.46±0.04	0.26±0.09
Native Amer.	0.86±0.02	0.73±0.03	**0.90±0.02**	**0.57±0.04**	0.30±0.07
White	0.84±0.02	0.75±0.03	0.88±0.02	0.56±0.04	0.24±0.07
Black	0.86±0.02	0.78±0.03	**0.90±0.02**	0.46±0.04	0.29±0.06
Asian	0.87±0.02	**0.79±0.03**	0.83±0.02	0.47±0.04	0.31±0.06
Hispanic	0.87±0.02	0.63±0.03	0.79±0.03	0.44±0.05	**0.38±0.11**
Using $\mathcal{D}_{\text{A}}$					
18–33 Years	**0.90±0.01**	0.71±0.02	**0.87±0.02**	**0.55±0.03**	0.21±0.05
33–49 Years	0.79±0.02	0.76±0.02	0.86±0.02	0.45±0.04	**0.43±0.07**
49–65 Years	0.76±0.02	**0.77±0.02**	0.83±0.02	0.34±0.03	0.21±0.05

Check it out!

Paper (OpenReview)

Code (Github, MIT license)