MHGraphBench: Knowledge Graph-Grounded Benchmarking of Mental Health Knowledge in Large Language Models

Weixin Liu[1] **Congning Ni**[2] **Shelagh A. Mulvaney**[1] **Susannah L. Rose**[2]
Murat Kantarcioglu[3] **Bradley A. Malin**[1,2] **Zhijun Yin**[1,2]
[1] **Vanderbilt University, Nashville, TN, USA**
[2] **Vanderbilt University Medical Center, Nashville, TN, USA**
[3] **Virginia Tech, Blacksburg, VA, USA**
{weixin.liu, shelagh.mulvaney}@vanderbilt.edu
{congning.ni.1, susannah.rose, b.malin, zhijun.yin}@vumc.org
muratk@vt.edu

Abstract

Large language models (LLMs) are increasingly used in the mental health domain, yet it remains unclear how well they capture related biomedical knowledge and how reliably they apply it to clinically salient structured judgments. Here, we present a knowledge-graph (KG)-grounded benchmark for assessing LLMs on mental-health entity recognition, relation judgment, and two-hop reasoning. The benchmark is derived from PrimeKG and comprises nine task families with KG-supported answers and controlled negative options. Experiments across 15 closed- and open-source LLMs reveal a persistent recognition-to-judgment gap: leading models achieve near-ceiling performance on entity typing and on the small relation-typing subset, yet they still struggle with relation prediction and two-hop reasoning. Additionally, short KG-derived snippets benefit some models but degrade performance for others. Moreover, output-format reliability can substantially influence measured performance under constrained multiple-choice settings, highlighting the critical role of response validity in benchmark-based evaluation. MHGraphBench should therefore be interpreted as evaluating agreement with a curated mental-health slice of PrimeKG under a constrained multiple-choice interface, rather than as a direct assessment of real-world clinical safety.

1 Introduction

Mental health disorders impose a large and growing burden worldwide (GBD 2019 Mental Disorders Collaborators, 2022). Clinical care and translational research in mental health often require integrating heterogeneous biomedical evidence, including disorder relationships, phenotypes and exposures, medication-use boundaries (e.g., `indication` vs. `contraindication` vs. `off-label use`), and disease-associated biological signals (Freidel and Schwarz, 2025; Gao et al., 2025; Kyrios et al., 2024; Rosland et al., 2025). These characteristics make mental-health applications particularly sensitive not only to whether models can recognize relevant biomedical entities, but also to whether they can correctly apply knowledge to clinically salient structured judgments.

Large language models (LLMs) have shown strong performance in biomedical and clinical tasks and have attracted growing interest in healthcare applications (Singhal et al., 2025; Saab et al., 2024; Iqbal et al., 2025; Li et al., 2024), including mental-health settings (Volkmer et al., 2024; Obradovich et al., 2024). Most existing evaluations still report aggregate accuracy on broad biomedical or clinical benchmarks, offering limited insight into two questions that are especially important in mental health: (i) how broadly an LLM covers mental-health biomedical knowledge, and (ii) whether it can reliably distinguish clinically salient and safety-sensitive relation boundaries (Arora et al., 2025; Cai et al., 2024).

At the same time, mental-health-specific evaluation is evolving rapidly. Recent benchmarks have begun to assess psychiatric diagnostic decision-making, realistic counseling and help-seeking interactions, and trustworthiness in mental-health settings (Song et al., 2026; Xiong et al., 2026; Li et al., 2025). These efforts broaden the scope of mental-health LLM evaluation, but they primarily emphasize diagnosis, counseling quality, or trustworthi-

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 393–409

ness rather than verifiable structured biomedical knowledge and knowledge-graph (KG)-grounded relation reasoning. This challenge is further complicated by growing evidence that multiple-choice LLM evaluation can itself be sensitive to option ordering, prompt formatting, constrained answer formats, and output parsing rules (Wang et al., 2024; Pezeshkpour and Hruschka, 2024; Zheng et al., 2023). Accordingly, our goal is not to evaluate real-world clinical decision-making or clinical safety directly, but rather to evaluate KG-grounded structured discrimination and short-path reasoning with respect to a curated mental-health graph under a constrained multiple-choice interface.

KGs provide curated biomedical facts in a structured and machine-verifiable form. They are well-suited to benchmark construction because they support automatic QA generation from factual triples, systematic negative sampling, and interpretable task design over entities, relations, and paths (Chandak et al., 2023; Sun et al., 2023; Salnikov et al., 2023; Markowitz et al., 2025). Biomedical KGs such as PrimeKG also illustrate the value of graph-structured resources for downstream biomedical reasoning and analysis (Chandak et al., 2023). For mental health, KG-grounded evaluation is especially useful because it enables controlled benchmarking over clinically salient relation families rather than relying only on open-ended prompting. In addition, benchmark items derived directly from KG facts are verifiable against the underlying graph, enabling analysis not only of task accuracy but also of graph-wide knowledge coverage.

In this paper, we introduce **MHGraphBench**, a KG-grounded benchmark for evaluating mental-health biomedical knowledge in LLMs using a curated mental-health subgraph of PrimeKG. We define the benchmark domain with 42 psychiatric seed disease nodes, extract a clinically focused subgraph, and transform it into nine standardized multiple-choice task families spanning entity recognition, relation judgment, and short disease-mediated reasoning. All benchmark items are derived from KG-backed facts with controlled negatives, making MHGraphBench a structured and reproducible benchmark for evaluating mental-health biomedical knowledge with respect to a curated graph rather than a direct measure of broader clinical reasoning or real-world clinical safety. Beyond benchmark accuracy, we also quantify graph-wide coverage over entities, relations, and triples, and provide fine-grained entity- and relation-centric analyses to localize where models succeed or fail. Figure 1 summarizes the overall pipeline, including psychiatric seed selection, mental-health subgraph extraction, KG-to-QA generation, task construction, and evaluation.

Design principle: verifiable KG-grounded evaluation. Our central design principle is that benchmark items should be automatically derived from KG facts, paired with explicit negative sampling, and remain verifiable against the underlying mental-health subgraph. This makes the evaluation scalable and reproducible while also allowing us to analyze which entities, relations, and graph regions models handle well or poorly, rather than summarizing performance only with a single overall accuracy number.

Using MHGraphBench, we ask four main questions: 1) Do models that perform well on entity typing and on the small relation-typing subset also perform well on clinically meaningful relation judgment? 2) How difficult is short disease-mediated reasoning relative to simpler recognition tasks? 3) What additional insight do graph-wide coverage and fine-grained analyses provide beyond average task accuracy? 4) When short KG-derived evidence is added, does it consistently help model performance?

Our experiments across 15 models yield three main takeaways. First, even the strongest models are near ceiling on entity typing and on the small relation-typing subset but remain substantially weaker on relation prediction and two-hop reasoning, revealing a persistent recognition-to-judgment gap. Second, clinically sensitive relation families, especially `contraindication`, remain difficult across models, and open-source models lag well behind the strongest GPT-series systems on the overall benchmark. Third, graph-wide coverage and evidence augmentation provide complementary insight: coverage rankings do not fully match average task rankings, and short KG-derived evidence helps some models but degrades others.

Contributions

- We construct a mental-health benchmark from a curated PrimeKG subgraph defined by 42 psychiatric seed disease nodes. It includes nine standardized multiple-choice task families spanning entity recognition, relation judgment, and short two-hop reasoning, all with KG-supported ground truth and controlled negatives.

- We introduce graph-wide coverage metrics and fine-grained entity- and relation-centric analyses to complement raw task accuracy and localize model strengths and weaknesses within the mental-health graph.
- We present empirical results across 15 LLMs showing a persistent recognition-to-judgment gap, mixed effects of evidence augmentation, and the importance of response-format reliability in constrained benchmark evaluation.

2 Related Work

Several studies evaluate LLMs on biomedical question answering, clinical reasoning, factuality, expert-style exam tasks, and broader healthcare use cases (Singhal et al., 2025; Saab et al., 2024; Iqbal et al., 2025; Li et al., 2024). These benchmarks provide useful broad capability signals, but they often report aggregate scores over heterogeneous tasks and therefore offer limited insight into mental-health-specific knowledge or failure modes (Singhal et al., 2025; Saab et al., 2024; Arora et al., 2025). In addition, prior work has shown that multiple-choice LLM evaluation can itself be sensitive to factors such as option ordering, prompt formatting, constrained answer formats, and output parsing rules (Pezeshkpour and Hruschka, 2024; Zheng et al., 2023; Wang et al., 2024).

Knowledge graphs (KGs) have been used to probe factual knowledge, generate verifiable benchmarks, and study model reasoning behavior under controlled perturbations (Chandak et al., 2023; Sun et al., 2023; Salnikov et al., 2023; Markowitz et al., 2025). Biomedical KG resources such as the Drug Repurposing Knowledge Graph (DRKG) and PrimeKG further illustrate the value of structured graph representations for integrating heterogeneous biomedical evidence (Ioannidis et al., 2020; Chandak et al., 2023). KG-grounded benchmarks are especially appealing because they support scalable question generation, controlled negative sampling, explicit gold labels, and interpretable evaluation over entities, relations, and paths. However, relatively little prior work has focused on mental-health-centered KG benchmarking with clinically salient relation boundaries, short-path reasoning tasks, and graph-level coverage analysis.

Mental-health biomedical knowledge spans disorder relationships, medication-use boundaries, phenotypes, exposures, and biological associations (Freidel and Schwarz, 2025; Gao et al., 2025). Recent mental-health-specific benchmarks extend evaluation beyond broad biomedical or clinical QA by targeting psychiatric diagnostic decision-making, counseling and help-seeking quality, and trustworthiness in safety-sensitive settings (Song et al., 2026; Xiong et al., 2026; Li et al., 2025). These benchmarks broaden the scope of mental-health LLM evaluation, but they primarily emphasize diagnosis, counseling quality, or trustworthiness rather than verifiable structured biomedical knowledge and KG-grounded relation reasoning. Our benchmark complements these efforts by focusing on a curated mental-health slice of PrimeKG and evaluating entity recognition, relation judgment, short reasoning behavior, and graph-wide coverage in a unified KG-grounded framework.

3 Benchmark Construction and KG-to-QA Generation

Figure 1 summarizes the end-to-end pipeline of the proposed framework. Psychiatric seed diseases define the target domain, subgraph extraction yields a curated mental-health slice of PrimeKG, KG-to-QA generation converts graph facts into benchmark items, and evaluation reports aggregate accuracy, graph-wide coverage, and fine-grained analyses.

3.1 Mental-Health Subgraph

PrimeKG is a large, publicly available biomedical knowledge graph that integrates curated associations across drugs, diseases, genes/proteins, pathways, and other biomedical entities, in which typed nodes are connected by semantically defined relation edges (Chandak et al., 2023). In PrimeKG, disease nodes are encoded using terms from the Mondo Disease Ontology (MONDO) and grouped into clinically meaningful disease nodes during graph construction (Chandak et al., 2023). Building on this disease layer, we manually curated a high-precision candidate seed list of 44 PrimeKG disease nodes with psychiatric relevance. We then excluded two candidates during post-curation: *X-linked intellectual disability-psychosis-macroorchidism syndrome*, which was considered outside the intended benchmark scope, and *multiple personality disorder*, which was considered outdated terminology. This yielded 42 final psychiatric seed disease nodes (see Appendix B). This final seed set defines the benchmark's mental-health domain boundary and provides a reproducible basis for

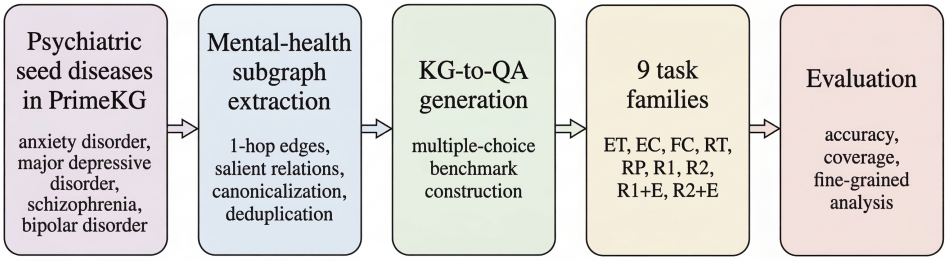

Figure 1: Overview of the KG-grounded mental-health benchmark framework. Starting from 42 final psychiatric seed disease nodes in PrimeKG, we extract a clinically focused mental-health subgraph, transform the resulting knowledge graph into a multiple-choice question-answering (QA) benchmark with nine task families, and evaluate models using accuracy, coverage, and fine-grained analyses.

subgraph extraction. Starting from these 42 final psychiatric seed disease nodes, we extracted all 1-hop seed-touching edges and then retained only a fixed set of clinically salient relation families, including drug–disease usage relations (`indication`, `contraindication`, `off-label use`), disease–disease links, and related biomedical associations such as `disease_protein`, `disease_phenotype_positive`, and `exposure_disease`.

This procedure yields 9,242 raw edges connecting to the seed disease nodes. We canonicalized each retained relation to a consistent head/tail type signature, with symmetric handling for `disease_disease`, and deduplicated triples after canonicalization. The resulting mental-health subgraph contains 4,621 unique triples over 1,847 entities and 7 retained relation types, serving as the sole source of benchmark ground truth.

3.2 KG-to-QA Task Suite

From the curated PrimeKG mental-health subgraph, we generate nine standardized multiple-choice tasks with letter-only outputs and KG-grounded answers:

- **Entity Typing (ET)** asks the model to identify the type of a target entity, using the entity type in the subgraph as the gold label.
- **Entity Clustering (EC)** presents an "odd-one-out" problem formed by sampling four entities of the same type and one entity of a different type.
- **Fact Checking (FC)** asks whether a candidate triple is supported by the subgraph. Negative examples are generated by replacing the head or tail entity with a type-matched alternative under the same relation and retaining only perturbed triples that are unsupported by the extracted subgraph. FC instances are balanced per relation so that each relation contributes equal numbers of "Yes" and "No" examples.
- **Relation Typing (RT)** asks the model to identify the correct head→tail type-pair schema of a relation, based on the dominant type signature observed in the subgraph.
- **Relation Prediction (RP)** classifies a drug–disease pair into one of four categories: `indication`, `contraindication`, `off-label use`, or `none`. Positive pairs are drawn from subgraph triples, while `none` examples are sampled from drug–disease pairs that do not appear in the subgraph.
- **Two-hop Verification (R1)** and **Two-hop Selection (R2)** are constructed from 2-hop contexts of the form Drug A → Disease B and Disease B → Disease C. Positive instances are created such that the queried Drug A → Disease C edge already exists in the subgraph. Negative instances preserve the same 2-hop scaffold but select a Disease C such that the queried edge is unsupported. R1 labels are sampled to achieve an approximately balanced ($\approx 50\%$) "Yes" rate, and R2 uses the same underlying 2-hop contexts.
- **Evidence-augmented Two-hop Verification (R1+E)** and **Evidence-augmented Two-hop Selection (R2+E)** extend the corresponding two-hop tasks by attaching short PrimeKG feature-table snippets for the involved entities. Controlled sanitization is applied to redact lexical forms overlapping with relation answer options, thereby reducing potential answer leakage.

The ground-truth answers for these questions are strictly defined by triples in the extracted PrimeKG

mental-health subgraph. Negative options are constructed in a task-specific but KG-consistent manner: FC negatives are created by type-matched head or tail replacement under the same relation and retained only when the perturbed triple is unsupported by the extracted subgraph; RP uses none examples drawn from drug–disease pairs that do not appear in the subgraph; and negative R1/R2 instances preserve the same 2-hop scaffold but query a drug–disease edge that is unsupported by the subgraph. The final benchmark comprises 1,847 ET items, 2,000 EC items, 4,000 FC items, 7 RT items, 1,634 RP items, and 1,200 items each for R1, R1+E, R2, and R2+E.

4 Experiments

4.1 Models

We evaluate 15 models spanning both closed- and open-source families: GPT-4.1, GPT-5.2-chat, GPT-4o, GPT-5-mini, GPT-5.1-chat, Qwen2.5-32B-Instruct, Mistral-7B-Instruct-v0.3, Qwen2.5-7B-Instruct, BioMistral-7B, Llama3-Med42-8B, DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Qwen-32B, Llama3.1-8B-Instruct, Meditron-7B, and Llama3-OpenBioLLM-8B. This set includes frontier GPT-series models, general-purpose open-source instruction-tuned models, and biomedical-domain variants. Representative technical reports and model cards for several evaluated families include GPT-4.1 (OpenAI, 2025), GPT-5-mini (OpenAI, 2026a), GPT-5.1-chat (OpenAI, 2026b), GPT-5.2-chat (OpenAI, 2026c), GPT-4o (OpenAI, 2024), Qwen2.5 (Qwen Team, 2024), Mistral-7B-Instruct-v0.3 (Mistral AI, 2024), BioMistral (Labrak et al., 2024), Med42 (Christophe et al., 2024), DeepSeek-R1 (DeepSeek-AI, 2025), Meditron (Chen et al., 2023), OpenBioLLM (Pal and Sankarasubbu, 2024), and Llama 3 (Llama Team AI @ Meta, 2024).

4.2 Evaluation Protocol

All tasks are evaluated under the same letter-only multiple-choice interface. For API-based models, we instruct the model to return a single option letter and apply strict answer parsing to recover one valid choice from the response. For local models, we use the same option-letter scoring setup as in the rest of the evaluation pipeline. Binary tasks use A/B labels rather than literal Yes/No to reduce lexical answer bias.

Because benchmark scoring under this setup depends on recovering a valid option letter, response validity is itself part of the evaluation problem in addition to raw task accuracy. We therefore treat output-format reliability as an important evaluation caveat in constrained multiple-choice assessment. Additional implementation details for benchmark construction, evidence sanitization, forced-choice local evaluation, API answer parsing, and randomness control are provided in Appendix D.

4.3 Metrics

We report task-level accuracy (%) and grouped averages for four benchmark dimensions:

$$\mathrm{Avg}_E = \mathrm{mean}(\mathrm{ET}, \mathrm{EC}), \tag{1}$$

$$\mathrm{Avg}_R = \mathrm{mean}(\mathrm{FC}, \mathrm{RT}, \mathrm{RP}), \tag{2}$$

$$\mathrm{Avg}_R^* = \mathrm{mean}(\mathrm{FC}, \mathrm{RP}), \tag{3}$$

$$\mathrm{Avg}_S = \mathrm{mean}(\mathrm{R1}, \mathrm{R2}), \tag{4}$$

$$\mathrm{Avg}_{S+E} = \mathrm{mean}(\mathrm{R1{+}E}, \mathrm{R2{+}E}), \tag{5}$$

$$\mathrm{Avg}_{All} = \mathrm{mean\ over\ all\ nine\ tasks}, \tag{6}$$

$$\mathrm{Avg}_{All}^* = \mathrm{mean\ over\ the\ eight\ tasks\ excluding\ RT}. \tag{7}$$

Because RT contains only one question per retained relation (7 items total), its score should be interpreted cautiously. In the results discussion below, we therefore emphasize the starred averages when drawing overall comparisons that are less influenced by the small RT set.

Beyond task accuracy, we also report graph-oriented coverage over entities, relations, and triples in the curated mental-health slice of PrimeKG. Specifically, we compute mean and degree-weighted correctness over entities and relations, together with a triple-level aggregate derived from entity and relation correctness. The none option in RP is treated as a task-specific no-relation label rather than as a KG relation and is therefore excluded from relation coverage. Full metric definitions are provided in Appendix C. In the current benchmark, all 1,847 entities and all 7 retained relations are measured for every model.

5 Results

5.1 Overall Performance

Table 1 reports the accuracy of the 15 selected LLMs on MHGraphBench. Because RT contains only 7 items, we focus first on the RT-excluded

Table 1: Benchmark accuracy (%) on the PrimeKG mental-health KG-to-QA tasks. We group tasks into four levels and report block averages: Avg_E=mean(ET, EC), Avg_R=mean(FC, RT, RP), Avg^*_R=mean(FC, RP), Avg_S=mean(R1, R2), Avg_{S+E}=mean(R1+E, R2+E), Avg_{All}=mean over all nine tasks, and Avg^*_{All}=mean over the eight tasks excluding RT. Because RT contains only 7 items, the starred summaries are often more informative for overall comparisons. Best values in each column are bolded; ties are jointly bolded.

Model	Entity			Relation					Subgraph			Evidence			Overall	
	ET	EC	Avg_E	FC	RT (n=7)	RP	Avg_R	Avg_{R^*}	R1	R2	Avg_S	R1+E	R2+E	Avg_{S+E}	Avg_{All}	Avg_{All^*}
GPT-4.1	97.62	91.85	94.73	63.32	**100.00**	54.96	72.76	59.14	57.58	64.00	**60.79**	**71.42**	61.50	**66.46**	**73.58**	**70.28**
GPT-5.2	98.05	90.10	94.07	63.35	**100.00**	**58.63**	**73.99**	**60.99**	50.17	**65.58**	57.88	61.08	**67.58**	64.33	72.73	69.32
GPT-4o	97.40	91.85	94.62	63.45	**100.00**	53.55	72.33	58.50	**62.08**	54.25	58.16	68.83	61.42	65.12	72.54	69.10
GPT-5-mini	**98.48**	91.75	95.12	**64.35**	**100.00**	57.28	73.88	60.81	50.25	59.83	55.04	59.67	65.42	62.55	71.89	68.38
GPT-5.1	98.27	**92.35**	**95.31**	62.50	**100.00**	58.08	73.53	60.29	50.17	60.83	55.50	60.00	64.42	62.21	71.85	68.33
Qwen2.5-32B	76.45	54.60	65.53	58.40	**100.00**	38.43	65.61	48.41	50.50	59.00	54.75	61.25	50.08	55.66	60.97	56.09
Mistral-7B	70.76	28.15	49.45	52.25	85.71	25.70	54.55	38.98	49.33	26.17	37.75	56.17	28.25	42.21	46.94	42.10
Qwen2.5-7B	49.32	36.95	43.14	53.18	57.14	25.89	45.40	39.53	50.08	38.92	44.50	56.75	37.75	47.25	45.11	43.61
BioMistral	29.24	14.80	22.02	49.90	28.57	20.99	33.15	35.45	53.92	22.92	38.42	53.17	38.42	45.80	34.66	35.42
Med42-8B	24.96	24.55	24.76	50.30	14.29	27.97	30.85	39.13	55.00	24.92	39.96	51.42	36.17	43.80	34.40	36.91
DeepSeek-R1-DQ-7B	9.42	20.40	14.91	49.60	28.57	32.62	36.93	41.11	49.75	27.17	38.46	49.75	25.00	37.38	32.48	32.96
DeepSeek-R1-DQ-32B	6.55	18.90	12.72	48.15	14.29	22.46	28.30	35.30	49.75	33.92	41.84	49.58	35.75	42.66	31.04	33.13
Llama3.1-8B	8.39	17.50	12.95	45.32	14.29	23.50	27.70	34.41	50.42	33.83	42.12	50.25	31.08	40.66	30.51	32.54
Meditron	20.25	18.00	19.12	50.00	14.29	25.34	29.88	37.67	49.17	21.42	35.30	51.00	22.08	36.54	30.17	32.16
OpenBioLLM-8B	16.57	23.35	19.96	50.25	0.00	23.68	24.64	36.97	49.75	24.58	37.16	56.50	24.33	40.41	29.89	33.63

Model abbreviations: GPT-5.2=GPT-5.2-chat; GPT-5.1=GPT-5.1-chat; GPT-4o=GPT-4o; GPT-5-mini=GPT-5-mini; Qwen2.5-32B=Qwen2.5-32B-Instruct; Qwen2.5-7B=Qwen2.5-7B-Instruct; Mistral-7B=Mistral-7B-Instruct-v0.3; BioMistral=BioMistral-7B; Med42-8B=Llama3-Med42-8B; DeepSeek-R1-DQ-7B/32B=DeepSeek-R1-Distill-Qwen-7B/32B; Llama3.1-8B=Llama3.1-8B-Instruct; Meditron=Meditron-7B; OpenBioLLM-8B=Llama3-OpenBioLLM-8B.

overall summary Avg^*_{All}. Under this summary, the strongest models in this evaluation are all GPT-series models: GPT-4.1 ranks first with $\mathrm{Avg}^*_{All} = 70.28\%$, followed by GPT-5.2-chat at 69.32% and GPT-4o at 69.10%. GPT-5-mini and GPT-5.1-chat follow closely at 68.38% and 68.33%, respectively. The RT-including summary Avg_{All} yields a similar top-level ordering, with GPT-4.1 achieving the highest score at 73.58%, followed by GPT-5.2-chat at 72.73% and GPT-4o at 72.54%.

Among open-source models, Qwen2.5-32B-Instruct is the strongest under both overall summaries, reaching 56.09% on Avg^*_{All} and 60.97% on Avg_{All}. Under the RT-excluded summary, this still leaves a gap of more than 12 percentage points relative to the leading GPT models. Below Qwen2.5-32B-Instruct, performance drops markedly: Mistral-7B-Instruct-v0.3 reaches 42.10% on Avg^*_{All} and Qwen2.5-7B-Instruct reaches 43.61%. The remaining models cluster in the low- to mid-30s on the RT-excluded overall summary. Taken together, these results suggest that the benchmark is challenging not only for smaller biomedical models, but for most open-source models in general.

5.2 Recognition vs. Relation Judgment

A central pattern in Table 1 is the gap between recognition-oriented tasks and relation-judgment tasks. For the top GPT-series models, recognition-oriented performance is very strong. GPT-5.1-chat achieves the highest Avg_E at 95.31%, closely followed by GPT-5-mini at 95.12%, GPT-4.1 at 94.73%, GPT-4o at 94.62%, and GPT-5.2-chat at 94.07%. ET is particularly strong, ranging from 97.40% to 98.48% across the top five models, while EC ranges from 90.10% to 92.35%. The RT subset is also saturated at 100.00% for these models, but this result should be interpreted cautiously because RT contains only 7 items and is therefore better treated as a small descriptive subset than as strong standalone evidence.

However, this recognition strength does not translate into equally strong performance on judgment-related tasks. The best RP score is only 58.63%, achieved by GPT-5.2-chat, followed by 58.08% for GPT-5.1-chat and 57.28% for GPT-5-mini. Even for the strongest models, these values remain far below ET and EC. The same separation appears in the grouped relation summaries. On the RT-excluded summary, GPT-5.2-chat reaches the highest Avg^*_R at 60.99%, followed by GPT-5-mini at 60.81% and GPT-5.1-chat at 60.29%. The RT-including summary Avg_R shows a similar ordering, but it should be interpreted more cautiously because it includes the 7-item RT subset.

This pattern indicates that a model may correctly identify entity types and relation schemas while still struggling to distinguish whether a drug is indicated, contraindicated, used off-label, or unsup-

ported for a target disorder.

5.3 Short-Chain Reasoning

Short disease-mediated reasoning is one of the task categories that most clearly separates stronger models from weaker ones in the benchmark. Even among the strongest models, subgraph reasoning scores remain well below entity-level performance. GPT-4o achieves the highest R1 score at 62.08%, while GPT-5.2-chat achieves the highest R2 score at 65.58%. The grouped reasoning score Avg_S is highest for GPT-4.1 at 60.79%, followed by GPT-4o at 58.16% and GPT-5.2-chat at 57.88%.

These values are notable because the reasoning tasks are tightly controlled: the 2-hop scaffold is explicitly provided, the answer space is constrained, and correctness is defined with respect to KG support. Even under these conditions, short-path composition remains substantially harder than ET or the small RT set. Among open-source models, the drop is sharper: Qwen2.5-32B-Instruct reaches 54.75% on Avg_S, whereas most others remain in the high-30s to mid-40s. This suggests that composing even two simple KG hops into a correct structured decision remains a major failure mode in constrained evaluation settings.

5.4 Evidence Augmentation Is Not Uniformly Helpful

Evidence augmentation affects models differently rather than providing a consistent benefit. On the positive side, several strong models improve when short KG-derived feature snippets are added. GPT-4.1 improves from 57.58% to 71.42% on R1, and GPT-4o improves from 62.08% to 68.83%. On the selection side, GPT-5.2-chat improves from 65.58% to 67.58% on R2, and GPT-5-mini improves from 59.83% to 65.42%. In grouped terms, GPT-4.1 achieves the best evidence-augmented reasoning score, with $\mathrm{Avg}_{S+E} = 66.46\%$, followed by GPT-4o at 65.12% and GPT-5.2-chat at 64.33%.

At the same time, evidence is not uniformly helpful across models. Qwen2.5-32B-Instruct, for example, improves strongly on R1, from 50.50% to 61.25%, but drops sharply on R2, from 59.00% to 50.08%, yielding only a modest evidence-grouped score of 55.66%. Smaller models also show inconsistent behavior, with some improving on one evidence-augmented task while remaining weak or deteriorating on the other. These results suggest that evidence augmentation is better interpreted as a diagnostic probe of whether a model can integrate short structured cues than as a universally corrective prompting strategy.

5.5 Response-Format Reliability

Response-format reliability is an important evaluation issue in MHGraphBench because all tasks use a constrained letter-only multiple-choice interface. In this setting, measured performance depends not only on whether a model knows the correct answer, but also on whether it can reliably return a single valid option letter that can be unambiguously parsed. This issue is especially relevant for API-based models, whose outputs may include extra explanation, multiple candidate letters, or other text that does not strictly follow the requested response format. As a result, benchmark accuracy can partially reflect output controllability in addition to underlying task knowledge, a broader concern that has also been noted in prior work on multiple-choice LLM evaluation (Wang et al., 2024; Pezeshkpour and Hruschka, 2024; Zheng et al., 2023).

This format issue also affects how chance-like scores on binary tasks such as FC or R1, where the positive and negative labels are approximately balanced, should be interpreted. Several weaker models remain close to 50% accuracy on FC or R1 while simultaneously performing poorly on ET, EC, or RP. Such behavior could reflect weak but genuine reasoning ability, but it may also arise in part from unstable constrained outputs, response biases, or instruction-following failures under the letter-only evaluation setup. For this reason, aggregate task accuracy alone can be misleading unless it is interpreted together with output-validity checks and inspection of prediction distributions.

5.6 Model-Family Observations

The results also offer a cautious perspective on biomedical-domain models. In this evaluation, biomedical or medically branded open-source models do not consistently outperform general-purpose instruction-tuned alternatives. BioMistral-7B (Labrak et al., 2024), Llama3-Med42-8B (Christophe et al., 2024), Meditron-7B (Chen et al., 2023), and Llama3-OpenBioLLM-8B (Pal and Sankarasubbu, 2024) all score below the strongest GPT models and below Qwen2.5-32B-Instruct on the overall summaries. Some of these models also exhibit unexpectedly weak entity-level performance: for example, Llama3-Med42-8B reaches only 24.76% on Avg_E, and Llama3-

Table 2: Compact knowledge coverage (%) on the PrimeKG mental-health subgraph. We report mean entity coverage, degree-weighted relation coverage, and triple coverage; full coverage metrics are provided in Appendix E.1. Models are sorted by $\mathrm{Cov}(T)$. Model abbreviations follow Table 1.

Model	$\mathrm{CovAvg}(E)$	$\mathrm{CovDeg}(R)$	$\mathrm{Cov}(T)$
GPT-5-mini	77.81	63.30	**65.27**
GPT-4o	77.36	61.18	64.77
GPT-4.1	77.91	61.24	63.57
GPT-5.1	77.84	59.62	61.48
GPT-5.2	63.92	44.56	54.97
Mistral-7B	51.19	49.21	53.83
Qwen2.5-7B	44.61	50.35	53.20
Qwen2.5-32B	61.47	55.09	52.31
Meditron	32.59	47.65	48.46
DeepSeek-R1-DQ-7B	29.24	47.25	47.34
DeepSeek-R1-DQ-32B	26.54	45.77	44.31
Med42-8B	38.23	48.34	38.89
OpenBioLLM-8B	34.02	48.10	37.26
BioMistral	36.74	46.34	37.12
Llama3.1-8B	26.67	42.98	36.71

OpenBioLLM-8B reaches 19.96%.

At the same time, these models are not uniformly weak across every dimension. Llama3-Med42-8B reaches 55.00% on R1, and BioMistral-7B reaches 53.92% on R1, despite their low entity scores. This uneven profile suggests that biomedical adaptation alone does not guarantee robust structured evaluation performance. However, this comparison should be interpreted cautiously, because the evaluated models also differ in parameter scale, base-model capability, instruction tuning, and output-format reliability. Taken together, the results suggest that performance in this benchmark reflects not only domain adaptation, but also the interaction among general model capacity, instruction following, constrained answer formats, and short reasoning requirements.

5.7 Knowledge Coverage

Coverage provides a complementary graph-wide view of model performance (Table 2; full metrics in Appendix E.1). GPT-5-mini achieves the strongest triple coverage, with $\mathrm{Cov}(T) = 65.27\%$, even though GPT-4.1 remains the top model by average task accuracy. This shows that benchmark averages and graph-wide coverage are not interchangeable. Coverage also changes the interpretation of open-source models: Qwen2.5-32B-Instruct is the best open-source model by Avg^{*}_{All}, but not by triple coverage, and GPT-5.2-chat shows lower coverage than the other GPT models despite ranking near the top on the main task table.

5.8 Refined Entity and Relation Analysis

To better understand where models succeed or fail, we also compute fine-grained entity- and relation-centric accuracy, with full tables reported in Appendix E.2. The fine-grained relation results show that `contraindication` is by far the hardest retained relation on average, whereas `indication` is comparatively easier. The fine-grained entity results further show that high benchmark incidence does not guarantee ease: anxiety-spectrum and psychotic-spectrum entities remain difficult despite their prominence in the benchmark. Taken together, these results suggest that model failures are concentrated in clinically important and diagnostically heterogeneous parts of the graph.

6 Discussion and Conclusion

We introduced MHGraphBench, a KG-grounded benchmark for evaluating mental-health biomedical knowledge in LLMs using a curated 1-hop mental-health subgraph of PrimeKG. The benchmark transforms KG-backed facts into nine standardized multiple-choice task families spanning entity recognition, relation judgment, and short two-hop reasoning, and complements task accuracy with graph-wide coverage and fine-grained entity- and relation-centric analyses.

Across 15 models, our results reveal a persistent recognition-to-judgment gap. Leading models achieve near-perfect performance on entity typing and very strong performance on the small relation-typing subset, yet they remain substantially weaker on relation prediction and short-chain reasoning. Coverage analysis further shows that average task accuracy and graph-wide coverage are not interchangeable: GPT-4.1 performs best on the main benchmark averages, whereas GPT-5-mini achieves the strongest triple coverage. Fine-grained analyses localize especially difficult graph regions, with clinically important relations such as `contraindication` and prominent entities such as *anxiety disorder* remaining challenging. We also find that evidence augmentation is not uniformly helpful across models and that response-format reliability can materially affect measured performance under constrained multiple-choice evaluation.

These findings suggest that broad biomedical competence should not be equated with reliable structured judgment in mental-health settings. They are consistent with recent evidence outside biomedicine. In a recent expert-led study

on high-temperature superconductivity, LLM systems grounded in curated literature outperformed more general systems, yet all evaluated systems still showed important limitations in expert-level scientific question answering (Guo et al., 2026). Taken together, these results suggest that current LLMs may read and organize scientific text fluently without consistently supporting the deeper judgment required for expert reasoning.

More broadly, our results show that KG-grounded benchmarking provides an interpretable and reproducible way to study what LLMs capture about mental-health biomedical structure, while highlighting limitations in safety-relevant relation distinctions and controlled reasoning. Future work should move toward more challenging but still controlled benchmarks that better connect structured knowledge evaluation with clinically relevant mental-health decision support. MHGraphBench should therefore be interpreted as a structured evaluation of a curated KG slice rather than as a direct assessment of real-world clinical safety.

Limitations

Our benchmark inherits the coverage limits and curation decisions of PrimeKG as well as those of our mental-health subgraph extraction process. As a result, the task suite is intentionally scoped and does not capture the full breadth of psychiatric care, longitudinal patient context, or individualized treatment decision-making.

All labels are defined with respect to the extracted PrimeKG mental-health subgraph. Because biomedical knowledge and clinical guidelines evolve over time, these KG-based labels may be incomplete or may lag behind the most up-to-date evidence. Accordingly, the benchmark measures agreement with a curated KG slice rather than absolute clinical truth.

In addition, we do not perform manual or expert validation of sampled benchmark items, negative instances, or evidence snippets beyond the KG-grounded construction pipeline itself. This means that benchmark validity depends on the quality of the underlying graph, the extraction procedure, and the task-generation rules. In particular, an edge being unsupported in the extracted subgraph should not be interpreted as evidence that the corresponding claim is false in the real world; it indicates only that the queried relation is absent from the curated benchmark graph. Similarly, although evidence snippets are sanitized to reduce direct answer leakage, they are not externally adjudicated by domain experts for completeness, clinical appropriateness, or real-world decision support value.

Although coverage and fine-grained analyses provide a richer view than average task accuracy alone, they still depend on benchmark construction choices and on how task items involve particular graph components. These analyses help localize strengths and weaknesses, but they should not be interpreted as exhaustive measurements of mental-health biomedical knowledge.

Finally, because evaluation relies on constrained multiple-choice outputs, models that fail to follow answer-format instructions may be penalized for reasons partly independent of their underlying biomedical reasoning ability. This is both a limitation and an empirical finding of the benchmark: output controllability is entangled with measured performance.

Ethics Statement

This work does not evaluate clinical safety, real-world mental-health decision-making, or patient-specific treatment appropriateness. The benchmark should not be used as a substitute for expert oversight, especially in settings involving treatment boundaries, contraindications, or crisis-related decisions (Agarwal et al., 2024; Zhu et al., 2025). More broadly, our results should be interpreted as a structured evaluation of KG-grounded mental-health biomedical knowledge with respect to a curated mental-health subgraph, rather than as evidence of clinical validity, real-world safety, or readiness for clinical deployment.

Acknowledgments

This research was, in part, funded by the Advanced Research Projects Agency for Health (ARPA-H). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Government.

References

Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. 2024. MedHalu: Hallucinations in responses to healthcare queries by large language models. *arXiv preprint arXiv:2409.19492*.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. MedBench: A large-scale Chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17709–17717.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70B: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical LLMs. *arXiv preprint arXiv:2408.06142*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Sebastian Freidel and Emanuel Schwarz. 2025. Knowledge graphs in psychiatric research: Potential applications and future perspectives. *Acta Psychiatrica Scandinavica*, 151(3):180–191.

Shan Gao, Kaixian Yu, Yue Yang, Sheng Yu, Chenglong Shi, Xueqin Wang, Niansheng Tang, and Hongtu Zhu. 2025. Large language model powered knowledge graph construction for mental health exploration. *Nature Communications*, 16(1):7526.

GBD 2019 Mental Disorders Collaborators. 2022. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*, 9(2):137–150.

Haoyu Guo, Maria Tikhanovskaya, Paul Raccuglia, Alexey Vlaskin, Chris Co, Daniel J. Liebling, Scott Ellsworth, Matthew Abraham, Elizabeth Dorfman, N. P. Armitage, Chunhan Feng, Antoine Georges, Olivier Gingras, Dominik Kiese, Steven A. Kivelson, Vadim Oganesyan, B. J. Ramshaw, Subir Sachdev, T. Senthil, and 4 others. 2026. Expert evaluation of LLM world models: A high-T_c superconductivity case study. *Proceedings of the National Academy of Sciences*, 123(11):e2533676123.

Vassilis N. Ioannidis, Xiang Song, Saurav Manchanda, Mufei Li, Xiaoqin Pan, Da Zheng, Xia Ning, Xiangxiang Zeng, and George Karypis. 2020. DRKG: Drug repurposing knowledge graph for COVID-19. https://github.com/gnn4dr/DRKG. Accessed: 2026-03-17.

Usman Iqbal, Afifa Tanweer, Annisa Ristya Rahmanti, David Greenfield, Leon Tsung-Ju Lee, and Yu-Chuan Jack Li. 2025. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. *Journal of Biomedical Science*, 32(1):45.

Mietta Kyrios, Jesse Levido, Daniel Talbot, and Anthony Harris. 2024. Off-label prescribing of psychotropics in a psychiatric patient population in australia. *Australasian Psychiatry*, 32(3):196–200.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Jianning Li, Amin Dada, Behrus Puladi, Jens Kleesiek, and Jan Egger. 2024. ChatGPT in healthcare: a taxonomy and systematic review. *Computer Methods and Programs in Biomedicine*, 245:108013.

Yahan Li, Jifan Yao, John Bosco S Bunyi, Adam C Frank, Angel Hsing-Chi Hwang, and Ruishan Liu. 2025. CounselBench: A large-scale expert evaluation and adversarial benchmarking of large language models in mental health question answering. *arXiv preprint arXiv:2506.08584*.

Llama Team AI @ Meta. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Elan Markowitz, Krupa Galiya, Greg Ver Steeg, and Aram Galstyan. 2025. KG-LLM-Bench: A scalable benchmark for evaluating LLM reasoning on textualized knowledge graphs. *arXiv preprint arXiv:2504.07087*.

Mistral AI. 2024. mistralai/mistral-7b-instruct-v0.3 (hugging face model card). https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. Accessed: 2026-01-13.

Nick Obradovich, Sahib S Khalsa, Waqas U Khan, Jina Suh, Roy H Perlis, Olusola Ajilore, and Martin P Paulus. 2024. Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry and Neuroscience*, 2(1):8.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

OpenAI. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2026-03-17.

OpenAI. 2026a. GPT-5 mini model. https://developers.openai.com/api/docs/models/gpt-5-mini. OpenAI API model documentation; Accessed: 2026-03-17.

OpenAI. 2026b. GPT-5.1 chat model. https://developers.openai.com/api/docs/models/gpt-5.1-chat-latest. OpenAI API model documentation; Accessed: 2026-03-17.

OpenAI. 2026c. GPT-5.2 chat model. https://developers.openai.com/api/docs/models/gpt-5.2-chat-latest. OpenAI API model documentation; Accessed: 2026-03-17.

Ankit Pal and Malaikannan Sankarasubbu. 2024. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/blog/aaditya/openbiollm. Accessed: 2026-03-17.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Haakon Gravanti Rosland, Gro Janne Wergeland, and Lone Holst. 2025. Off-label use of psychotropic drugs in youth. *BMC Psychiatry*, 25(1):739.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, and 48 others. 2024. Capabilities of Gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Mikhail Salnikov, Hai Le, Prateek Rajput, Irina Nikishina, Pavel Braslavski, Valentin Malykh, and Alexander Panchenko. 2023. Large language models meet knowledge graphs to answer factoid questions. *arXiv preprint arXiv:2310.02166*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Hoyun Song, Migyeong Kang, Jisu Shin, Jihyun Kim, Chanbi Park, Hangyeol Yoo, Jihyun An, Alice Oh, Jinyoung Han, and KyungTae Lim. 2026. MentalBench: A benchmark for evaluating psychiatric diagnostic capability of large language models. *arXiv preprint arXiv:2602.12871*.

Table 3: Summary statistics of the PrimeKG mental-health subgraph used in this study. HP = high-precision; canon./dedup. = canonicalization and deduplication.

Item	Count
HP candidate disease nodes	44
Post-curation exclusions	2
Final seed disease nodes	42
Raw seed-touching edges	9,242
Canon./dedup. unique triples	4,621
Unique entities	1,847
Retained relation types	7

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M Ni, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *arXiv preprint arXiv:2307.07697*.

Sebastian Volkmer, Andreas Meyer-Lindenberg, and Emanuel Schwarz. 2024. Large language models in psychiatry: Opportunities and challenges. *Psychiatry Research*, 339:116026.

Haochun Wang, Sendong Zhao, Zewen Qiang, Bing Qin, and Ting Liu. 2024. Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models. *CoRR*.

Zixin Xiong, Ziteng Wang, Haotian Fan, Xinjie Zhang, and Wenxuan Wang. 2026. TrustMH-Bench: A comprehensive benchmark for evaluating the trustworthiness of large language models in mental health. *arXiv preprint arXiv:2603.03047*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882*.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing Long, Yefeng Zheng, and Xian Wu. 2025. Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769.

A Subgraph Statistics

This appendix section summarizes the size and composition of the curated PrimeKG mental-health subgraph used throughout the benchmark.

A.1 Summary Statistics

A.2 Entity and Relation Breakdown

B Mental-Health Seed Disease Nodes

This appendix section documents how the psychiatric seed disease list was defined and reports the final set of seed nodes used for subgraph extraction.

Table 4: Entity-type and relation-type counts in the PrimeKG mental-health subgraph.

Entity type	**Count**
Disease	75
Drug	345
Gene/protein	1,326
Effect/phenotype	66
Exposure	35
Relation type	**Count**
`disease_protein`	3,565
`contraindication`	556
`indication`	234
`disease_disease`	84
`disease_phenotype_positive`	82
`off-label use`	54
`exposure_disease`	46

B.1 Seed Selection Procedure

We began with a manually curated high-precision candidate seed list of 44 PrimeKG disease nodes, stored in `mental_health_seed_diseases_HP.csv`. PrimeKG disease nodes are encoded using terms from the Mondo Disease Ontology (MONDO) and grouped into clinically meaningful disease nodes during PrimeKG construction (Chandak et al., 2023). Candidate selection was restricted to psychiatric disorders and closely related conditions intended to define the benchmark scope. We then applied two manual post-curation exclusions: one out-of-scope entry (*X-linked intellectual disability-psychosis-macroorchidism syndrome*), which appeared in the initial candidate list but was excluded at post-curation because it was outside the intended psychiatric benchmark scope, and one outdated entry (*multiple personality disorder*). This yielded the final set of 42 psychiatric seed disease nodes used for subgraph extraction.

B.2 Final Seed List

The final seed list, stored in `mental_health_seed_diseases_FINAL.csv`, is shown in Table 5.

C Coverage Metric Definitions

This section defines the coverage metrics used in the main paper. We first define per-entity and per-relation correctness and then derive entity-, relation-, and triple-level coverage scores.

C.1 Per-Entity and Per-Relation Correctness

Beyond task accuracy, we quantify how well a model covers the mental-health slice of PrimeKG in terms of correctness over entities, relations, and triples. Let the curated mental-health graph be $G = (V, R, T)$, where V is the entity set, R is the relation set, and $T \subseteq V \times R \times V$ is the triple set.

For each entity $e \in V$, let $\mathcal{Q}(e)$ denote the set of benchmark items whose gold annotation involves e. We define empirical entity correctness as

$$a_E(e) = \frac{1}{|\mathcal{Q}(e)|} \sum_{q \in \mathcal{Q}(e)} \mathbf{1}[\hat{y}_q = y_q]. \tag{8}$$

Similarly, for each relation $r \in R$, let $\mathcal{Q}(r)$ denote the set of benchmark items whose gold annotation involves r, and define empirical relation correctness as

$$a_R(r) = \frac{1}{|\mathcal{Q}(r)|} \sum_{q \in \mathcal{Q}(r)} \mathbf{1}[\hat{y}_q = y_q]. \tag{9}$$

The `none` option is treated as a task-specific no-relation label rather than a KG relation. It is therefore excluded from relation coverage and contributes only indirectly through entity-level correctness.

C.2 Coverage Scores

We then define five coverage scores. Mean entity coverage is

$$\mathrm{CovAvg}(E) = \frac{1}{|V_{\mathrm{meas}}|} \sum_{e \in V_{\mathrm{meas}}} a_E(e). \tag{10}$$

For degree-weighted entity coverage, we define entity degree as the number of incident triples in T,

$$\deg(e) = |\{(h, r, t) \in T : h = e \text{ or } t = e\}|, \tag{11}$$

with normalizing constant

$$Z_E = \sum_{e \in V} \deg(e). \tag{12}$$

The resulting degree-weighted entity coverage is

$$\mathrm{CovDeg}(E) = \frac{1}{Z_E} \sum_{e \in V} \deg(e)\, a_E(e). \tag{13}$$

For relations, mean relation coverage is

$$\mathrm{CovAvg}(R) = \frac{1}{|R_{\mathrm{meas}}|} \sum_{r \in R_{\mathrm{meas}}} a_R(r), \tag{14}$$

Table 5: Final set of 42 psychiatric seed disease nodes used to extract the PrimeKG mental-health subgraph.

Node index	Node name	Node index	Node name
27933	anxiety disorder	84190	binge eating disorder
28249	major affective disorder	84195	narcissistic personality disorder
28313	schizophrenia	84204	mixed anxiety and depressive disorder
28592	bulimia nervosa, susceptibility to	84208	alcoholic psychosis
28899	obsessive-compulsive disorder	84226	postpartum depression
32965	early-onset schizophrenia	84288	antisocial personality disorder
33572	psychotic disorder	84294	paranoid schizophrenia
35758	specific phobia	94658	neurotic depression
36021	personality disorder	95043	avoidant personality disorder
38242	bipolar disorder	95044	dependent personality disorder
38945	bulimia nervosa	95390	schizotypal personality disorder
38957	major depressive disorder	95419	schizoid personality disorder
39833	agoraphobia	95420	paranoid personality disorder
83763	manic bipolar affective disorder	95557	atypical depressive disorder
83779	schizoaffective disorder	95941	histrionic personality disorder (disease)
83840	unipolar depression	96890	substance-induced psychosis
83841	endogenous depression	97059	treatment-refractory schizophrenia
83842	anorexia nervosa	97074	methamphetamine-induced psychosis
83903	post-traumatic stress disorder	97811	anorexia nervosa, susceptibility to, 1
83904	social phobia	98548	postpartum psychosis
83910	drug psychosis	99866	panic disorder without or with agoraphobia

where relation degree is defined as the number of triples in T that use relation r,

$$\deg(r) = \left|\{(h, r', t) \in T : r' = r\}\right|. \quad (15)$$

The corresponding degree-weighted relation coverage is

$$\mathrm{CovDeg}(R) = \frac{1}{|T|} \sum_{r \in R} \deg(r)\, a_R(r). \quad (16)$$

Finally, for each triple $(h, r, t) \in T$, we define an auxiliary triple score

$$s(h, r, t) = \frac{a_E(h) + a_R(r) + a_E(t)}{3}, \quad (17)$$

and triple coverage as

$$\mathrm{Cov}(T) = \frac{1}{|T|} \sum_{(h,r,t) \in T} s(h, r, t). \quad (18)$$

Here, V_{meas} and R_{meas} denote the sets of measured entities and relations. In the current benchmark, all 1,847 entities and all 7 retained relations are measured for every model.

D Benchmark Construction, Evidence, and Evaluation Details

This appendix section provides implementation-level details that extend, rather than repeat, the benchmark overview in Section 3 and the evaluation description in Section 4.2. Unless otherwise noted, all statements in this section are derived directly from the benchmark-generation and evaluation code used in our experiments.

D.1 Benchmark Construction Details

The benchmark is generated from PrimeKG using four input files: `kg.csv`, `mental_health_seed_diseases_FINAL.csv`, `disease_features.tab`, `drug_features.tab`. The final seed file contains the 42 psychiatric seed disease nodes reported in Appendix B. Starting from these seeds, we extract all 1-hop seed-touching edges from `kg.csv`, retain only a fixed set of seven clinically salient relations, canonicalize relation direction to predefined head/tail type signatures, and deduplicate the resulting triples.

The retained relations are `disease_protein`, `contraindication`, `indication`, `off-label use`, `disease_disease`, `disease_phenotype_positive`, and `exposure_disease`. Canonical relation signatures are fixed during preprocessing: `disease_protein` is canonicalized as disease→gene/protein; `contraindication`, `indication`, and `off-label use` as drug→disease; `disease_disease` as disease→disease; `disease_phenotype_positive` as disease→effect/phenotype; and `exposure_disease` as exposure→disease. For `disease_disease`, symmetric duplicates are additionally removed by lexicographic canonicalization of the two disease names.

All task instances are generated programmatically from fixed English templates and use a unified letter-only answer interface. Entity Typing (ET) is a 5-way multiple-choice question over entity types. Entity Clustering (EC) is constructed

as an odd-one-out task with four entities of one type and one entity of another type. Relation Typing (RT) asks for the dominant head→tail type signature of a relation. Relation Prediction (RP) is a 4-way multiple-choice task over `indication`, `contraindication`, `off-label use`, and `none`. Two-hop Verification (R1) is a binary A/B task, and Two-hop Selection (R2) is a 4-way multiple-choice task. Evidence-augmented variants (R1+E and R2+E) use the same underlying task structure but append short feature-table evidence snippets to the question.

The two-hop tasks are constructed from contexts in which a drug is linked to disease A by one of the three drug–disease usage relations and disease A is linked to disease B by `disease_disease`. Positive R1 instances are those for which the queried drug–disease B relation already exists in the retained subgraph. Negative R1 instances preserve the same 2-hop scaffold but require that the queried drug–disease B relation be absent from the subgraph. The generator targets an approximately balanced R1 label distribution with a 50% Yes rate and does not allow the intermediate disease and queried disease to be identical. R2 instances are built from the same contexts and ask the model to select the most appropriate relation label for the queried drug–disease pair.

The Fact Checking (FC) task is balanced *per relation*. For each retained relation r, the benchmark generator samples positive triples from that relation and constructs an equal number of negative examples under the same relation. FC negatives are created by type-matched head or tail replacement while preserving the original relation label, and the perturbed triple is kept only if it does not appear in the retained mental-health subgraph. This design avoids relation-replacement negatives and makes per-relation FC behavior easier to interpret.

D.2 Evidence Construction and Sanitization

Evidence-augmented tasks draw text snippets from `disease_features.tab` and `drug_features.tab`. For disease nodes, the pipeline attempts to use available fields such as `mondo_name`, `mondo_definition`, `umls_description`, `orphanet_clinical_description`, `mayo_symptoms`, `mayo_causes`, `mayo_risk_factors`, and `orphanet_management_and_treatment`. For drug nodes, it attempts to use fields such as `description`, `indication`, `mechanism_of_action`, `pharmacodynamics`, `half_life`, `state`, and `category`. When multiple rows are available for the same node, the generator keeps the first non-empty value for each field. Long text fields are truncated to at most 220 characters per field before question assembly.

To reduce answer leakage, evidence text is sanitized before insertion into the question. The sanitization step redacts lexical forms overlapping with relation answer options, including patterns matching `indication`, `contraindication`, and `off-label`, and replaces them with a neutral placeholder `[REL]`. In addition, the original drug-table field name `indication` is rendered with the more neutral display label *Clinical use* when evidence blocks are assembled. The evidence block is then attached in a fixed order: drug evidence first, followed by evidence for disease A and disease B.

D.3 Evaluation Protocol Details

All benchmark tasks are evaluated under the same letter-only interface. For binary tasks, the benchmark uses A/B labels rather than literal Yes/No strings, with A corresponding to Yes and B corresponding to No. The evaluation code records per-task accuracy for all tasks and additionally computes diagnostic quantities such as prediction-A rate and balanced accuracy for A/B tasks.

For local Hugging Face models, evaluation is performed by forced-choice scoring rather than free-form generation. The prompt is constructed by appending the fixed anchor `\nAnswer:` to each benchmark question. If the tokenizer provides a chat template, the prompt is wrapped using the tokenizer's chat-template interface; otherwise, the raw question text is used directly. Candidate answer letters are scored through multi-token log-probability accumulation over several surface forms, including `(A)`, bare-letter forms with a trailing newline, parenthesized-letter forms with a trailing newline, and several short prefixes. Scores for multiple surface forms corresponding to the same letter are merged by taking the maximum score for that letter, and the highest-scoring allowed option is selected as the model prediction. This procedure makes local evaluation deterministic given fixed model weights, tokenizer behavior, and benchmark inputs.

The local evaluation code uses batch size 1, maximum sequence length 4096, `bfloat16` model loading, and automatic device placement via

device_map="auto". Model loading is performed with trust_remote_code=True. For API-based models, the evaluation pipeline queries the model with temperature set to 0 and a maximum completion length of 120 tokens, then applies strict answer parsing to recover a single option letter from the returned text. The parser first searches for explicit answer patterns such as Answer: (X) and then falls back to leading-letter or in-text letter matching when necessary. Because some API-based models occasionally return outputs that do not perfectly follow the requested constrained format, response validity is itself an important part of the measured evaluation behavior.

D.4 Randomness and Deterministic Settings

Benchmark generation uses a fixed Python random seed of 42. This seed controls the sampling operations used during task construction, including option shuffling, entity selection, negative sampling, and task-instance ordering. Local model evaluation by forced-choice scoring does not sample from the model and is therefore deterministic given fixed model weights, tokenizer behavior, and benchmark inputs. API-based evaluation is configured with temperature 0 to reduce generation variability.

E Extended Coverage and Fine-Grained Results

This appendix section reports the full coverage tables and the detailed entity- and relation-level analyses that complement the compact summaries in the main text.

E.1 Knowledge Coverage Results

Table 6 provides a complementary graph-wide view of performance. Because all 1,847 entities and all 7 retained relations are measured for every model, these metrics summarize behavior over the full benchmark graph rather than over a partial subset. The strongest triple coverage is achieved by GPT-5-mini, with $\mathrm{Cov}(T) = 65.27\%$, followed by GPT-4o at 64.77% and GPT-4.1 at 63.57%. This ranking differs from the main task average, where GPT-4.1 is the top model. The discrepancy suggests that average task accuracy and graph-wide coverage capture different aspects of model behavior.

At the entity level, GPT-4.1 achieves the highest mean entity coverage, with $\mathrm{CovAvg}(E) = 77.91\%$, closely followed by GPT-5.1-chat and GPT-5-mini. However, GPT-5-mini attains the highest degree-weighted relation coverage, with $\mathrm{CovDeg}(R) = 63.30\%$, which helps explain why it leads on triple coverage. In other words, GPT-5-mini is especially strong on graph-central relation mass, even though GPT-4.1 remains slightly stronger on average benchmark accuracy.

Coverage also changes the interpretation of open-source models. Qwen2.5-32B-Instruct is the best open-source model by Avg_{All} in Table 1, but it does not have the strongest open-source triple coverage. Mistral-7B-Instruct-v0.3 and Qwen2.5-7B-Instruct both slightly exceed Qwen2.5-32B-Instruct on $\mathrm{Cov}(T)$, suggesting that correctness on high-degree graph components can differ from overall task-level performance. GPT-5.2-chat is another notable case: despite ranking near the top on the main benchmark averages, its coverage scores are substantially lower than those of the other GPT models, indicating that its correctness is less evenly distributed across the graph.

E.2 Fine-Grained Entity and Relation Analysis

For a model m and relation r, we define

$$\mathrm{Acc}_m(r) = \frac{c_m(r)}{n(r)}, \tag{19}$$

where $n(r)$ is the number of benchmark items whose gold annotation involves relation r, and $c_m(r)$ is the number of those items answered correctly. Similarly, for an entity e, we define

$$\mathrm{Acc}_m(e) = \frac{c_m(e)}{n(e)}. \tag{20}$$

We report high-incidence entities and retained relations because these components exert a strong influence on observed benchmark behavior and help localize where model performance concentrates.

The fine-grained relation results in Table 7 reveal substantial variation across clinically salient relation families. The easiest relation on average is indication, with a mean accuracy of 59.1%, followed by disease_disease at 56.6% and exposure_disease at 55.7%. In contrast, contraindication is by far the hardest relation, with a mean accuracy of only 35.8%. This is notable because contraindication marks one of the most safety-sensitive boundaries in mental-health pharmacotherapy. Its difficulty helps explain why RP remains much weaker than ET or the small RT set, even for the strongest models.

Table 6: Knowledge coverage (%) on the PrimeKG mental-health subgraph. Higher is better. All models measure all 1,847 entities and all 7 retained relations. Models are sorted by $\mathrm{Cov}(T)$.

Model	$\mathrm{CovAvg}(E)$	$\mathrm{CovDeg}(E)$	$\mathrm{CovAvg}(R)$	$\mathrm{CovDeg}(R)$	$\mathrm{Cov}(T)$	**Meas. E**	**Meas. R**
GPT-5-mini	77.81	66.26	64.34	63.30	**65.27**	1847	7
GPT-4o	77.36	66.57	61.23	61.18	64.77	1847	7
GPT-4.1	77.91	64.74	61.33	61.24	63.57	1847	7
GPT-5.1-chat	77.84	62.40	62.30	59.62	61.48	1847	7
GPT-5.2-chat	63.92	60.17	44.86	44.56	54.97	1847	7
Mistral-7B-Instruct-v0.3	51.19	56.14	52.52	49.21	53.83	1847	7
Qwen2.5-7B-Instruct	44.61	54.63	54.92	50.35	53.20	1847	7
Qwen2.5-32B-Instruct	61.47	50.92	61.62	55.09	52.31	1847	7
Meditron-7B	32.59	48.87	44.68	47.65	48.46	1847	7
DeepSeek-R1-Distill-Qwen-7B	29.24	47.39	44.89	47.25	47.34	1847	7
DeepSeek-R1-Distill-Qwen-32B	26.54	43.58	46.11	45.77	44.31	1847	7
Llama3-Med42-8B	38.23	34.17	46.46	48.34	38.89	1847	7
Llama3-OpenBioLLM-8B	34.02	31.84	45.81	48.10	37.26	1847	7
BioMistral-7B	36.74	32.50	45.84	46.34	37.12	1847	7
Llama3.1-8B-Instruct	26.67	33.58	44.39	42.98	36.71	1847	7

Table 7: Fine-grained accuracy (%) on retained relations in the mental-health subgraph. Relations are ordered by benchmark incidence in the evaluation set. The highest model score in each row is bolded. The final column reports the mean accuracy across models.

Relation	**Items**	**BioM7B**	**DS32**	**DS7**	**Med42**	**OpenBio**	**Meditron**	**Mistral7B**	**Q32**	**Q7**	**GPT4.1**	**GPT4o**	**GPT5m**	**GPT5.1**	**GPT5.2**	**L3.1-8B**	**Mean**
contraindication	3348	24.5	38.1	35.7	39.8	38.9	36.8	27.0	33.1	30.3	**43.6**	41.8	37.8	38.0	31.8	39.8	35.8
disease_protein	2983	49.5	47.4	49.7	49.9	50.1	50.0	50.6	56.1	51.0	62.9	62.9	**65.9**	61.4	46.0	43.3	53.1
indication	1629	45.0	32.7	35.8	45.0	39.7	38.0	75.3	80.4	**83.9**	78.1	81.9	81.7	77.9	51.9	38.6	59.1
off-label use	239	50.5	49.5	41.7	41.7	41.7	39.8	40.8	66.0	49.5	63.1	62.1	67.0	**70.9**	49.5	41.7	51.7
disease_disease	95	50.5	48.4	49.5	50.5	51.6	49.5	65.3	**72.6**	55.8	63.2	64.2	64.2	69.5	44.2	49.5	56.6
disease_phenotype_positive	93	51.6	52.7	49.5	50.5	49.5	49.5	51.6	**61.3**	47.3	58.1	57.0	59.1	58.1	39.8	54.8	52.7
exposure_disease	63	49.2	54.0	52.4	47.6	49.2	49.2	57.1	61.9	66.7	60.3	58.7	**74.6**	60.3	50.8	42.9	55.7

Model abbreviations: BioM7B=BioMistral-7B; DS32=DeepSeek-R1-Distill-Qwen-32B; DS7=DeepSeek-R1-Distill-Qwen-7B; Med42=Llama3-Med42-8B; OpenBio=Llama3-OpenBioLLM-8B; Meditron=Meditron-7B; Mistral7B=Mistral-7B-Instruct-v0.3; Q32=Qwen2.5-32B-Instruct; Q7=Qwen2.5-7B-Instruct; GPT4.1=GPT-4.1; GPT4o=GPT-4o; GPT5m=GPT-5-mini; GPT5.1=GPT-5.1-chat; GPT5.2=GPT-5.2-chat; L3.1-8B=Llama3.1-8B-Instruct.

The relation-level results also show that overall model quality does not imply uniform strength across relation families. GPT-4.1 is strongest on `contraindication` at 43.6%, GPT-5-mini is strongest on `disease_protein` and `exposure_disease` at 65.9% and 74.6%, respectively, GPT-5.1-chat is strongest on `off-label use` at 70.9%, and Qwen2.5-7B-Instruct is strongest on `indication` at 83.9%. Qwen2.5-32B-Instruct achieves the best score on `disease_disease` and `disease_phenotype_positive`. These row-wise inversions reinforce the idea that performance is relation-dependent rather than uniformly ordered by overall benchmark average.

The fine-grained entity results in Table 8 show that high benchmark incidence does not guarantee ease. Among the Top-15 high-incidence mental-health entities, the highest mean accuracies are observed for *major depressive disorder* (57.1%), *schizophrenia* (55.7%), and *unipolar depression* (54.7%). However, several prominent entities remain difficult, most notably *anxiety disorder*, which has the highest benchmark incidence in this subset but only 40.1% mean accuracy. *Psychotic disorder* (43.0%) and *schizoaffective disorder* (40.4%) are also challenging despite their high incidence. At the lower end, *anorexia nervosa* (37.8%) and *obsessive-compulsive disorder* (38.6%) are among the hardest entities in this set.

Taken together, the fine-grained analyses suggest that model failures are concentrated in clinically important and diagnostically heterogeneous parts of the graph. High-incidence depressive disorders are often handled reasonably well by the stronger models, but anxiety-spectrum, psychotic-spectrum, and medication-boundary distinctions remain much less stable.

Table 8: Fine-grained accuracy (%) on Top-15 high-incidence mental-health entities in the benchmark. Entities are ordered by benchmark incidence in the evaluation set. The final column reports the mean accuracy across models.

Entity	Items	BioM7B	DS32	DS7	Med42	OpenBio	Meditron	Mistral7B	Q32	Q7	GPT4.1	GPT4o	GPT5m	GPT5.1	GPT5.2	L3.1-8B	Mean
anxiety disorder	1178	24.8	57.0	53.9	28.6	28.3	57.7	46.4	30.0	54.2	36.5	41.4	38.3	30.3	45.7	28.7	40.1
bipolar disorder	758	28.5	34.6	61.7	30.7	30.2	62.0	60.2	40.0	62.2	59.5	59.8	57.6	51.7	61.2	31.7	48.8
major affective disorder	681	28.8	61.8	62.6	33.1	33.3	61.3	58.5	36.9	58.0	48.9	53.2	50.4	44.0	59.0	59.3	49.9
schizophrenia	673	24.2	71.1	71.4	23.5	23.1	71.1	74.4	48.7	74.2	62.7	65.2	62.2	55.6	66.7	42.0	55.7
psychotic disorder	609	28.2	34.8	39.2	37.0	33.7	37.0	24.9	51.4	35.9	72.4	61.9	53.0	60.8	45.3	29.8	43.0
schizoaffective disorder	556	30.1	40.7	39.4	39.4	36.1	42.6	32.4	41.2	41.7	47.2	43.1	49.1	46.8	39.8	36.1	40.4
unipolar depression	480	30.4	59.8	63.1	31.8	31.5	61.9	67.3	50.0	68.5	62.2	70.5	69.0	60.1	61.9	32.7	54.7
major depressive disorder	475	36.1	55.1	55.5	38.8	35.7	55.1	67.8	55.9	64.8	62.1	69.6	75.3	70.5	59.5	55.1	57.1
endogenous depression	358	31.8	56.6	58.5	32.2	30.6	57.8	64.0	51.2	65.5	57.0	61.2	63.6	55.8	57.8	31.8	51.7
obsessive-compulsive disorder	262	36.7	30.0	36.7	40.0	41.1	33.3	25.6	42.2	35.6	48.9	35.6	52.2	53.3	43.3	24.4	38.6
anorexia nervosa	187	26.7	32.0	32.0	40.0	36.0	34.7	21.3	40.0	24.0	46.7	48.0	54.7	52.0	41.3	37.3	37.8
drug psychosis	178	32.0	41.8	40.2	39.3	38.5	43.4	32.8	41.8	35.2	49.2	55.7	39.3	40.2	41.0	36.9	40.5
manic bipolar affective disorder	163	41.7	43.7	43.7	47.6	42.7	49.5	49.5	55.3	49.5	63.1	62.1	65.0	66.0	50.5	47.6	51.8
mixed anxiety and depressive disorder	135	56.9	21.6	21.6	62.7	56.9	23.5	39.2	68.6	35.3	62.7	58.8	70.6	70.6	37.3	31.4	47.8
neurotic depression	133	49.1	42.1	38.6	42.1	40.4	42.1	36.8	42.1	40.4	70.2	73.7	63.2	59.6	49.1	33.3	48.2

Model abbreviations are identical to Table 7.