No Man Is an Island: Explainable Graph-LLM Agents for Real-Time Clinical Reasoning

Ratna Kandala[1] Akshata Kishore Moharir[2] Niva Manchanda[3] Samantha Adorno [4]
[1,3,4] University of Kansas
[2]Independent Researcher
[ratnanirupama,akshatankishore5]@gmail.com, nmanchanda@ku.edu,
samantha.adorno30@gmail.com

Abstract

Mental health is inherently relational, encompassing interactions between social factors, longitudinal treatment dynamics, and interdependencies between symptoms. However, current AI systems fail to capture this complexity by treating patient data as independent features. This creates a fundamental mismatch between the relational nature of mental health and the capabilities of current AI systems. The result is models that miss critical interdependencies, such as how social isolation exacerbates depressive symptoms or how medication side effects interact with existing conditions, limiting their clinical utility and accuracy. Addressing this mismatch requires rethinking how AI systems represent and reason over clinical data. This position paper argues for a fundamental architectural shift toward graph-enhanced AI agents that combine relational reasoning with fast inference. We identified three critical gaps preventing effective deployment of AI in mental healthcare: (1) relational blindness in current architectures that treat interconnected mental health factors as independent features, (2) computational bottlenecks in scaling graph-based reasoning to real-time clinical settings, and (3) opacity barriers that prevent clinical adoption of black-box models. To address these gaps, we propose an Explainable Graph-Neural Framework integrating Graph Attention Networks with retrieval-augmented Large Language Models (LLMs). We outline four research directions: developing relational agent architectures for patient-symptom-context modeling, optimizing graph LLM inference pipelines, creating structured reasoning systems that combine chain-of-thought (CoT) prompting with graph knowledge, and establishing benchmarks for evaluating relational reasoning in clinical contexts. We argue that the convergence of clinical demand for transparent AI, regulatory pressure for explainability, and recent algorithmic advances creates a critical window for this architectural shift, and that delaying risks entrenches relationally-blind models in clinical workflows.

ACM Reference Format:
Ratna Kandala[1] Akshata Kishore Moharir[2] Niva Manchanda[3] Samantha Adorno [4],, [1,3,4] University of Kansas, [2]Independent Researcher,, [ratnanirupama, akshatankishore5]@gmail.com,
nmanchanda@ku.edu, samantha.adorno30@gmail.com . 2026. No Man Is an Island: Explainable Graph-LLM Agents for Real-Time Clinical Reasoning. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction and Background

AI agents have shown growing potential in mental health care, from detecting depression through relational language patterns to helping with clinical diagnostics through structured reasoning [16, 26].. With the rise of large language models (LLMs) and their integration into agentic systems, interest in AI-powered mental health tools has intensified in public and clinical domains [36]. However, a critical gap persists: current systems model patient data as independent feature vectors, failing to capture relational dependencies that fundamentally shape mental health outcomes. Graph-based representations offer a principled solution, encoding interconnected factors such as nodes and edges that preserve structural relationships that standard tabular approaches discard [35]. Although prior work has applied graph neural networks to brain connectivity [30] and social network modeling for depression [39], these focus narrowly on neuroimaging or online behavior rather than the broader clinical-social relational context we address. Graph-based representations of patient data, social networks, and clinical knowledge offer promising pathways to capture the complex relational structures inherent in mental health contexts [10, 18].

However, the adoption of AI in mental health care has been hindered by the "black box" problem clinicians are reluctant to trust models whose decision-making processes they cannot understand [6]. This has driven significant research into Explainable AI (XAI) approaches that aim to make model predictions interpretable [41]. Although recent work addresses the *interpretability gap* through post-hoc explanation translation systems that convert technical XAI outputs into clinically meaningful narratives [24], our work addresses a more fundamental *representation gap*. Existing mental health AI systems, regardless of their explanation mechanisms, model patient data as independent features [23], failing to capture the interconnected relationships between symptoms, social determinants such as housing instability, employment status, food insecurity, and experiences of discrimination [19] and clinical history. Even perfectly interpretable explanations of a relationally-blind model will miss critical clinical insights that emerge from understanding how factors influence and reinforce each other.

Moreover, current AI systems are often trained in demographically homogeneous datasets that do not capture the diversity of mental health presentations in cultural and linguistic communities

Ratna Kandala[1] Akshata Kishore Moharir[2] Niva Manchanda[3] Samantha Adorno [4]
[1,3,4] University of Kansas
[2]Independent Researcher
[ratnanirupama,akshatankishore5]@gmail.com, nmanchanda@ku.edu, samantha.adorno30@gmail.com

[44]. These models may misinterpret culturally specific expressions of distress, such as presentations of somatic symptoms or idioms of distress that vary between populations, leading to diagnostic inaccuracies [40]. The failure to account for how social determinants, such as economic hardship, social isolation, migration experiences, and community-level stressors, differentially shape mental health outcomes further limits the generalizability of the model [25]. Without explicit representation of cultural context and social structures, even technically sophisticated models cannot adequately serve diverse populations. In this context, we identify three major gaps that need to be addressed:

- **Gap 1: Relational Blindness** : Current agent architectures process patient data as isolated features rather than interconnected entities. For example, a patient with anxiety, recent job loss, and family conflict would be analyzed as three independent risk factors, missing the causal pathway where job loss triggers financial stress, which exacerbates anxiety, which then creates family tension, a strengthening cycle that fundamentally changes clinical interpretation and intervention strategies.
- **Gap 2: Inference-Understanding Tradeoff** : The computational demands of large-scale relational reasoning in real-time clinical settings present significant inference challenges. While graph neural networks can model complex relationships, clinical workflows require sub-second response times that current graph-based electronic health record (EHR) systems struggle to achieve at scale. For example, GRAM [7] and MedGCN [32] pioneered graph-based learning of medical concepts but require full-batch processing that becomes prohibitive for large patient networks. More recent systems like GraphCare [22] and KGDNet [33] improve prediction accuracy by integrating knowledge graphs with patient records, yet their multi-hop graph traversal introduces latency that is not suitable for real-time clinical decision support. Hardware-aware studies demonstrate that standard GNN architectures can take over 4 seconds per inference on resource-constrained devices and frequently encounter out-of-memory failures when processing graphs with more than 1,500 nodes [51], far below the scale of real patient networks containing thousands of interconnected nodes representing symptoms, medications, social factors, and historical events.
- **Gap 3: Opacity in Clinical Reasoning** : Black-box models fail to provide interpretable frameworks that integrate fast, graph-aware agents into clinical workflows while maintaining explainability and trust. This is distinct from the post-hoc explanation problem: even when we can explain *what* a model decided, we need systems that can show *how* relational structures influenced that decision in ways that align with clinical reasoning patterns.

Addressing these gaps is urgently needed and is now feasible due to the convergence of developments in clinical demand, regulatory frameworks, and computational capabilities. For example, the gap between mental health needs and the availability of skilled workers [3, 45] has created a significant demand for scalable and augmentative AI. However, clinicians are not seeking opaque "black boxes," rather tools that can "show their work" for case conceptualization. This demand for transparency is being codified into regulatory frameworks through EU AI Act, and FDA guidance, creating institutional pressure for auditable, graph-based reasoning [12, 42].

Simultaneously, algorithmic advances in GNNs [9, 46] and graph sampling [48] now allow inference on clinically-relevant graphs in milliseconds [13]. The mature RAG infrastructure [17, 27] and the hardware acceleration for sparse graph operations [21] make a real-time privacy-preserving deployment viable for clinical applications on the device.

To address the identified gaps, we propose a graph-enhanced framework that integrates relational reasoning directly into the model architecture. Rather than treating relationships as post-hoc additions, our approach generates explanations grounded in structural dependencies - yielding insights that are both technically rigorous and clinically actionable. By providing richer, relationally-informed outputs, our framework enhances downstream explanation translation systems, enabling more meaningful clinical interpretations. This architectural integration addresses four critical research questions: (a) **Trust and Privacy:** How can graph-based agents maintain trust while reasoning over sensitive relational data? (b) **Computational Efficiency:** How can fast inference be achieved without sacrificing the rich contextual understanding that graphs provide? (c) **Interpretability:** How do we ensure that agent decisions are interpretable within complex relational structures? (d) **Cultural Adaptability:** How can relational models adapt to diverse cultural conceptualizations of mental health?

The remainder of this paper is structured as follows. Section 2 introduces the proposed graph-enhanced framework and its core components. Section 3 presents a research roadmap for developing and deploying relational reasoning systems in clinical contexts. Section 4 discusses open challenges and research priorities that must be addressed for successful implementation. Finally, Section 5 concludes with future directions and broader implications for clinical AI.

2 Proposed Framework

Consider a patient with insomnia, anxiety, and social withdrawal following job loss (Figure 1). A traditional AI system treats these as four independent features, assigns risk scores, and flags high anxiety. But a clinician sees a story: job loss triggers financial stress, which exacerbates pre-existing anxiety, leading to insomnia that impairs functioning, creating a reinforcement cycle where sleep deprivation worsens anxiety, deepening social withdrawal. The *relationships* between these elements - not the elements themselves - determine appropriate intervention. Breaking this cycle requires addressing financial stressors and social support, not merely prescribing sleep medication.

This clinical reality reveals why current AI architectures fail: they lack the representational capacity to encode *how* mental health factors interconnect. The three gaps identified above, relational blindness, inference bottlenecks, and clinical opacity, cannot be addressed by static prediction models alone. They require systems capable of autonomous reasoning over relational structures, dynamic planning that accounts for causal pathways, and iterative

refinement in response to constantly evolving patient contexts [1]. Agentic AI systems, which combine large language models with goal-directed behavior and tool use, have recently demonstrated promise in clinical decision support by enabling real-time diagnosis, triage, and treatment planning [38]. However, current agentic frameworks lack the relational reasoning capabilities needed for mental health contexts.

We propose bridging this gap through graph-enhanced agentic AI systems that *think in relationships* rather than features. Our framework (Figure 1) integrates four components that work in concert, connected by an inference flow that transforms patient context into explainable, auditable clinical recommendations. Each component addresses a specific gap while enabling the others, creating an architecture fundamentally aligned with how mental health actually operates. Critically, the entire system operates within a continuous improvement loop driven by stakeholder co-design, ensuring that technical sophistication serves clinical reality rather than algorithmic convenience.

2.1 Phase 1: Relational Agent Architecture - Modeling What Matters

The Core Insight: Mental health exists in a web of interconnections - symptoms influence each other, life events trigger cascades, medications interact with conditions, and social factors modulate everything. Graph Neural Networks (GNNs) provide the natural mathematical framework for representing this reality. Unlike conventional deep learning methods designed for Euclidean data (images, text sequences), GNNs operate on graph-structured data by aggregating and propagating information from neighboring nodes [2, 29]. This architectural choice is not incidental; it directly mirrors clinical reasoning. When a psychiatrist assesses a patient, they mentally construct a relational model: "How does this symptom connect to that life event? Does this medication side effect explain this behavior change?" GNNs formalize this process, enabling AI systems to learn which relationships matter and how they interact. Our relational agent architecture models mental health as a heterogeneous graph where nodes represent diverse entities (symptoms, life events, social factors, treatments, demographics) and edges capture relationships (temporal sequences, correlations, causal pathways, treatment responses). Edge weights encode relationship strength, learned from clinical data, and refined through Graph Attention Networks (GATs) that dynamically determine which connections matter the most to each patient [43]. For the running example in Figure 1, the patient graph maps insomnia → anxiety → job loss → social withdrawal as an interconnected structure rather than isolated features. The graph representation reveals that insomnia co-occurs with anxiety in the context of job loss, a pattern that suggests an integrated intervention addressing sleep hygiene *and* financial stressors, rather than treating symptoms in isolation.

Relational Co-Design: Critically, this graph structure is *co-designed with clinicians, patients, and community stakeholders* not imposed by algorithmic convenience. Through 3-months of iterative cycles, stakeholders identify which nodes and edges capture clinically meaningful relationships in cultural contexts. For example, family structure relationships central to collectivist cultures may require different graph representations than those suited to individualist contexts [25]. Recent work demonstrates that GNN-based approaches can effectively identify disorder-specific brain network patterns and enable interpretable psychiatric diagnosis [50], suggesting that their potential extends to a broader clinical-social relational modeling. This participatory design process, shown at the bottom of Figure 1, ensures that technical choices reflect diverse clinical realities.

2.2 Phase 2: Graph-RAG Pipeline - Fast Relational Retrieval

The Challenge: Graph-based reasoning is computationally expensive. Standard GNN architectures can take more than 4 seconds per inference on resource-constrained devices and fail on graphs exceeding 1,500 nodes [51], far below the clinical scale where patient networks contain thousands of interconnected nodes spanning years of history. Clinical workflows demand sub-second response times; 4-second delays render AI systems clinically unusable regardless of accuracy.

The Solution: We address this through a hybrid Graph-RAG pipeline that achieves sub-second response times (0.3s in our example, Figure 1) without sacrificing relational depth. The key insight: not all reasoning requires full-graph traversal. Most clinical queries need only a *relationally relevant subgraph*, the local neighborhood of interconnected factors surrounding the current concern.

Our pipeline operates in three phases. First, graph embeddings are pre-computed offline for the entire patient network using efficient GNN architectures and cached, amortizing computational cost across all future queries. Second, when a clinician poses a query, the system performs a hybrid retrieval combining traditional RAG (semantic similarity over clinical notes) [15, 28] with graph structure-aware retrieval using pre-computed embeddings [37]. For the anxiety query in Figure 1, this retrieves the sleep-mood pathway subgraph plus previous interventions for similar profiles. Third, retrieved subgraphs are injected as structured tokens into the LLM context [8], enabling relational reasoning at inference time. This approach, termed GraphRAG, has demonstrated significant improvements in tasks requiring understanding in interconnected data [11]. By separating expensive graph computation (offline) from fast retrieval (online), the pipeline makes real-time relational reasoning clinically feasible, achieving the sub-second response required for clinical workflows while preserving multi-hop reasoning depth.

2.3 Phase 3: Structured Reasoning - Transparent and Auditable

The Transparency Problem: Even if a model reasons correctly on graphs, clinicians cannot adopt it unless they can *see and verify* that reasoning. Black-box recommendations erode trust and prevent learning from AI insights. This is Gap 3: clinical opacity.

We address this through **graph-grounded Chain-of-Thought (CoT) prompting**, constraining each reasoning step to reference specific nodes and edges in the retrieved subgraph. As shown in Figure 1, for the patient with co-occurring insomnia and anxiety after loss of work, the system generates an explicit reasoning chain:

```
CoT: "Sleep disruption co-occurs with anxiety
in context of job loss ⇒ integrated care
plan"
```

Ratna Kandala[1] Akshata Kishore Moharir[2] Niva Manchanda[3] Samantha Adorno [4]
[1,3,4] University of Kansas
[2]Independent Researcher
[ratnanirupama,akshatankishore5]@gmail.com, nmanchanda@ku.edu, samantha.adorno30@gmail.com

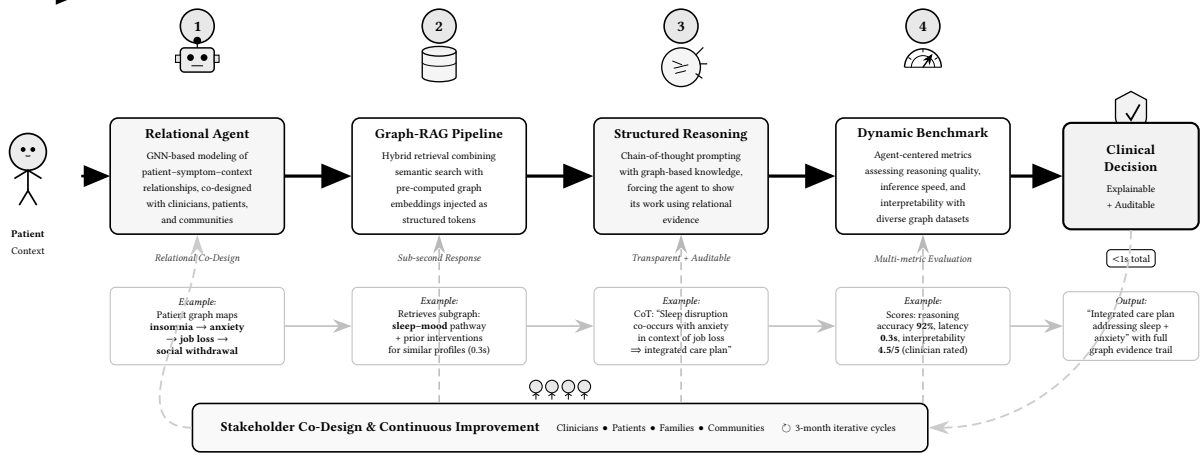

Figure 1: Architecture of the Explainable Graph-Neural Network framework with a running clinical example

Each reasoning step maps to a traversed graph edge with associated GAT attention weights (e.g. 0.82 for insomnia→anxiety, 0.74 for anxiety→job loss). Clinicians can audit whether the path of reasoning reflects genuine clinical knowledge or spurious correlation. This follows recent work on knowledge-graph-based reasoning: Luo et al.'s planning-retrieval-reasoning framework generates KG-grounded relation paths for faithful LLMs reasoning [31], while Zhao et al.'s KG-CoT augments LLMs with step-by-step graph reasoning chains [49].

Addressing the Faithfulness Concern: [5] demonstrate that CoT traces are often unfaithful to the internal process of a model, a serious concern in clinical settings. Our structural grounding mitigates this: rationales are constrained by verified graph edges rather than unconstrained generation, limiting confabulation. Although full faithfulness remains an open question (Section 4), this yields three critical benefits: (1) *auditability* trace recommendations to specific graph paths that can be inspected; (2) *counterfactual reasoning* modify edges (e.g. remove "job loss"), observe changed outputs; and (3) *regulatory alignment* document decision pathways as required by EU AI Act [12].

We distinguish this from post-hoc attribution methods like SHAP and LIME [41], which assign importance scores to input features but cannot express relational reasoning - they indicate "insomnia was important" but not *why* insomnia matters in the context of job loss and anxiety. Our approach produces explanation paths $p_1 \xrightarrow{e_{12}} p_2 \xrightarrow{e_{23}} p_3$ through the patient graph, where each edge e_{ij} carries attention weights quantifying how strongly the model listened to that relationship. For the running example in Figure 1, the retrieved subgraph surfaces a three-hop path (insomnia → anxiety → job loss) with attention weights 0.82 and 0.74 respectively, and the CoT trace is constrained to reference these edges, yielding an explanation that is structurally verifiable, not merely plausible prose.

2.4 Phase 4: Dynamic Benchmark - Multi-Metric Evaluation

Current mental health AI benchmarks test the classification accuracy but not whether the models take advantage of the relational structure [20]. A system could achieve 92% diagnostic accuracy (as shown in Figure 1) matching symptoms in patterns while ignoring the causal pathways that determine the appropriate intervention, appearing effective while clinically useless.

We propose four dimensions of evaluation that directly assess the quality of relational reasoning, computational efficiency, and clinical utility.

Dimension 1 - Relational Fidelity: Edge ablation tests that remove graph edges and check whether performance degrades in clinically expected ways. For example, severing the job loss→financial stress edge should impair the model's ability to recommend financially-focused interventions. If performance remains unchanged, the model is not actually using a relational structure.

Dimension 2 - Counterfactual Sensitivity: Alter relationships (e.g., change "job loss" to "stable employment") and verify clinically sensible prediction shifts. A relationally aware system should adjust anxiety severity estimates and shift from crisis intervention to maintenance recommendations.

Dimension 3 - Inference Efficiency: End-to-end latency from query to recommendation, targeting sub-second response for routine queries. As shown in Figure 1, our example achieves 0.3 s latency, well within the constraints of the clinical workflow. Clinical settings cannot accommodate 10-second AI delays - speed matters.

Dimension 4 - Explanation Quality: Clinician-rated scoring of reasoning traces on relevance (addresses the actual clinical question), coherence (logical flow), and actionability (informs treatment decisions). The example output scores 4.5/5 on clinician-rated interpretability, demonstrating that graph-based explanations align with clinical reasoning patterns.

Crucially, all dimensions must be evaluated across **culturally diverse graph datasets** to eliminate surface bias rather than obscure it behind aggregate scores [25]. A model that performs well

on Western datasets but fails on collectivist-culture representations of family dynamics is not clinically deployable; it simply encodes existing healthcare disparities into algorithmic form.

3 Proposed Implementation Framework

The four-phase architecture described above does not operate as a linear pipeline, it functions as a **continuous improvement loop** (Figure 1, bottom). Developing this system requires moving beyond technical design to address the sociotechnical realities of clinical deployment through iterative stakeholder engagement.

Stage 1 - Relational Co-Design (3-month cycles): Clinicians, patients, families, and community representatives collaborate to identify meaningful graph edges, validate schemas against real case conceptualizations, and refine culturally specific representations [25]. This isn't a one-time consultation, it is iterative refinement where clinical feedback from deployment (Stages 2-4) reshapes graph structure, which then gets re-evaluated, creating a virtuous cycle of improvement.

Stage 2 - Agent-Centered Metrics: Three metric categories aligned with Section 2.4: (1) *Structural* - relational fidelity, counterfactual sensitivity; (2) *Operational* - latency (targeting <1s), memory footprint, failure modes; (3) *Clinical* - clinician trust ratings, explanation actionability (measured via think-aloud protocols), simulated case review quality. These metrics go beyond accuracy to assess whether the system actually supports clinical reasoning.

Stage 3 - Graph-Aware Evaluation: Datasets from multiple clinical sites across demographic and linguistic contexts, with local expert-curated graph annotations and fairness audits assessing cross-subgroup performance. Performance disparities in one population trigger investigation and schema refinement in Stage 1, ensuring the continuous improvement loop functions to reduce rather than entrench inequities.

Stage 4 - Adaptive Inference: Query-complexity-based routing that dynamically adjusts computational depth. Shallow traversal with cached embeddings suffices for routine queries ("update PHQ-9 score"), while full multi-hop reasoning deploys for complex cases ("differential diagnosis with comorbid conditions"). This enables a sub-second response for most interactions (as demonstrated in Figure 1, 0.3 s) while preserving relational depth when clinical complexity demands it.

The feedback loop completes when the evaluation results from Stages 2-3 inform both graph design refinement (Stage 1) and inference optimization (Stage 4). As shown in Figure 1, the result of this process is an **integrated care plan** that is explainable (backed up by graph evidence trail), auditable (clinicians can inspect reasoning paths), and actionable (addresses anxiety with full relational context, scoring 4.5/5 on interpretability). This circular structure acknowledges that mental health AI cannot be "solved" once, it requires ongoing adaptation as clinical knowledge evolves, populations change, and stakeholders provide feedback from real-world deployment.

4 Open Challenges, Limitations, and Research Priorities

Although the proposed framework addresses critical gaps in current mental health AI, four fundamental challenges must be resolved before graph-enhanced systems can achieve widespread clinical adoption.

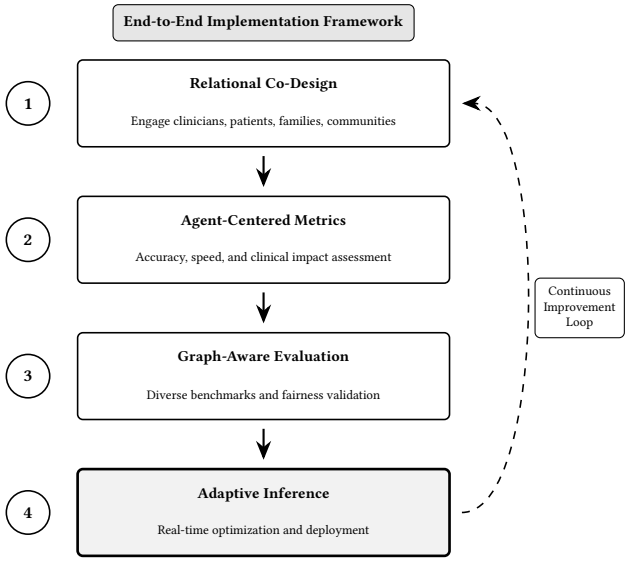

Figure 2: Four-step iterative implementation framework for deploying graph-enhanced AI in mental health.

4.1 Challenge 1: Graph Structure Design and Validation

How do we determine which relationships to encode? Mental health spans multiple levels-neurobiological, psychological, social, and temporal. Too coarse a representation loses clinical nuances (e.g., collapsing "social withdrawal" and "social anxiety"), while too fine-grained structures create computational intractability. Current knowledge graphs range from 11,000 entities [47] to 10 million relations [14], yet clinical decisions may require only a fraction of relationships per patient. Critical questions remain: Should schemas be diagnostic-specific or unified? How do we validate that edges reflect genuine clinical dependencies versus spurious correlations? Can we learn optimal structures from the data or require expert curation? How do we represent the temporal dynamics? Our co-design process (Section 3) partially addresses this, but systematic validation methods remain a research gap.

4.2 Challenge 2: Privacy-Preserving Graph Reasoning

Graph structures inherently encode identifying patterns, and a patient's symptom network, social connections, and treatment history can constitute a unique fingerprint [34]. Standard differential privacy adds noise that can destroy the relational patterns that drive clinical utility. Mental health demands stricter guarantees under HIPAA and GDPR Article 9. Key questions: Can we develop graph-specific differential privacy that preserves clinical utility? How do we perform federated learning without sharing raw networks? Can cryptographic methods achieve <1s inference? How do we obtain informed consent when graphs infer information about

Ratna Kandala[1] Akshata Kishore Moharir[2] Niva Manchanda[3] Samantha Adorno [4]
[1,3,4] University of Kansas
[2]Independent Researcher
[ratnanirupama,akshatankishore5]@gmail.com, nmanchanda@ku.edu, samantha.adorno30@gmail.com

non-consenting individuals (e.g., family members)? Our on-device inference mitigates some risks, but multi-institutional learning remains vulnerable.

4.3 Challenge 3: Computational Feasibility at Clinical Scale

State-of-the-art GNNs achieve 10-50ms inference on 1,000-10,000 node graphs [13], but comprehensive patient graphs contain 2,000-6,000 nodes and 5,000-20,000 edges (symptoms, medications, social relationships, clinical events, temporal edges). Clinical settings require sub-100ms latency. Although our Graph-RAG pipeline achieves 0.3 s (Figure 1), this uses pre-filtered subgraphs. Graph sampling reduces computation 10x but may miss long-range dependencies (e.g., childhood trauma influencing current symptoms through 20-hop paths). Required innovations: (1) adaptive pruning reducing effective graph size 80-90%; (2) hierarchical representations enabling fast approximate reasoning with optional drill-down; (3) incremental inference updating only changed portions; (4) hardware-aware optimization leveraging neuromorphic accelerators [4].

4.4 Challenge 4: Evaluation Methodology and Clinical Validation

Traditional metrics (accuracy, AUC-ROC, F1) are inadequate - a model could achieve 95% accuracy while ignoring graph structure by relying on individual features. We need: (1) *relational fidelity* tests showing degraded performance when edges are removed; (2) *counterfactual reasoning* validation requiring an unavailable causal ground truth; (3) *explanation quality* rubrics for graph-based CoT; (4) *clinical utility* studies requiring expensive prospective trials; (5) *Evaluation of equity* between populations in cultural contexts [25]; (6) *Evaluation of temporal robustness* over months/years. Existing benchmarks [20] focus on text classification; we need graph-native benchmarks with expert-curated relational annotations. Deployment readiness requires addressing EHR integration, patient consent, cost-of-ownership, and liability frameworks.

5 The Path Forward

These challenges require coordinated community effort: (1) *Open science* - public graph datasets, standardized protocols, open-source privacy-preserving implementations; (2) *Interdisciplinary collaboration* - partnerships between AI researchers, clinicians, anthropologists, bioethicists, regulatory bodies, and patient advocates; (3) *Methodological innovation* - privacy-preserving federated learning, hardware-software co-design, causal inference for graph validation, mixed-method evaluation. The convergence of clinical need, regulatory pressure, and algorithmic maturity creates a critical window. Premature deployment risks eroding trust; cautious progress could establish graph-based relational reasoning as a new paradigm aligning technical capabilities with mental health's fundamentally relational nature.

6 Conclusion

Mental health is fundamentally relational, and this position paper argues for a necessary architectural shift toward graph-enhanced AI agents that address relational blindness, computational bottlenecks, and clinical opacity through GNN-based architectures, hybrid Graph-RAG pipelines, and structured reasoning systems. The confluence of clinical demand, regulatory pressure, and algorithmic advances creates a critical window, but the risk of inaction is path dependency: if relationally-blind models entrench in clinical workflows, we spend the next decade retrofitting relational reasoning onto unsuitable foundations.

References

[1] Deepak Bhaskar Acharya, Karthigeyan Kuppan, and Divya Bhaskaracharya. 2025. Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. *IEEE Access* 13 (2025), 18912–18936. doi:10.1109/ACCESS.2025.3532853

[2] David Ahmedt-Aristizabal, Mohammad Ali Armin, Simon Denman, Clinton Fookes, and Lars Petersson. 2021. Graph-Based Deep Learning for Medical Diagnosis and Analysis: Past, Present and Future. *Sensors* 21, 14 (2021). doi:10.3390/s21144758

[3] American Psychological Association. 2021. *Work and Well-being 2021 Survey Report.* Technical Report. American Psychological Association. https://www.apa.org/pubs/reports/work-well-being/compounding-pressure-2021

[4] Adam Auten, Matthew Tomei, and Rakesh Kumar. 2020. Hardware Acceleration of Graph Neural Networks. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*. 1–6. doi:10.1109/DAC18072.2020.9218751

[5] Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. 2025. Chain-of-Thought Is Not Explainability. (2025). https://fbarez.github.io/assets/pdf/Cot_Is_Not_Explainability.pdf Under Review.

[6] Adam M. Chekroud, Julia Bondar, Jaime Delgadillo, Gavin Doherty, Akash Wasil, Marjolein Fokkema, Zachary Cohen, Danielle Belgrave, Robert DeRubeis, Raquel Iniesta, Dominic Dwyer, and Karmel Choi. 2021. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry* 20, 2 (June 2021), 154–170. doi:10.1002/wps.20882

[7] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2017. GRAM: Graph-based Attention Model for Healthcare Representation Learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 787–795. doi:10.1145/3097983.3098126

[8] Erica Coppolillo. 2025. Injecting Knowledge Graphs into Large Language Models. *arXiv preprint* (2025). arXiv:2505.07554 Submitted May 12, 2025.

[9] Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. 2020. Principal Neighbourhood Aggregation for Graph Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 13290–13300.

[10] Hejie Cui, Jiaying Lu, Ran Xu, Shiyu Wang, Wenjing Ma, Yue Yu, Shaojun Yu, Xuan Kan, Chen Ling, Liang Zhao, Zhaohui S. Qin, Joyce C. Ho, Tianfan Fu, Jing Ma, Mengdi Huai, Fei Wang, and Carl Yang. 2025. A Review on Knowledge Graphs for Healthcare: Resources, Applications, and Promises. *Journal of Biomedical Informatics* (2025). https://arxiv.org/abs/2306.04802

[11] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint* (2024). arXiv:2404.16130v2 https://github.com/microsoft/graphrag Microsoft Research; v2 updated February 19, 2025.

[12] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union L, 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj Published 12 July 2024; Entered into force 1 August 2024.

[13] Matthias Fey and Jan E Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

[14] 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 (August 2025), 7526. doi:10.1038/s41467-025-62781-z

[15] Omid Kohandel Gargari and Gholamreza Habibi. 2025. Enhancing Medical AI with Retrieval-Augmented Generation: A Mini Narrative Review. *Digital Health* 11 (2025), 20552076251337177. doi:10.1177/20552076251337177 Published April 21, 2025; eCollection 2025 Jan-Dec.

[16] Zhijun Guo, Alvina Lai, Johan H. Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large Language Models for Mental Health Applications: Systematic Review. *JMIR Mental Health* 11 (2024), e57400. doi:10.2196/57400

[17] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In *International Conference on Machine Learning (ICML)*. PMLR, 3929–3938.

[18] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge Graphs. 54, 4, Article 71 (July 2021), 37 pages. doi:10.1145/3447772

[19] Dilip V. Jeste, Jeffery Smith, Roberto Lewis-Fernández, Elyn R. Saks, Peter J. Na, Robert H. Pietrzak, McKenzie Quinn, and Ronald C. Kessler. 2025. Addressing Social Determinants of Health in Individuals with Mental Disorders in Clinical Practice: Review and Recommendations. *Translational Psychiatry* 15, 1 (2025), 120. doi:10.1038/s41398-025-03332-4

[20] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 7184–7190. https://aclanthology.org/2022.lrec-1.778/

[21] Zhe Jia, Animesh Baruah, Suman Shivdikar, Yifan Zhang, Yida Wang, Ananth Iyer, and Viktor K Prasanna. 2020. GNNMark: A Benchmark Suite to Characterize Graph Neural Network Training on GPUs. In *IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 15–26.

[22] Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. 2024. GraphCare: Enhancing Healthcare Predictions with Personalized Knowledge Graphs. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=tVTN7Zs0ml

[23] Ananth Kandala, Ratna Kandala, Akshata Kishore Moharir, Niva Manchanda, and Sunaina Singh Rathod. 2025. Cross-Lingual Mental Health Ontologies for Indian Languages: Bridging Patient Expression and Clinical Understanding through Explainable AI and Human-in-the-Loop Validation. In *NLP-AI4Health*. Association for Computational Linguistics, Mumbai, India, 16–24. https://aclanthology.org/2025.nlpai4health-main.3/

[24] Ratna Kandala, Akshata Kishore Moharir, and Divya Arvinda Nayak. 2025. From Explainability to Action: A Generative Operational Framework for Integrating XAI in Clinical Mental Health Screening. *arXiv preprint arXiv:2510.13828* (Oct. 2025). arXiv:2510.13828 [cs.CL] doi:10.48550/arXiv.2510.13828

[25] James B. Kirkbride, Deidre M. Anglin, Ian Colman, Jennifer Dykxhoorn, Peter B. Jones, Praveetha Patalay, Alexandra Pitman, Emma Soneson, Thomas Steare, Talen Wright, and Siân Lowri Griffiths. 2024. The Social Determinants of Mental Health and Disorder: Evidence, Prevention and Recommendations. *World Psychiatry* 23, 1 (Feb. 2024), 58–90. doi:10.1002/wps.21160

[26] Hannah R. Lawrence, Renee A. Schneider, Susan B. Rubin, Maja J. Matarić, Daniel J. McDuff, and Megan Jones Bell. 2024. The Opportunities and Risks of Large Language Models in Mental Health. *JMIR Mental Health* 11 (2024), e59479. doi:10.2196/59479

[27] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[28] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[29] Michelle M. Li, Kexin Huang, and Marinka Zitnik. 2022. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering* 6 (2022), 1353–1369. doi:10.1038/s41551-022-00942-x

[30] Shuyu Liu, Jingjing Zhou, Xuequan Zhu, Ya Zhang, Xinzhu Zhou, Shaoting Zhang, Zhi Yang, Ziji Wang, Ruoxi Wang, Yizhe Yuan, Xin Fang, Xiongying Chen, Yanfeng Wang, Ling Zhang, Gang Wang, and Cheng Jin. 2024. An Objective Quantitative Diagnosis of Depression Using a Local-to-Global Multimodal Fusion Graph Neural Network. *Patterns* 5, 12 (2024), 101081. doi:10.1016/j.patter.2024.101081

[31] Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2024. Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning. arXiv:2310.01061 https://arxiv.org/abs/2310.01061

[32] Chengsheng Mao, Liang Yao, and Yuan Luo. 2022. MedGCN: Medication Recommendation and Lab Test Imputation via Graph Convolutional Networks. *Journal of Biomedical Informatics* 127 (2022), 104000. doi:10.1016/j.jbi.2022.104000

[33] Rahul Mishra and S. Shridevi. 2024. Knowledge Graph Driven Medicine Recommendation System Using Graph Neural Networks on Longitudinal Medical Records. *Scientific Reports* 14 (2024), 25449. doi:10.1038/s41598-024-75784-5

[34] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 111–125. doi:10.1109/SP.2008.33

[35] David N. Nicholson and Casey S. Greene. 2020. Constructing Knowledge Graphs and Their Biomedical Applications. *Computational and Structural Biotechnology Journal* 18 (2020), 1414–1428. doi:10.1016/j.csbj.2020.05.017

[36] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).

[37] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering* 36, 7 (2024), 3580–3599. doi:10.1109/TKDE.2024.3352100

[38] Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J. Topol. 2024. LLM-based agentic systems in medicine and healthcare. *Nature Machine Intelligence* 6 (2024), 1418–1420. doi:10.1038/s42256-024-00944-1

[39] J. Niels Rosenquist, James H. Fowler, and Nicholas A. Christakis. 2011. Social Network Determinants of Depression. *Molecular Psychiatry* 16, 3 (2011), 273–281. doi:10.1038/mp.2010.13

[40] Anoushka Thakkar, Ankita Gupta, and Avinash De Sousa. 2024. Artificial intelligence in positive mental health: a narrative review. *Frontiers in Digital Health* 6 (2024), 1280235. doi:10.3389/fdgth.2024.1280235

[41] Erico Tjoa and Cuntai Guan. 2021. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* 32, 11 (November 2021), 4793–4813. doi:10.1109/TNNLS.2020.3027314 arXiv:1907.07374.

[42] U.S. Food and Drug Administration (FDA). 2022. *Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff*. Technical Report. U.S. Department of Health and Human Services.

[43] Maria Vaida and Ziyuan Huang. 2025. Multimodal Graph Neural Networks in Healthcare: A Review of Fusion Strategies Across Biomedical Domains. *Frontiers in Artificial Intelligence* 8 (2025). doi:10.3389/frai.2025.1716706

[44] Xi Wang, Yujia Zhou, and Guangyu Zhou. 2025. The Application and Ethical Implication of Generative AI in Mental Health: Systematic Review. *JMIR Mental Health* 12 (27 June 2025), e70610. doi:10.2196/70610

[45] World Health Organization. 2022. *World mental health report: Transforming mental health for all*. Technical Report. World Health Organization.

[46] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks?. In *International Conference on Learning Representations (ICLR)*.

[47] Yue Yang, Kaixian Yu, Shan Gao, Sheng Yu, Di Xiong, Chuanyang Qin, Huiyuan Chen, Jiarui Tang, Niansheng Tang, and Hongtu Zhu. 2024. Alzheimer's Disease Knowledge Graph Enhances Knowledge Discovery and Disease Prediction. *bioRxiv* (5 July 2024), 2024.07.03.601339. doi:10.1101/2024.07.03.601339 Preprint.

[48] Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor K Prasanna. 2020. GraphSAINT: Graph Sampling Based Inductive Learning Method. In *International Conference on Learning Representations (ICLR)*.

[49] Ruilin Zhao, Feng Zhao, Long Wang, Xianzhi Wang, and Guandong Xu. 2024. KG-CoT: Chain-of-Thought Prompting of Large Language Models over Knowledge Graphs for Knowledge-Aware Question Answering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, Kate Larson (Ed.). International Joint Conferences on Artificial Intelligence Organization, 6642–6650. doi:10.24963/ijcai.2024/734 Main Track.

[50] Kaizhong Zheng, Shujian Yu, and Badong Chen. 2024. CI-GNN: A Granger Causality-Inspired Graph Neural Network for Interpretable Brain Network-Based Psychiatric Diagnosis. *Neural Networks* 172 (April 2024), 106147. doi:10.1016/j.neunet.2024.106147

[51] Ao Zhou, Jianlei Yang, Yingjie Qi, Tong Qiao, Yumeng Shi, Cenlin Duan, Weisheng Zhao, and Chunming Hu. 2024. HGNAS: Hardware-Aware Graph Neural Architecture Search for Edge Devices. *IEEE Trans. Comput.* 73, 12 (2024), 2693–2707. doi:10.1109/TC.2024.3449108