EPIGRAPH: BUILDING GENERALISTS FOR EVIDENCE-INTENSIVE EPILEPSY REASONING IN THE WILD

Anonymous authors
Paper under double-blind review

ABSTRACT

Epilepsy diagnosis and treatment require evidence-intensive reasoning across heterogeneous clinical knowledge, including biosignal patterns, genetic mechanisms, pharmacogenomics, treatment strategies, and patient outcomes. In this work, we present EPIGRAPH, a large-scale epilepsy knowledge graph and benchmark for evaluating knowledge-augmented clinical reasoning. EPIGRAPH integrates 48,166 peer-reviewed papers and seven clinical resources into a heterogeneous graph containing 24,324 entities and 32,009 evidence-grounded triplets across five clinical layers. Built upon this graph, EPIBENCH defines five clinically motivated tasks spanning clinical decision-making, EEG report generation, pharmacogenomic precision medicine, treatment recommendation, and deep research planning. We evaluate six LLMs under both standard and Graph-RAG settings. Results show that integrating EPIGRAPH consistently improves performance across all tasks, with the largest gains observed in pharmacogenomic reasoning (+30–41%). Our findings demonstrate that structured epilepsy knowledge substantially enhances evidence-grounded clinical reasoning and provides a practical benchmark framework for evaluating knowledge-augmented LLMs in real-world neurological settings. See implementation details at: https://anonymous.4open.science/r/Epi-2E84.

1 INTRODUCTION

Epilepsy is a long-standing challenge in neuroscience and clinical medicine, affecting over 50 million individuals Kotoge et al. (2025). Although many disease mechanisms can lead to epilepsy, accurate diagnosis remains challenging Kwan et al. (2011). To answer diagnostic questions, clinicians must integrate multiple sources of evidence, including seizure semiology, electrophysiological signals, and patient history Chen et al. (2025b). This process requires extensive training to construct a comprehensive knowledge framework. Moreover, the etiology of epilepsy remains unknown in 50% of cases, and understanding disease progression requires identifying subtle symptoms, determining which causal genes are associated with the disease, and characterizing how patients respond to treatments Kuo et al. (2024). This requires integrating heterogeneous knowledge across biological, clinical, and pharmacological perspectives, as well as reasoning evidence from a large body of literature. *Precise curation and reasoning of such diverse evidence* is therefore crucial for both clinical practice and scientific discovery.

To support such evidence-intensive investigation, heterogeneous biomedical knowledge must be represented in a structured and integrative form. Knowledge graphs (KGs) provide a natural framework for this purpose, organizing entities and their relationships across multiple domains and enabling multi-hop reasoning Cui et al. (2025). Several biomedical KGs have been proposed across diverse domains such as precision medicine Chandak et al. (2023), drug repurposing Lu et al. (2025), oncology Cao et al. (2024), and disease–gene–drug modeling Bonner et al. (2022). Prior work includes disease-agnostic resources with limited epilepsy-specific depth Bodenreider (2004) and semantic-annotation resources that identify medical terms in text (e.g., symptoms, seizure types, and electrophysiological patterns) and convert them into a consistent vocabulary Sahoo et al. (2014); Sargsyan et al. (2023). Therefore, there remains a gap in developing a KG that comprehensively captures epilepsy-specific knowledge and enables interpretation of complex underlying disease mechanisms.

Recently, large language models (LLMs) have demonstrated strong capabilities for advancing evidence curation in scientific research, leveraging both internal knowledge corpora and the ability to interpret external scientific publications Singhal et al. (2023); Kung et al. (2023); Nori et al. (2023);

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

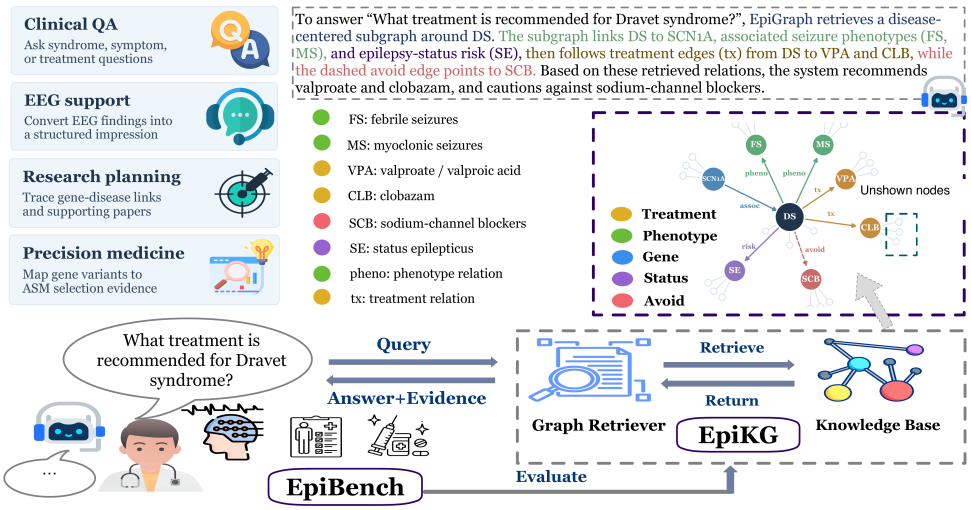


Figure 1: **Overview of the EPIGRAPH framework**, comprising two core components: EPIKG, an epilepsy knowledge graph, and EPIBENCH, a multi-task evaluation benchmark. Given a clinical query (e.g., "What treatment is recommended for Dravet syndrome?"), the Graph Retriever queries EPIKG to extract a subgraph linking the syndrome to associated phenotypes, genes, treatments, and contraindications. The retrieved reasoning paths are returned to the LLM to generate a grounded answer with supporting evidence, which is then evaluated by EPIBENCH across five clinical tasks.

Shao et al. (2025). Recent work further combines KGs with LLMs for clinical reasoning Gao et al. (2025); Rezaei et al. (2025); Wu et al. (2025); Li et al. (2026a). A growing body of benchmarks has been proposed to evaluate LLMs on complex biomedical question answering and multi-step reasoning Al-Breiki et al. (2025); Lucas et al. (2024); Ma et al. (2024); Arora et al. (2025); Zhu et al. (2025); Xiong et al. (2024). Existing benchmarks often focus on contrived tasks within narrow domains, offering limited coverage of epilepsy practice. More importantly, epilepsy research extends beyond isolated clinical question answering, requiring integration across biomarker discovery, disease mechanisms, treatment strategies, and broader clinical practice Li et al. (2026b). Therefore, a critical gap exists in datasets to evaluate whether LLMs can perform expert-level evidence curation and reasoning in realistic epilepsy settings.

Present work. These gaps motivate EPIGRAPH, a unified framework for evidence-intensive reasoning in epilepsy. Figure 1 presents its two connected components. ❶ EPIKG, a large-scale epilepsy KG constructed via a structured evidence-to-graph pipeline with three stages. First, we define the graph schema based on seven authoritative resources (e.g., NCBI-MeSH National Library of Medicine (2024)), organizing entities into five layers (genes, phenotypes, syndromes, treatments, and outcomes) and defining 1,370 cross-layer relation types (e.g., *caused by gene*, *treated with*). This ensures that the graph is grounded in clinically validated resources. We then collect large-scale evidence over 120,000 PubMed papers and applying a two-stage screening process, where an LLM-assisted classifier filters candidates and domain experts adjudicate borderline cases, yielding 48,166 papers for evidence extraction. We finally map the extracted evidence into the predefined graph schema, aligning relations to predefined types, and validating triplets. The resulting graph contains 24,324 entities and 32,009 triplets, including 14,576 cross-layer connections, enabling evidence-grounded, multi-hop reasoning in epilepsy. ❷ EPIBENCH, a multi-task benchmark for evaluating LLM reasoning and the effectiveness of EPIKG in epileptology. Our evaluation covers five tasks, including biomedical question answering, as well as a practical clinical task that supports clinicians in EEG interpretation and report generation. To construct the benchmark, we curate evidence from 1,700 research papers involving over 7,000 clinical cases, and transform them into structured evaluation formats. This results in 6,199 QA pairs, 151 pharmacogenomic multiple-choice questions, 472 treatment recommendation tasks, and 163 research planning cases. For the practical EEG-to-report generation task, LLMs take EEG inputs and incorporate external knowledge in EPIKG to interpret the data and generate report.

Our contributions are as follows:
❑ *A New Epilepsy Knowledge Graph.* We propose the first large-scale, multi-hop relational knowledge graph dedicated to epileptology, spanning phenotypes, genetic mechanisms, and treatments.

❑ *A Comprehensive Benchmark Dataset.* We propose a publicly released, modular benchmark spanning five clinical reasoning tasks to support plug-and-play evaluation of any LLM or retriever.
❑ *A Clinically Grounded Evaluation with Practical Impact.* We validate on decades of real-world Harvard clinical EEG notes, grounding evaluation in authentic clinical practice and demonstrating that graph-augmented LLMs can meaningfully support AI-assisted epilepsy care.

2 EPIKG: EPILEPSY KNOWLEDGE GRAPH

2.1 PROBLEM FORMULATION

We define EPIKG as a multi-relational heterogeneous knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of epilepsy-related entities and $\mathcal{E}$ denotes the set of typed relations between entities. Each entity belongs to one of five layers, including *genes, phenotypes, syndromes, treatments, and outcomes*. We collect a set of associated scientific documents $\mathcal{D} = \bigcup_{v \in \mathcal{V}} \mathcal{D}_v$, to extract various scientific findings or clinical observations for representing v, via a mapping function $\mathcal{P}$. Each relation is supported by evidence collected from clinical guidelines, where $\mathcal{D}_v$ denotes the collection of scientific documents associated with entity v. Therefore, each fact in $\mathcal{G}$ is represented as a triplet $\langle h, r, t \rangle$, where h and t are entities and r denotes the relation type between them. Cross-layer triplets connect entities across distinct clinical layers, enabling multi-hop reasoning for epilepsy diagnosis, treatment, and discovery.

2.2 DATA COLLECTION AND PROCESSING

Data Source. To construct the comprehensive entities $\mathcal{V}$ in EPIKG, we collect large-scale epilepsy-related scientific literature $\mathcal{D}$ from PubMed/PMC National Library of Medicine (2024). Using a comprehensive MeSH-based retrieval strategy covering epilepsy, seizure, epileptic encephalopathy, antiseizure medication, EEG, and pharmacogenomics, we retrieve over 120,000 candidate publications indexed between 1990 and 2024. These serve as the primary evidence source for extracting cross-layer clinical relations across genes, phenotypes, syndromes, treatments, and outcomes.
Ontology and Guideline Resources. To ensure clinically grounded and comprehensive relational edges $\mathcal{E}$, the schema of EPIKG is constructed from seven authoritative clinical resources, including ILAE 2022 Fisher et al. (2014), MeSH National Library of Medicine (2024), OMIM Hamosh et al. (2005), ChEBI Hastings et al. (2016), HPO Köhler et al. (2021), AES 2024 American Epilepsy Society (2024), and UMLS Bodenreider (2004). These resources provides curated term lists, concept hierarchies, and cross-resource identifiers. ILAE 2022 provides the authoritative classification of epilepsy syndromes and seizure types. OMIM provides gene–disease associations as the primary source for Gene entities. ChEBI provides standardized chemical identifiers for antiseizure medications in Treatment. HPO provides standardized phenotype vocabulary for seizure types and developmental outcomes. AES 2024 provides evidence-based treatment recommendations grounding Treatment relations. UMLS serves as the cross-ontology linking hub, integrating 200+ controlled vocabularies to enable synonym resolution and entity alignment.
Preprocessing. Both the literature corpus and ontology resources preprocessed through a two-stage pipeline. In the *automated filtering stage*, an LLM-based classifier applies five retention criteria to titles and abstracts: (i) primary focus on epilepsy or seizure disorders; (ii) clinical relevance; (iii) adequate generalizability; (iv) English-language publication; and (v) a unique record. In the *domain expert review stage*, borderline cases are adjudicated using established epilepsy systematic-review criteria Beghi et al. (2019); Kwan et al. (2011), retaining studies reporting original findings on syndrome classification, genetic etiology, ASM efficacy and safety, EEG biomarkers, pharmacogenomics, and treatment outcomes. Concurrently, ontology resources are normalized by resolving cross-resource identifier conflicts through UMLS CUI mapping and curating over 100 term lists covering entity aliases, abbreviations, and relation patterns across all layers, yielding a corpus of 48,166 papers and a validated entity vocabulary for graph construction.

2.3 KNOWLEDGE GRAPH CONSTRUCTION

Entity Extraction. We extract five layers (**L**) of entities for EPIKG, each corresponding to one clinical layer. For example, given the sentence *"Patients with Dravet Syndrome carrying SCN1A loss-of-function variants showed seizure freedom after treatment with Stiripentol combined with Valproate"*, the pipeline extracts: *Dravet Syndrome* (L1), *SCN1A* (L3), *Stiripentol* and *Valproate* (L4),

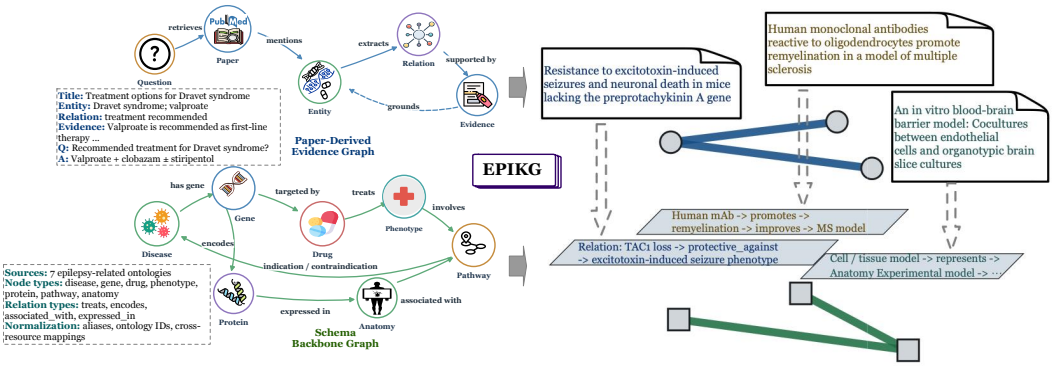


Figure 2: Pipeline overview of EPIKG, comprising two components. *Left:* paper-derived evidence graph is processed through an extraction pipeline that identifies entities, relations, and supporting evidence, mapped into relation graph. *Right:* An example of how EPIKG grounds the paper-derived evidence graph: given the query paper "Resistance to excitotoxin-induced seizures...", EPIKG retrieves the supporting reasoning path, linking the retrieved evidence back to the relation graph.

and *Seizure Freedom* (L5), which are subsequently linked as typed relations in EPIKG (Figure 2). Therefore, the five layers are as follows:

- **L1 Syndrome.** This layer includes epilepsy syndromes and seizure types following the ILAE 2022 taxonomy Fisher et al. (2014), including Dravet Syndrome, Lennox-Gastaut Syndrome, and West Syndrome. Entities are extracted from ILAE 2022 and MeSH National Library of Medicine (2024), and normalized through UMLS CUI mapping.
- **L2 Diagnostic.** Diagnostic entities capture EEG patterns, neuroimaging findings, and biomarkers, such as Spike-Wave Discharge, Hypsarrhythmia, and MRI T2 Signal Abnormality. Entities are extracted from MeSH and HPO Köhler et al. (2021).
- **L3 Gene.** Gene entities encode causal genetic factors associated with epilepsy, including genes and pathogenic variants. Examples include SCN1A, KCNQ2, TSC1, and CDKL5. Entities are extracted from OMIM gene–disease association Hamosh et al. (2005) and normalized using HGNC identifiers.
- **L4 Treatment.** Treatment entities cover antiseizure medications, surgical procedures, and neuromodulation therapies, including Valproate, Stiripentol, Vagus Nerve Stimulation, and Ketogenic Diet. Entities are extracted from ChEBI Hastings et al. (2016) chemical identifiers and AES 2024 American Epilepsy Society (2024) guideline-listed interventions, with abbreviation normalization handled through curated aliases.
- **L5 Outcome.** Outcome entities represent clinical endpoints such as seizure control, adverse effects, and developmental outcomes, including Seizure Freedom, Drug Resistance, Cognitive Impairment, and SUDEP. Entities are extracted from HPO Köhler et al. (2021) and MeSH.

Relation Construction. Relations are extracted through two complementary pipelines. (i) Rule-based extraction applies ontology-derived pattern matching to sentences containing co-occurring entity pairs. For example, given *"Valproate is recommended as first-line treatment for Dravet Syndrome but should be avoided in patients with SCN1A gain-of-function variants"*, the pipeline matches the pattern `[Treatment] + {recommended for / first-line} + [Syndrome]` to extract the triplet (Valproate, `treats`, Dravet Syndrome), and `[Treatment] + {avoid / contraindicated} + [Gene]` to extract (Valproate, `contraindicated_with`, SCN1A). (ii) LLM-based extraction employs MiniMax-Text-01 Li et al. (2025) on full-text articles with structured prompts specifying entity layers and relation types, enabling extraction of novel associations beyond rule-based coverage.

- **Etiology Relations.** These relations connect genetic factors with epilepsy syndromes and encode causal disease mechanisms, e.g., $\langle$*SCN1A*, CAUSED_BY_GENE, *Dravet Syndrome*$\rangle$.
- **Diagnostic Relations.** These relations connect syndromes with clinical manifestations and diagnostic findings, e.g., $\langle$*West Syndrome*, CHARACTERIZED_BY, *Hypsarrhythmia*$\rangle$.
- **Treatment Relations.** These relations encode therapeutic associations between syndromes, genes, and interventions, e.g., $\langle$*Dravet Syndrome*, TREATED_WITH, *Stiripentol*$\rangle$.
- **Outcome Relations.** These relations capture prognostic outcomes and treatment effects, e.g., $\langle$*Valproate*, CAUSES_TERATOGENICITY, *Neural Tube Defects*$\rangle$.

Table 1: Knowledge & Clinical Reasoning results. **T1a**: MCQ, Acc: Top-1 Accuracy (%). **T1b**: Open-ended QA, LJ: LLM-as-Judge (1–5). **T3**: Precision Medicine, GC: Guideline Concordance (%). **T4**: Treatment Recommendation, DFS: Drug Safety Score, KGEC: KG Evidence Coverage. R-L: ROUGE-L; BS: BERTScore F1; Reas.: Reasoning Accuracy. Results are mean ± std; Δ: avg. relative improvement. Claude S4 refers to Claude Sonnet 4.

		T1a: MCQ				T1b: Open-ended QA				T3: Prec. Med.		T4: Treatment Rec.			
Model	**RAG**	**Acc**	**R-L**	**BS**	**Reas.**	**LJ**	**R-L**	**BS**	**Reas.**	**Acc**	**GC**	**Acc**	**DFS**	**GC**	**KGEC**
GPT-4o	–	68.0±1.4	0.38±0.02	0.73±0.01	0.46±0.03	3.61±0.08	0.34±0.02	0.69±0.01	0.35±0.03	53.0±3.2	51.0±3.5	72.0±2.1	0.38±0.03	68.0±2.4	0.73±0.02
	w/	**75.0±1.2**	0.41±0.02	0.80±0.01	0.59±0.02	**4.33±0.06**	0.38±0.02	0.74±0.01	0.40±0.02	69.0±2.6	65.0±2.8	**81.0±1.7**	0.41±0.03	76.0±2.0	0.77±0.02
Claude S4	–	69.0±1.3	0.41±0.02	0.77±0.01	0.49±0.03	3.74±0.07	0.39±0.02	0.75±0.01	0.34±0.03	66.0±2.8	61.0±3.0	70.0±2.2	0.41±0.03	66.0±2.5	0.74±0.02
	w/	73.0±1.2	**0.45±0.01**	**0.82±0.01**	**0.64±0.02**	4.21±0.05	**0.42±0.01**	**0.80±0.01**	**0.42±0.02**	**82.0±2.1**	**79.0±2.3**	79.0±1.8	**0.45±0.02**	**78.0±1.9**	**0.82±0.01**
Gemini	–	63.0±1.5	0.36±0.02	0.74±0.01	0.47±0.03	3.54±0.09	0.35±0.02	0.71±0.02	0.37±0.03	58.0±3.1	51.0±3.4	65.0±2.4	0.36±0.03	61.0±2.7	0.67±0.02
	w/	68.0±1.4	0.42±0.02	0.81±0.01	0.56±0.02	3.96±0.07	0.37±0.02	0.76±0.01	0.41±0.02	74.0±2.5	74.0±2.6	74.0±2.0	0.42±0.03	72.0±2.2	0.76±0.02
Llama	–	57.0±1.6	0.33±0.02	0.70±0.02	0.30±0.03	3.42±0.10	0.29±0.02	0.63±0.02	0.29±0.03	44.0±3.5	48.0±3.3	58.0±2.6	0.33±0.04	55.0±2.8	0.70±0.02
	w/	66.0±1.4	0.36±0.02	0.76±0.01	0.41±0.03	3.95±0.07	0.35±0.02	0.66±0.01	0.34±0.03	60.0±2.9	65.0±2.8	67.0±2.2	0.36±0.03	64.0±2.4	0.74±0.02
Qwen	–	59.0±1.5	0.32±0.02	0.67±0.02	0.28±0.04	3.44±0.11	0.27±0.03	0.61±0.02	0.24±0.04	41.0±3.6	44.0±3.4	55.0±2.7	0.32±0.04	52.0±2.9	0.63±0.03
	w/	64.0±1.4	0.37±0.02	0.74±0.01	0.39±0.03	3.74±0.08	0.33±0.02	0.69±0.01	0.31±0.03	58.0±3.0	61.0±2.9	64.0±2.3	0.37±0.03	62.0±2.5	0.69±0.02
Mistral	–	51.0±1.8	0.31±0.03	0.64±0.02	0.24±0.04	3.54±0.12	0.26±0.03	0.58±0.02	0.25±0.04	38.0±3.8	44.0±3.5	51.0±2.9	0.31±0.04	49.0±3.0	0.61±0.03
	w/	61.0±1.5	0.35±0.02	0.72±0.01	0.33±0.03	3.85±0.09	0.31±0.02	0.62±0.02	0.30±0.03	51.0±3.2	58.0±3.1	62.0±2.4	0.35±0.03	60.0±2.6	0.68±0.02
Avg. Δ		**+11.3**	**+0.04**	**+0.07**	**+0.13**	**+0.51**	**+0.05**	**+0.05**	**+0.06**	**+34.8**	**+33.2**	**+15.6**	**+12.4**	**+14.1**	**+4.2**

2.4 Graph Mapping and EpiKG Statistics

After entity extraction and relation construction, raw literature evidence must be mapped to the curated entity vocabulary and validated before integration into EpiKG. This process consists of three stages. *First, Entity Normalization.* Literature mentions are normalized to canonical entity identifiers through a four-stage pipeline: (i) a exact match maps mentions to curated term lists; (ii) an alias match resolves abbreviations and synonyms (*e.g.*, VPA → Valproate, Nav1.1 → SCN1A); (iii) a semantic match uses sentence-transformer embeddings Reimers & Gurevych (2019) and cosine similarity to retrieve the closest entity for unresolved mentions; and (iv) a quality control verifies consistency through UMLS CUI alignment across ontologies. *Second, Relation Type Matching.* Extracted relation candidates are aligned to the 1,370 predefined cross-layer relation types. Rule-based candidates are matched through ontology-derived templates, while LLM-extracted relations are normalized using ontology-grounded semantic matching. Low-confidence candidates are flagged for expert review. *Finally, Triplet Validation.* Validated triplets are deduplicated, assigned to entity layers, and annotated with paper count as an evidence-frequency indicator. Low-evidence triplets supported by fewer than two independent sources are retained but flagged accordingly.

The resulting EpiKG contains 32,009 triplets across 24,324 unique entities and 1,370 cross-layer relation types. Rule-based extraction contributed 9,670 triplets (30.2%) and LLM-based extraction 22,339 triplets (69.8%). Cross-layer triplets number 14,576 (45.5%), covering all pairwise layer combinations. The densest cross-layer connections are between L1 Syndrome and L4 Treatment (3,217 triplets) and between L3 Gene and L1 Syndrome (2,845 triplets), capturing the gene–syndrome–treatment reasoning chains central to clinical decision-making. The median paper count per triplet is 3 (IQR: 1–8), and 4,612 triplets are supported by at least 10 independent publications.

3 EpiBench

3.1 Problem Formulation

We formulate EpiBench as a unified benchmark for evaluating evidence-intensive reasoning in epilepsy across both clinical and scientific settings. Each task is defined as:

$$\hat{y}_i = f_{\mathrm{LM}}(x_i, \mathcal{E}_i, \mathcal{C}_i), \tag{1}$$

where x_i denotes the primary task input, including clinical cases, EEG recordings, or scientific documents; $\mathcal{E}_i$ denotes the evidence context retrieved from EpiKG through Graph-RAG; $\mathcal{C}_i$ denotes optional task-specific context such as candidate options, supporting entities, or document metadata; and $\hat{y}_i$ denotes the model prediction evaluated against the gold-standard output y_i^*.

The benchmark covers diverse reasoning scenarios, including diagnosis-oriented question answering, pharmacogenomic treatment, biomarker discovery, research planning, and EEG-to-report generation.

3.2 Task Design

Epilepsy management requires jointly reasoning over syndrome classification, seizure phenotype, genetic factors, treatment contraindications, and patient-specific context. Accordingly, EpiBench

aims to evaluate reasoning, including clinical decision-making from phenotypic evidence, biomarker-driven precision medicine, treatment recommendations, and deep research planning from scientific literature. Beyond conventional QA settings, we further introduce a practical EEG-to-report generation task that reflects real-world clinical workflow Tatum IV et al. (2016); Pradeepkumar et al. (2026). We investigate whether LLMs, through interaction with EPIGRAPH, can leverage external epilepsy knowledge to support deeper EEG interpretation, clinical reasoning, and automatic neurologist-style report generation. The tasks are as follows.

Task 1: Clinical Decision Accuracy (CDA). This task evaluates epilepsy-specific clinical reasoning using recently published papers outside the EPIKG source snapshot, creating source separation for leakage control Singhal et al. (2023; 2025). The benchmark includes 1,000 multiple-choice questions and 5,199 open-ended questions, covering syndrome diagnosis, EEG interpretation, treatment selection, and phenotype reasoning. For example, given the question "A child with febrile seizures and *SCN1A* mutation is diagnosed with Dravet syndrome. What is the first-line treatment?", the model must select corresponding treatment from different options. Ground-truth answers are derived from expert-curated source literature.

Task 2: Clinical Report Generation (CRG). This task focuses on clinical impression generation from patient information and EEG descriptions (i.e., a summary of EEG signals). Given these inputs, LLMs must identify important findings, including characteristic EEG waveforms, dominant frequency patterns, and their associations with disease states, and then generate an impression for real-time patient documentation. Different from conventional QA benchmarks, CRG evaluates whether LLMs can perform evidence-grounded clinical reasoning or through interaction with EPIKG, integrating EEG findings with syndrome knowledge and evidence.

Task 3: Biomarker-Driven Precision Medicine (BPM). This task evaluates whether LLMs can select appropriate antiseizure medications (ASMs) from genetic variants and patient phenotypes. It requires multi-step reasoning over gene function, disease mechanisms, drug targets, and contraindication evidence Kuo et al. (2024). The benchmark contains 151 multiple-choice questions constructed from CPIC and ILAE 2022 guidelines. For example, given a patient with a *TSC2* variant and refractory focal seizures, the model should select *Everolimus* over *Carbamazepine* by reasoning through the path: *TSC2* $\rightarrow$ *mTOR Pathway* $\rightarrow$ *Everolimus*, and flag Carbamazepine as contraindicated.

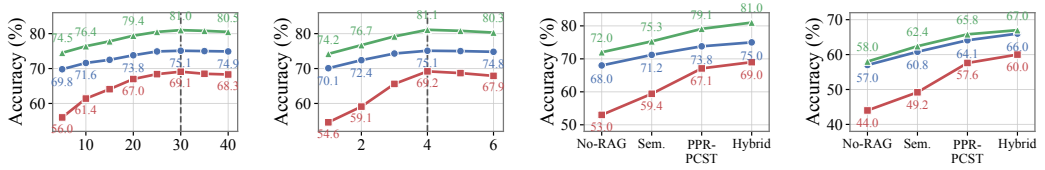
Task 4: Treatment Recommendation (TR). This task evaluates whether LLMs can recommend guideline-consistent therapies under patient-specific constraints Glauser et al. (2006). Questions are constructed from neurology subsets of *MedQA-USMLE* Jin et al. (2021) and *MMLU Professional Medicine* Hendrycks et al. (2020), covering 472 treatment recommendation cases in total. For example, given "a woman of childbearing age with juvenile myoclonic epilepsy", the model should recommend *Levetiracetam* while recognizing the risk of *Valproate*.

Task 5: Deep Research Planning (DRP). This task evaluates whether LLMs can perform scientific reasoning from epilepsy literature by proposing feasible research directions Wang et al. (2025). The benchmark is constructed from 163 PMC epilepsy papers, with expert annotations available for 30 papers. For example, given a paper on *KCNQ2*-related neonatal epilepsy, the model should identify unresolved questions regarding long-term neurodevelopmental outcomes and propose a longitudinal cohort study grounded in prior evidence from EPIKG. The generated plan should specify meaningful endpoints, such as neurodevelopmental outcomes.

Evaluation and Metrics. We evaluate EPIBENCH using both task accuracy and reasoning-oriented metrics. For **CDA**, we report Top-1 Accuracy for MCQs and ROUGE-L, BERTScore F1, and LLM-as-Judge for open-ended responses, evaluating both answer correctness and reasoning quality. For **CRG**, we use ROUGE-L and provide text alignment human evaluation. For **BPM** and **TR**, we report Top-1 Accuracy, Guideline Concordance, and Drug Safety Score to evaluate treatment selection, safety, and alignment with clinical recommendations. TR additionally includes KG Evidence Coverage, measuring whether predictions are supported by retrieved evidence paths from EPIKG. For **DRP**, we use ROUGE-L, BERTScore F1, Alignment Score, and LLM-as-Judge to evaluate scientific validity, coherence, and feasibility of generated research plans.

4 EXPERIMENTS

Models. We evaluate six LLMs spanning closed- and open-source families. *Closed-source:* GPT-4o Hurst et al. (2024), Claude Sonnet 4 Anthropic (2024), and Gemini 2.0 Flash Comanici et al.



(a) Max Subgraph Nodes (b) Max Path Depth (c) GPT-4o Ablation (d) Llama Ablation

Figure 3: Sensitivity analysis and ablation results of Graph-RAG. Blue denotes T1 MCQ, red denotes T3 Precision Medicine, and green denotes T4 Treatment Recommendation. Dashed vertical lines mark the selected optimal hyperparameter settings.

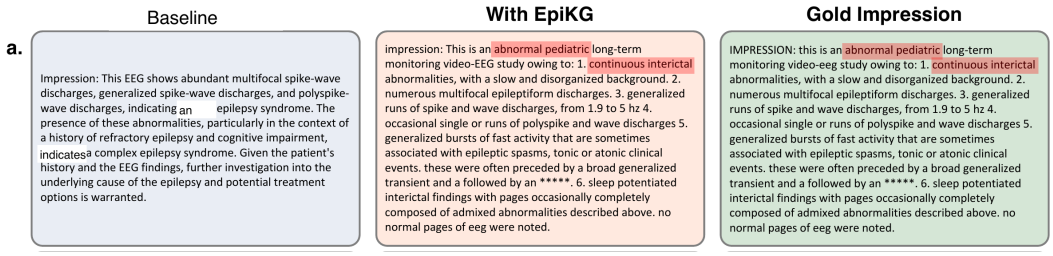


Figure 4: Impression generation example on S0001. Three column refer to medgemma model result, medgemma w/ EpiKG, and ground truth report.

(2025). *Open-source:* Llama-3.3-70B Grattafiori et al. (2024), Qwen2.5-72B Hui et al. (2024), and Mistral Small 3.1 Mistral AI (2025). All six models are evaluated on Task 1,3,4,5. For Task2 (**CRG**), the dataset access terms define an evaluation on *four locally deployed smaller models:* Gemma-3-4B Google DeepMind (2024), Llama-3.2-3B Grattafiori et al. (2024), MedGemma-4B Sellergren et al. (2025), and Qwen3-4B Hui et al. (2024).

Prompting. We construct prompts for all tasks following three principles: (1) chain-of-thought prompting Wei et al. (2022) to elicit step-by-step clinical reasoning; (2) role-playing prompts informing the model it is a skilled epileptologist; and (3) full task context as defined in §3.2, including the retrieved EpiKG subgraph serialised as structured reasoning paths. For MCQ tasks, temperature is set to 0.0; for generation tasks, to 0.3. All closed-source models are accessed via the OpenRouter API; all open-source models are deployed locally under standardised inference settings to ensure reproducibility. Full details see Appendix G.

Evaluation. Standard metrics include Top-1 Accuracy (MCQ), ROUGE-L, BERTScore F1, and LLM-as-Judge (GPT-4.1-mini, 1–5 scale). Domain-specific metrics include Clinical NER F1, Hallucination Rate (NLI-based), Guideline Concordance (ILAE 2022 / CPIC), Drug Safety Score (contraindication avoidance), KG Evidence Coverage, and Alignment Score. Full metric definitions see Appendix E.

Baselines. We compare against three knowledge-augmented systems: *MedRAG* Xiong et al. (2024) (flat dense retrieval over PubMed and textbooks), *DR.KNOWS* Gao et al. (2025) (UMLS-based KG paths for diagnostic tasks), and *AMG-RAG* Rezaei et al. (2025) (dynamic KG construction via LLM agents, no multi-task evaluation). None combines a curated domain-specific KG with multi-task clinical evaluation. As shown in Figure 8, Graph-RAG consistently dominates all baselines across all five task forms, with the largest margin on T3 Precision Medicine where domain-specific multi-hop reasoning is essential. Graph-RAG also remains competitive in inference efficiency (Figure 6).

Task 1: Clinical Decision Accuracy (CDA). (Tables 1) Graph-RAG yields consistent MCQ gains across all six models (avg. +11.3 pp). Open-source models benefit most (Mistral +19.6%, Llama +15.8%), while closed-source gains are smaller (GPT-4o +10.3%, Claude S4 +5.8%), showing that models with stronger parametric epilepsy knowledge have less marginal benefit from KG augmentation. Notably, LLM-as-Judge reasoning accuracy improves by +31.9% on average, substantially larger than ROUGE-L (+14.6%) and BERTScore (+7.6%) gains, indicating deeper gains in clinical reasoning chains beyond surface-level lexical overlap with reference answers. On open-ended QA, GPT-4o achieves 4.33/5.0 (+19% over No-RAG). Graph-RAG outperforms MedRAG (62%)

and AMG-RAG (66.3%) by substantial margins, confirming the advantage of domain-specific KG curation over general-purpose retrieval.

Task 2: Clinical Report Generation (CRG) (Table 2). We evaluate CRG using the Harvard EEG database (site `s0001`) Zafar et al. (2025); Sun et al. (2025) processed with the CLEM preprocessing pipeline Pradeepkumar et al. (2026) to extract structured EEG impression data. Each instance consists of EEG descriptions, patient information, and neurologist-written impression reports as ground truth. Access terms specify local evaluation with Gemma, Llama, MedGemma, and Qwen.

Table 2 shows that generating clinically meaningful EEG impressions remains highly challenging for current LLMs, with overall METEOR scores remaining relatively low. Integrating EPIKG consistently improves performance for all models. Specifically, MedGemma-4B shows the largest gain (+30.8% METEOR), demonstrating the largest benefit for domain-pretrained models when structured clinical context is provided. Figure 4 further presents two representative EEG impression generation examples using the MedGemma baseline and its EPIKG-augmented configuration. Integrating EPIKG enables the model to produce more precise impressions, including detailed waveform descriptions, abnormal slowing patterns, and clinically relevant interpretations that more closely align with neurologist-written reports.

Table 2: Generation tasks for neurology reports on Harvard Electroencephalography Database v4.1 Zafar et al. (2025); Sun et al. (2025) processed using the pipeline from Pradeepkumar et al. (2026). Results are reported as mean ± standard deviation; Δ denotes relative improvement (%).

Model	RAG	METEOR	Δ
Gemma-3-4b-it	–	0.23±0.06	–
	w/	**0.26±0.08**	+13.0
Llama-3.2-3B-it	–	0.29±0.08	–
	w/	**0.31±0.11**	+6.8
Medgemma-4b-it	–	0.26±0.13	–
	w/	**0.34±0.17**	+30.8
Qwen3-4B-it-2507	–	0.26±0.06	–
	w/	**0.27±0.07**	+3.8

Task 3: Biomarker-Driven Precision Medicine (BPM) (Table 1). Graph-RAG produces the largest relative gains across all tasks. *Claude Sonnet 4* reaches 82% (+24%); open-source models improve even more dramatically, *Qwen* +42%, *Llama* +36%, reversing the usual closed/open-source performance gap and showing that pharmacogenomic knowledge is more uniformly absent from all LLM parameter spaces, making KG augmentation equally essential regardless of model scale. Critically, the no-KG *Mistral* baseline scores 38%, near the 25% random baseline for four-option MCQ, confirming insufficient pharmacogenomic reasoning in general-purpose LLM parameters; prompting alone fails to recover this capability.

Task 4: Treatment Recommendation (TR) (Table 1). On MedQA-USMLE, Graph-RAG improves Top-1 Accuracy +15.6%, Drug Safety Score +12.4%, and Guideline Concordance +14.1%. On MMLU Professional Medicine, DFS and GC improve far more substantially (+28.1% and +28.4%), while raw accuracy gains are similar (+17.2%). This divergence indicates that MMLU questions, which require broader multi-step clinical reasoning, expose safety and guideline gaps more clearly than narrow factual recall tasks. Graph-RAG improves clinical safety alongside answer correctness in treatment recommendation. The positive correlation between KGEC and DFS improvements (+4.2% vs +13.2% across the two datasets) further confirms that stronger KG utilisation directly translates to better contraindication avoidance.

Table 3: Deep Epi-Research results. LJ: LLM-as-Judge (1–5); R-L: ROUGE-L. Results are reported as mean ± standard deviation; Δ denotes relative improvement (%).

Model	RAG	LJ	Δ	R-L	Δ
GPT-4o	–	3.56±0.12	–	0.34±0.02	–
	w/	**4.25±0.09**	+19.4	0.41±0.02	+20.6
Claude S4	–	3.69±0.11	–	0.36±0.02	–
	w/	4.13±0.08	+11.9	**0.43±0.02**	+19.4
Gemini	–	3.49±0.13	–	0.33±0.03	–
	w/	3.88±0.10	+11.2	0.39±0.02	+18.2
Llama	–	3.37±0.14	–	0.31±0.03	–
	w/	3.87±0.11	+14.8	0.37±0.02	+19.4
Qwen	–	3.39±0.15	–	0.32±0.03	–
	w/	3.66±0.12	+8.0	0.36±0.02	+12.5
Mistral	–	3.49±0.16	–	0.30±0.03	–
	w/	3.77±0.13	+8.0	0.35±0.03	+16.7

Task 5: Deep Research Planning (DRP) (Table 3). Graph-RAG improves LLM-as-Judge scores by +12.2% on average; *GPT-4o* achieves 4.25/5.0 (+19.4%). The closed/open-source performance gap narrows substantially relative to T1 and T3: *Llama* reaches 3.87 vs *GPT-4o*'s 4.25 under Graph-RAG (gap of 0.38), compared to a gap of 9 pp on T1 MCQ. This convergence indicates that research plan generation depends more on structural reasoning ability, the capacity to follow retrieved KG paths and synthesise coherent hypotheses, than on parametric domain knowledge, and that Graph-RAG can substantially fill the capability gap between open and closed source models on this dimension.

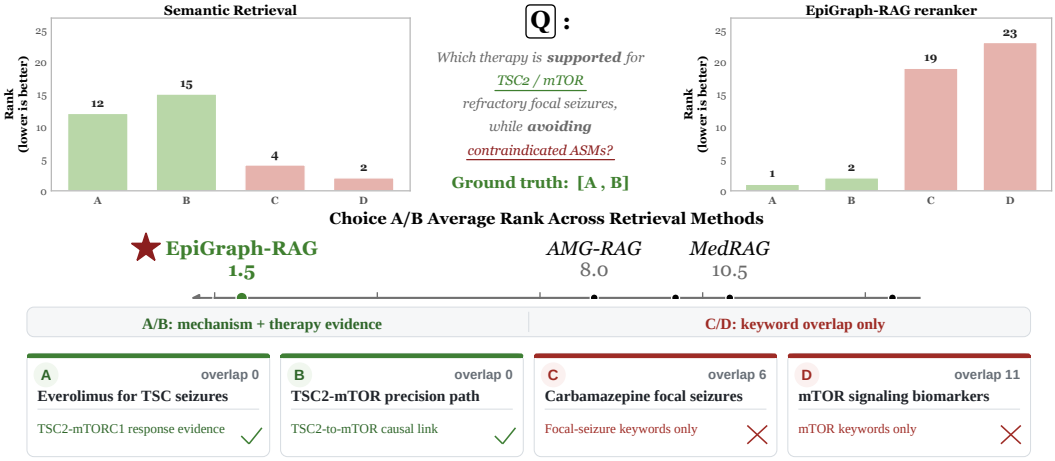


Figure 5: Case study of a TSC2 precision-medicine query. Semantic Retrieval ranks distractor candidates C and D highly due to surface keyword overlap, whereas EpiGraph-RAG reranks the true supporting evidence A and B to the top by following the TSC2/mTOR therapeutic path and avoiding contraindicated ASMs. Green denotes ground-truth supporting evidence, and dark red denotes non-ground-truth evidence.

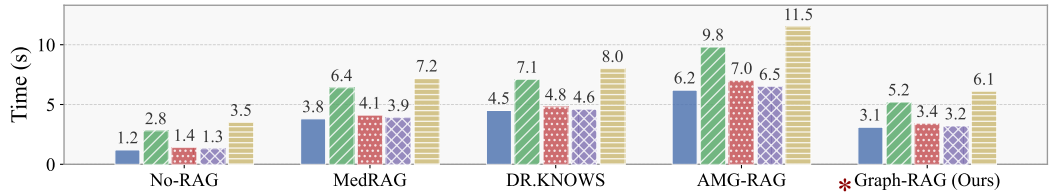


Figure 6: Overview of **EPIBENCH** Running Time. Blue denotes T1 Knowledge QA, green denotes T2 Report Generation, red denotes T3 Precision Medicine, purple denotes T4 Treatment Recommendation, and olive denotes T5 Deep Research. The red star highlights Graph-RAG (Ours).

Ablation and Sensitivity (Figure 3). PPR-PCST outperforms semantic retrieval by +2.6–7.7 pp on T1 and +7.7–8.4 pp on T3; Hybrid adds a further +1.2–2.4 pp. The advantage of graph-based retrieval is most pronounced on T3, where multi-hop reasoning from gene to drug mechanism requires traversing KG topology beyond flat retrieval. Optimal subgraph size is 30 nodes and path depth is 4 hops; T3 shows the strongest depth sensitivity (depth 2: 58.7% vs depth 4: 69.0% for GPT-4o), confirming that pharmacogenomic reasoning inherently requires 3–4 hop chains.

5 CONCLUSION AND FUTURE WORK

EPIKG and EPIBENCH together establish a knowledge-grounded evaluation framework **EPIGRAPH** for epilepsy. EPIKG provides the first open-source domain-specific epilepsy KG integrating seven ontologies across five clinical layers, and EPIBENCH demonstrates that Graph-RAG over EPIKG consistently improves LLM performance across all five clinical task forms relative to both direct-inference baselines and general-purpose retrieval systems. The case study and retrieval analyses localize these gains in decisive cross-layer paths: disease–gene–pathway–treatment links elevate clinically relevant evidence, and provenance fields connect each retrieved claim to its supporting source. The task-specific outcomes further connect retrieval quality to answer accuracy, report quality, medication safety, guideline concordance, and research-plan quality. These results support three concrete directions. First, versioned EPIKG updates integrate newly published evidence and clinician adjudication, enabling temporal evaluation as diagnostic and treatment knowledge evolves. Second, retrieved subgraphs and evidence paths supervise retrieval-augmented fine-tuning and persistent memory, converting graph structure into reusable clinical reasoning trajectories. Third, multilingual, multi-institutional, and longitudinal extensions of EPIBENCH evaluate transfer across care settings, rare syndromes, and treatment trajectories. Together, these directions position EPIGRAPH as infrastructure for tracing clinical claims from evidence retrieval through model reasoning to outcome-oriented evaluation, advancing auditable neurological AI.

ETHICS STATEMENT

Intended use and clinical safety. EPIGRAPH and EPIBENCH are research and educational resources for studying knowledge-grounded language models. They are not medical devices and have not been validated for diagnosis, treatment, clinical documentation, medication selection, or autonomous clinical decision making. They must not be deployed for patient care or used to replace qualified clinical judgment. Any research involving clinical scenarios requires review by a licensed clinician with access to the complete patient history and applicable local guidance. Guideline Concordance, Drug Safety Score, and all other reported benchmark metrics measure performance on the specified test cases; they do not certify clinical safety, efficacy, or regulatory compliance.

Knowledge quality, provenance, and correction. Automated extraction can introduce incorrect, incomplete, or outdated entities and relations, including clinically consequential treatment and contraindication edges. Released graph records therefore carry source identifiers, publication or guideline versions, evidence-strength indicators, and extraction and validation status. Expert-reviewed subsets are distinguished from automatically extracted records. Versioned releases preserve superseded edges and their supporting evidence. A public issue channel supports reports of incorrect or harmful edges, with documented review, correction, and withdrawal procedures; treatment-related reports receive priority review. The resource must not be interpreted as a current clinical guideline.

Clinical data governance and privacy. The Harvard EEG component originates from a credentialed clinical-data resource and remains subject to its Data Use Agreement, access controls, and applicable institutional review requirements. The public release does not redistribute raw records, derived EEG descriptions, neurologist-written impressions, or other restricted clinical text. It provides preprocessing and evaluation code, schemas, aggregate summaries, and source record identifiers so authorized users can reconstruct the benchmark only after obtaining access through the original credentialed process. Any future release of record-level clinical material requires written authorization covering the exact fields, the required institutional approval or exemption, and residual-identifier screening. A designated reporting channel supports prompt investigation and removal if sensitive information is discovered.

Licensing and responsible release. Each released field is accompanied by source-level provenance and applicable license or access terms. Author-created code and metadata are licensed separately from third-party resources. Content that cannot be redistributed under compatible terms is omitted or represented only by source identifiers; it is not relicensed under the project license. Downstream use remains subject to the original terms of OMIM, UMLS, clinical guidelines, Harvard EEG, and all other source resources.

Access controls, misuse prevention, and incident response. Public artifacts exclude patient-level content and executable clinical decision interfaces. Any hosted demonstration uses synthetic or fully public examples, displays research-use notices, limits automated bulk access, and records security-relevant events under institutional policy. Users are expected to preserve source access controls, avoid re-identification, and prevent outputs from being presented as patient-specific advice. Reports of privacy leakage, unsafe recommendations, licensing violations, or exploitable model behavior trigger triage, temporary artifact withdrawal when warranted, documented investigation, and a versioned corrective release.

Evaluation validity, coverage, and fairness. LLM-generated questions, LLM-judged outputs, and clinician-authored references are explicitly labeled and reported separately. For Task 5, the 30 expert-annotated papers are distinguished from the 133 LLM-judged papers; LLM-as-Judge results provide supplemental scalable evaluation and are not evidence of clinical validity. The current evidence base is English-language, reflects predominantly Western clinical practice, contains a single-site EEG evaluation, and has lower coverage for rare syndromes and underrepresented populations. These data do not establish performance across languages, regions, institutions, demographic groups, or uncommon conditions. Coverage and subgroup results are reported when permitted by the source data, and unavailable groups are explicitly marked as unevaluated.

Conflicts of interest. The authors declare no financial or commercial conflicts of interest related to the proprietary or open-source models evaluated in this work. Model comparisons were conducted for research benchmarking and do not constitute product endorsement.

AI USE STATEMENT

Generative AI was not used to formulate the research questions, design the methods, construct the datasets or knowledge graph, conduct the experiments, analyze the results, or produce the scientific claims in this paper. AI-assisted tools were used solely for grammar checking and presentation-quality review, including the identification of sentence fragments, dangling words, discontinuities, and formatting inconsistencies. The authors reviewed and approved every resulting edit and retain full responsibility for the manuscript.

REPRODUCIBILITY STATEMENT

The anonymous supplementary release contains the EPIKG construction pipeline, Graph-RAG retrieval code, five EPIBENCH task runners, preprocessing schemas, prompt templates, evaluation metrics, and experiment configurations. Appendix F documents the data sources and licenses, graph schema, task construction, retrieval hyperparameters, model settings, metric definitions, and complete prompts. Released graph records retain source identifiers and evidence provenance, allowing retrieved paths to be traced to their supporting resources. All public benchmark components and code are specified in Appendix F.4. The Harvard EEG component remains under credentialed access and its Data Use Agreement.

REFERENCES

Almuntasar Al-Breiki, Said Al-Sinani, Ahmed Elsharaawy, Mohamed Usama, and Tariq Al-Saadi. Artificial intelligence in epilepsy: a systemic review. *Journal of Epilepsy Research*, 15(1):2–22, 2025.

American Epilepsy Society. AES clinical practice guidelines, 2024. URL `https://www.aesnet.org`. Accessed: 2025.

Anthropic. Claude Sonnet 4. Technical report, Anthropic, 2024. URL `https://www.anthropic.com/claude`.

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

Ettore Beghi, Giorgia Giussani, Emma Nichols, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, Haftom Niguse Abraha, Mina G Adib, Sutapa Agrawal, Fares Alahdab, et al. Global, regional, and national burden of epilepsy, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, 18(4):357–375, 2019.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Stephen Bonner, Ian P Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, Andreas Bender, Charles Tapley Hoyt, and William L Hamilton. A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Briefings in Bioinformatics*, 23(6):bbac404, 2022.

Lang Cao, J. Sun, and Adam Cross. Autord: An automatic and end-to-end system for rare disease knowledge graph construction based on ontology-enhanced large language models (preprint). *JMIR Medical Informatics*, 12, 2024.

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

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, 16 (1):3280, 2025a.

Zheng Chen, Yasuko Matsubara, Yasushi Sakurai, and Jimeng Sun. Long-term eeg partitioning for seizure onset detection. In *Proc. AAAI Conf. Artif. Intell.*, pp. 14221–14229, 2025b.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Hejie Cui, Jiaying Lu, Ran Xu, Shiyu Wang, Wenjing Ma, Yue Yu, Shaojun Yu, Xuan Kan, Chen Ling, Liang Zhao, et al. A review on knowledge graphs for healthcare: Resources, applications, and promises. *Journal of biomedical informatics*, pp. 104861, 2025.

Robert S Fisher, Carlos Acevedo, Alexis Arzimanoglou, Alicia Bogacz, J Helen Cross, Christian E Elger, Jerome Engel Jr, Lars Forsgren, Jacqueline A French, Mike Glynn, et al. Ilae official report: a practical clinical definition of epilepsy. *Epilepsia*, 55(4):475–482, 2014.

Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *Jmir Ai*, 4:e58670, 2025.

Tracy Glauser, Elinor Ben-Menachem, Blaise Bourgeois, Avital Cnaan, David Chadwick, Carlos Guerreiro, Reetta Kälviäinen, Richard Mattson, Emilio Perucca, and Torbjorn Tomson. Ilae treatment guidelines: evidence-based analysis of antiepileptic drug efficacy and effectiveness as initial monotherapy for epileptic seizures and syndromes. *Epilepsia*, 47(7):1094–1120, 2006.

Google DeepMind. Gemma 3 technical report. Technical report, Google, 2024. URL `https://ai.google.dev/gemma`.

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

Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl_1):D514–D517, 2005.

Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, 44(D1): D1214–D1219, 2016.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M Brower, et al. The human phenotype ontology in 2021. *Nucleic acids research*, 49(D1):D1207–D1217, 2021.

Rikuto Kotoge, Zheng Chen, Tasuku Kimura, Yasuko Matsubara, Takufumi Yanagisawa, Haruhiko Kishima, and Yasushi Sakurai. Evobrain: Dynamic multi-channel EEG graph modeling for time-evolving brain networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.

Chin-Wei Kuo, Ching-Yun Huang, Hsuan-Ming Chen, Jing-Jane Tsai, and Chin-Wei Huang. Review of pharmacogenetics of antiseizure medications: focusing on genetic variants of mechanistic targets. *Frontiers in Pharmacology*, 15:1411487, 2024.

Patrick Kwan, Steven C Schachter, and Martin J Brodie. Drug-resistant epilepsy. *New England Journal of Medicine*, 365(10):919–926, 2011.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, et al. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*, 2025.

Lincan Li, Zheng Chen, and Yushun Dong. Llm as clinical graph structure refiner: Enhancing representation learning in eeg seizure diagnosis. *arXiv preprint arXiv:2604.28178*, 2026a.

Lincan Li, Rikuto Kotoge, Xihao Piao, Zheng Chen, and Yushun Dong. Optimizing eeg graph structure for seizure detection: An information bottleneck and self-supervised learning approach. *arXiv.2604.01595*, 2026b.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Yuxing Lu, Sin Yee Goi, Xukai Zhao, and Jinzhuo Wang. Biomedical knowledge graph: A survey of domains, tasks, and real-world applications. *arXiv preprint arXiv:2501.11632*, 2025.

Alfredo Lucas, Andrew Revell, and Kathryn A Davis. Artificial intelligence in epilepsy—applications and pathways to the clinic. *Nature Reviews Neurology*, 20(6):319–336, 2024.

Mingyu Derek Ma, Chenchen Ye, Yu Yan, Xiaoxuan Wang, Peipei Ping, Timothy S Chang, and Wei Wang. Clibench: Multifaceted evaluation of large language models in clinical decisions on diagnoses, procedures, lab tests orders and prescriptions. *arXiv preprint arXiv*, 2406, 2024.

Mistral AI. Mistral small 3.1, 2025. URL `https://mistral.ai/news/mistral-small-3-1`. Accessed: 2025.

National Library of Medicine. Medical subject headings (MeSH), 2024. URL `https://www.nlm.nih.gov/mesh`. Accessed: 2025.

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

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Jathurshan Pradeepkumar, Zheng Chen, and Jimeng Sun. Neural signals generate clinical notes in the wild. *arXiv preprint arXiv:2601.22197*, 2026.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Mohammad Reza Rezaei, Reza Saadati Fard, Jayson L Parker, Rahul G Krishnan, and Milad Lankarany. Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between llms and evolving medical knowledge. *arXiv preprint arXiv:2502.13010*, 2025.

Satya S Sahoo, Samden D Lhatoo, Deepak K Gupta, Licong Cui, Meng Zhao, Catherine Jayapandian, Alireza Bozorgi, and Guo-Qiang Zhang. Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care. *Journal of the American Medical Informatics Association*, 21(1):82–89, 2014.

Astghik Sargsyan, Philipp Wegner, Stephan Gebel, Abish Kaladharan, Priya Sethumadhavan, Vanessa Lage-Rupprecht, Johannes Darms, Bruce Schultz, Jürgen Klein, Marc Jacobs, et al. The epilepsy ontology: a community-based ontology tailored for semantic interoperability and text mining. *Bioinformatics advances*, 3(1):vbad033, 2023.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Priyanka Sen, Sandeep Mavadia, and Amir Saffari. Knowledge graph-augmented language models for complex question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pp. 1–8, 2023.

Ruosi Shao, Md Shamim Seraj, Kangyi Zhao, Yingtao Luo, Lincan Li, Bolin Shen, Averi Bates, Yue Zhao, Chongle Pan, Lisa Hightow-Weidman, et al. Llm-empowered patient-provider communication: A data-centric survey from a clinical perspective. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 684–705, 2025.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Chenxi Sun, Jin Jing, Niels Turley, Callison Alcott, Wan-Yee Kang, Andrew J Cole, Daniel M Goldenholz, Alice Lam, Edilberto Amorim, Catherine Chu, et al. Harvard electroencephalography database: A comprehensive clinical electroencephalographic resource from four boston hospitals. *Epilepsia*, 2025.

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

William O Tatum IV, Olga Selioutski, Juan G Ochoa, Heidi Munger Clary, Janna Cheek, Frank W Drislane, and Tammy N Tsuchida. American clinical neurophysiology society guideline 7: guidelines for eeg reporting. *The Neurodiagnostic Journal*, 56(4):285–293, 2016.

Jesper Tveit, Harald Aurlien, Sergey Plis, Vince D Calhoun, William O Tatum, Donald L Schomer, Vibeke Arntsen, Fieke Cox, Firas Fahoum, William B Gallentine, et al. Automated interpretation of clinical electroencephalograms using artificial intelligence. *JAMA neurology*, 80(8):805–812, 2023.

Chong Wang, Mengyao Li, Junjun He, Zhongruo Wang, Erfan Darzi, Zan Chen, Jin Ye, Tianbin Li, Yanzhou Su, Jing Ke, et al. A survey for large language models in biomedicine. *Artificial Intelligence in Medicine*, pp. 103268, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, 2024.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 535–546, 2021.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*, pp. 102–120. Springer, 2024.

S Zafar, T Loddenkemper, JW Lee, A Cole, D Goldenholz, J Peters, A Lam, E Amorim, C Chu, S Cash, et al. Harvard electroencephalography database (version 4.1). *Brain Data Science Platform*, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. Greaselm: Graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860*, 2022.

Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. Diagnosisarena: benchmarking diagnostic reasoning for large language models. *arXiv preprint arXiv:2505.14107*, 2025.

APPENDIX CONTENTS

A Related Work 17

B EpiKG Construction Details 18
- B.1 Ontology Sources and Coverage 18
- B.2 Entity Normalisation Protocol 18
- B.3 Relation Extraction Details 18
- B.4 Knowledge Graph Statistics 19

C EpiBench Dataset Details 20
- C.1 Dataset Construction Protocols 20
- C.2 Gold Standard Annotation 20
- C.3 Dataset Statistics 20

D Graph-RAG Retriever Details 21
- D.1 PPR-PCST Implementation 21
- D.2 Semantic Retrieval Implementation 21
- D.3 Hyperparameter Settings 21

E Evaluation Metrics 22

F EpiBench and EpiKG Documentation 23

G Prompts 25

H Additional Experimental Results 27

I Scope and Future Work 28

J Reviewer-Audit Addendum 29

K Discussion 30

A Related Work

Epilepsy knowledge resources. Epilepsy-specific knowledge has been encoded in various forms, ranging from clinical ontologies to curated databases. The ILAE 2022 classification system Fisher et al. (2014) provides a standardised taxonomy of seizure types and epilepsy syndromes. OMIM Hamosh et al. (2005) and HPO Köhler et al. (2021) encode gene–disease associations and phenotypic descriptions respectively, while ChEBI Hastings et al. (2016) provides chemical identifiers for antiseizure medications. Epilepsy-specific ontologies have also been developed to support semantic interoperability and clinical text mining Sahoo et al. (2014); Sargsyan et al. (2023). CPIC guidelines formalise pharmacogenomic rules for drug selection based on genetic variants (*Clinical Pharmacogenomics Implementation Consortium, cpicpgx.org*). EpiKG integrates these resources into a unified relational structure, enabling multi-hop clinical reasoning through their combined coverage.

Biomedical knowledge graphs. Knowledge graphs have been widely adopted in biomedicine for drug discovery, disease characterisation, and clinical decision support Cui et al. (2025); Bonner et al. (2022). General-purpose biomedical KGs such as UMLS Bodenreider (2004) and PrimeKG Chandak et al. (2023) provide broad coverage with general-purpose relation types and clinical granularity. Domain-specific KGs have been applied across biomedical domains Lu et al. (2025), but no comparable resource exists for epilepsy. EpiKG fills this gap by combining ontology seeding with LLM-based relation extraction to produce a clinically grounded, epilepsy-specific KG.

Graph-RAG and knowledge-augmented LLMs. Retrieval-augmented generation has emerged as a dominant paradigm for grounding LLM outputs in external knowledge Lewis et al. (2020). Graph-based retrieval extends flat document retrieval by exploiting relational structure to recover multi-hop reasoning chains Yasunaga et al. (2021); Zhang et al. (2022). Subgraph-based context augmentation has been shown to improve factual grounding and reasoning coherence over dense retrieval alone Sun et al. (2023); Sen et al. (2023). Medical applications include DR.KNOWS Gao et al. (2025), MedRAG Xiong et al. (2024), and AMG-RAG Rezaei et al. (2025), which apply KG-augmented retrieval to general clinical tasks. More recently, MedReason Wu et al. (2025) demonstrates that KG-elicited reasoning steps improve factual accuracy in medical LLMs. EpiGraph differs from these systems by combining a curated domain-specific KG with a multi-task evaluation benchmark, enabling controlled ablation of retrieval quality across diverse epilepsy clinical reasoning tasks.

Clinical NLP benchmarks. Existing biomedical NLP benchmarks include MedQA Jin et al. (2021), MMLU Professional Medicine Hendrycks et al. (2020), PubMedQA Jin et al. (2019), and MedMCQA Pal et al. (2022), which evaluate factual recall and clinical reasoning on general medical knowledge. Epilepsy-specific evaluation has centered on narrow tasks such as seizure classification from EEG signals Tveit et al. (2023) or pharmacogenomic variant interpretation Kuo et al. (2024). EpiBench is the first benchmark to provide multi-task evaluation of LLMs across the full epilepsy clinical reasoning pipeline, from syndrome identification and treatment recommendation to research planning.

LLM evaluation in clinical settings. Recent work has evaluated LLMs on clinical tasks Singhal et al. (2023; 2025), finding that even large models struggle with specialised clinical knowledge and guideline adherence. LLM-as-Judge approaches have been proposed for evaluating open-ended clinical generation Chen et al. (2025a), and domain-specific metrics such as guideline concordance have been introduced to capture clinically meaningful performance beyond standard NLP metrics Yu et al. (2024). EpiBench adopts and extends these evaluation practices with epilepsy-specific metrics including Drug Safety Score and KG Evidence Coverage.

B EpiKG Construction Details

B.1 Ontology Sources and Coverage

Table 4 lists the seven ontology sources integrated in EpiKG, together with the entity layer each source contributes, the number of seed entities extracted, and the license under which each resource is used.

Table 4: Ontology sources used in EpiKG construction. Seed counts are after epilepsy-specific filtering.

Source	Layer	Seed entities	License
ILAE 2022	L1 Syndrome	312	Open access
MeSH	L1, L2, L5	2,841	Public domain
HPO	L2, L5	1,654	CC-BY 4.0
OMIM	L3 Gene	3,892	Academic use
HGNC	L3 Gene	4,107	CC-BY 4.0
ChEBI	L4 Treatment	2,876	CC-BY 4.0
AES 2024	L4 Treatment	643	Open access
Total		**24,324**	

B.2 Entity Normalisation Protocol

Entity normalisation maps surface mentions extracted from literature to canonical ontology identifiers. The pipeline proceeds in three steps. First, named entity recognition using a biomedical NER model identifies entity spans in epilepsy full-text articles. Second, candidate spans are matched to ontology entries via exact string matching, then fuzzy matching with a threshold of 0.85, and finally UMLS CUI lookup for unresolved spans. Third, aliases and abbreviations are resolved through curated synonym lists derived from each source ontology. Entities unmatched to a canonical identifier after all three steps are discarded.

B.3 Relation Extraction Details

Rule-based extraction. Pattern templates are defined for each of the six relation types. Each template specifies a syntactic pattern (subject entity layer, trigger phrase set, object entity layer) applied to dependency-parsed sentences containing co-occurring entity pairs. Table 5 lists the trigger phrase sets for each relation type.

Table 5: Rule-based extraction trigger phrases for each relation type.

Relation	Subject layer	Trigger phrases (examples)
treats	Treatment	recommended for, first-line, effective in
contraindicated_with	Treatment	avoid, contraindicated, avoid use
associated_with	Gene	associated with, linked to, implicated in
characteristic_of	Diagnostic	characteristic of, consistent with, seen in
encodes	Gene	encodes, produces, results in
expressed_in	Gene	expressed in, detected in, localised to

LLM-based extraction. MiniMax-Text-01 Li et al. (2025) is prompted with a structured template specifying the six relation types, their definitions, and the five entity layers. The prompt instructs the model to extract triplets in the form `(head entity, relation, tail entity)` from each full-text passage, restricted to entity pairs where both head and tail have been normalised to canonical identifiers. Extracted triplets are deduplicated and merged with rule-based extractions; conflicts are resolved by retaining the triplet with the higher paper count $\mathcal{P}$.

B.4 KNOWLEDGE GRAPH STATISTICS

Table 6: EPIKG statistics by entity layer and relation type.

Category	**Count**	**Avg. paper count $\mathcal{P}$**
Entity layers		
L1 Syndrome	312	6.3
L2 Diagnostic	1,654	4.8
L3 Gene	4,107	5.2
L4 Treatment	3,412	7.1
L5 Outcome	2,198	3.9
Protein	5,841	4.1
Anatomy	4,683	3.6
Other	2,117	2.8
Total nodes	**24,324**	**4.6**
Relation types		
treats	8,214	4.2
contraindicated_with	1,843	3.1
associated_with	12,307	3.8
characteristic_of	3,891	2.9
encodes	4,012	2.6
expressed_in	1,742	2.4
Total edges	**32,009**	**3.0 (median)**

C EPIBENCH DATASET DETAILS

C.1 DATASET CONSTRUCTION PROTOCOLS

T1 EpiBench-MCQ and EpiBench-QA. Questions are generated from 2025–2026 epilepsy papers retrieved from PubMed using epilepsy-specific MeSH terms, restricted to papers published after the EPIKG construction cutoff to prevent knowledge leakage. MCQ distractors are generated by substituting semantically similar but clinically incorrect entities from EPIKG. Open-ended questions are generated by prompting GPT-4.1-mini with the paper abstract and instructing it to produce factual questions whose answers are explicitly stated in the source text.

T2 Harvard EEG dataset. EEG text descriptions and computed statistics (band power, spike rate) are extracted from the Harvard Electroencephalography Database v4.1 Zafar et al. (2025); Sun et al. (2025) using the preprocessing pipeline from Pradeepkumar et al. (2026). Neurologist-written clinical impressions are used as gold standards verbatim within the credentialed-access environment; neither the impressions nor derived clinical text are included in the public release.

T3 Pharmacogenomic MCQs. 151 MCQs are constructed by clinical experts from established gene–drug rules in CPIC guidelines and ILAE 2022 gene-specific recommendations, spanning six pharmacogenomic categories: ion channel, mTOR pathway, metabolic, pharmacokinetic safety, EEG-guided, and multi-hop reasoning. Each question requires selecting one ASM from four candidates; distractors are drawn from the same drug class to require mechanistic discrimination.

T4 MedQA-USMLE and MMLU. Epilepsy-relevant questions are filtered from MedQA-USMLE Jin et al. (2021) and MMLU Professional Medicine Hendrycks et al. (2020) using keyword matching on epilepsy syndromes, ASM names, and EEG findings. Retained questions require resolution against ILAE 2022 or CPIC guidelines.

T5 PMC Research Planning. 163 epilepsy full-text papers are sampled from PubMed Central using epilepsy MeSH terms, restricted to original research articles published between 2020–2024. Expert annotations are provided by domain collaborators for 30 papers; LLM-as-Judge evaluations cover the remaining 133.

C.2 GOLD STANDARD ANNOTATION

Expert annotations for T5 are produced by neurologists with >5 years of epilepsy research experience. Annotators are provided with the full paper text and instructed to write (i) a focused research question the paper addresses, (ii) a study design rationale, and (iii) a list of required data sources. Inter-annotator agreement is measured on a 20-paper overlap subset using Cohen's κ; $\kappa = 0.81$ for research question quality and $\kappa = 0.76$ for feasibility ratings.

C.3 DATASET STATISTICS

Table 7: EPIBENCH dataset statistics by task.

Task	Dataset	Instances	Gold standard	Split
T1a MCQ	EpiBench-MCQ	1,000	Source papers	Test
T1b QA	EpiBench-QA	5,199	Reference answers	Test
T2 Report	Harvard EEG v4.1	40,000+	Neurologist impressions	Test
T3 Prec.	CPIC/ILAE MCQ	151	Guidelines	Test
T4 Treat.	MedQA-USMLE	200	Guidelines	Test
T4 Treat.	MMLU Prof. Med.	272	Guidelines	Test
T5 Research	PMC papers	163	Expert (30) + LLM-Judge (133)	Test

D GRAPH-RAG RETRIEVER DETAILS

D.1 PPR-PCST IMPLEMENTATION

The PPR-PCST retriever proceeds in three stages. First, named entity recognition identifies seed entities $\mathcal{S}$ in the query using the same biomedical NER model used in EPIKG construction. Second, Personalized PageRank Yasunaga et al. (2021) is run from $\mathcal{S}$ over the EPIKG adjacency matrix with restart probability $\alpha = 0.15$ and damping factor $1 - \alpha$, producing a relevance score $r(v)$ for each node v. Third, a Prize-Collecting Steiner Tree (PCST) Zhang et al. (2022) approximation extracts a connected subgraph by assigning prize $r(v)$ to each node and minimising edge costs, subject to a maximum node budget of 30 and maximum path depth of 4 hops. The extracted subgraph is serialised into structured reasoning paths by enumerating all source-to-sink paths and formatting each as `(head, relation[Np], tail)` where N is the paper count annotation.

D.2 SEMANTIC RETRIEVAL IMPLEMENTATION

The semantic retriever encodes the query using `all-MiniLM-L6-v2` Reimers & Gurevych (2019) and retrieves the top-k most similar EPIKG nodes by cosine similarity over pre-computed node embeddings. Node embeddings are constructed by encoding the concatenation of the entity name and its ontology definition. Local neighbourhoods (depth 1) of the top-k nodes are extracted and serialised in the same format as PPR-PCST paths. $k = 10$ is used in all experiments.

D.3 HYPERPARAMETER SETTINGS

Table 8: Graph-RAG retriever hyperparameters. Optimal values are selected based on T1 MCQ validation performance.

Hyperparameter	**Range tested**	**Selected value**
Max subgraph nodes	10, 20, 30, 40	30
Max path depth	2, 3, 4, 5	4
PPR restart prob. α	0.10, 0.15, 0.20	0.15
Semantic top-k	5, 10, 15	10
NER confidence threshold	0.7, 0.8, 0.9	0.8

E EVALUATION METRICS

Top-1 Accuracy. For MCQ tasks, the model prediction is the option letter with the highest generation probability or the option explicitly named in the generated response. Accuracy is the proportion of correct predictions over all instances.

ROUGE-L. Longest common subsequence F1 between the generated output and the reference answer Lin (2004).

BERTScore F1. Token-level semantic similarity between generated and reference text using contextual embeddings Zhang et al. (2019). We use `roberta-large` as the backbone model.

LLM-as-Judge. GPT-4.1-mini scores generated outputs on a 1–5 Likert scale across three dimensions: factual correctness, clinical relevance, and reasoning quality Chen et al. (2025a). The final score is the average across dimensions.

Clinical NER F1. Entity-level F1 measuring overlap between syndrome, drug, and finding mentions in generated outputs and reference texts, using the biomedical NER model from EPIKG construction.

Hallucination Rate. Proportion of generated sentences classified as contradicting the source document by a Natural Language Inference (NLI) model fine-tuned on biomedical text.

Guideline Concordance (GC). Binary indicator of whether the recommended treatment is consistent with ILAE 2022 and CPIC guidelines for the given syndrome and genetic context, averaged over all instances.

Drug Safety Score (DFS). Proportion of generated treatment recommendations that exclude every contraindicated ASM for the given patient context, as defined by CPIC and ILAE 2022 guidelines.

KG Evidence Coverage (KGEC). Proportion of EPIKG entities in the retrieved subgraph that appear in the generated output, measuring active utilisation of retrieved KG context Yu et al. (2024).

Alignment Score. Composite score measuring consistency between generated research plans and source papers, combining LLM-as-Judge ratings on clinical impression alignment, research question quality, and feasibility.

F EPIBENCH AND EPIKG DOCUMENTATION

F.1 DATASET DOCUMENTATION AND INTENDED USE

EPIKG and EPIBENCH are intended for academic research on epilepsy clinical AI, knowledge graph construction, and LLM evaluation. Their sources are accessed under source-specific licenses, credentialing requirements, and data-use terms. No restricted Harvard EEG record, derived clinical text, or neurologist-written impression is included in the public package. EPIKG will be released on Hugging Face at [https://anonymous.4open.science/r/EpiVerse-BF8E/] and EPIBENCH evaluation scripts will be released on GitHub at [https://anonymous.4open.science/r/EpiVerse-BF8E/]. The released package includes dataset summaries, a non-restricted metadata preview, and evaluation scripts.

Data organisation. The release is structured as follows:

- `epikg/nodes/`: entity files per layer (L1–L5) in JSON format, each record containing entity name, canonical identifier, ontology source, and layer label.
- `epikg/edges/`: triplet files per relation type in JSON format, each record containing head entity, relation, tail entity, paper count $\mathcal{P}$, source identifiers, source version, evidence strength, and extraction and validation status.
- `epikg/graph.pkl`: full EPIKG graph serialised as a NetworkX DiGraph with node and edge attributes.
- `epibench/t1/`: EpiBench-MCQ (1,000 questions) and EpiBench-QA (5,199 questions) in JSON format.
- `epibench/t2/`: access instructions, preprocessing and evaluation code, schemas, aggregate summaries, and source record identifiers. Authorized users reconstruct EEG descriptions, computed statistics, and gold-standard impressions only after obtaining access through the original credentialed process.
- `epibench/t3/`: 151 pharmacogenomic MCQs with CPIC/ILAE gold standards.
- `epibench/t4/`: epilepsy-filtered questions from MedQA-USMLE (200) and MMLU Professional Medicine (272), with guideline-aligned gold standard answers.
- `epibench/t5/`: 163 PMC paper instances with expert annotations (30) and LLM-as-Judge gold standards (133).
- `scripts/`: retriever implementation (PPR-PCST, Semantic, Hybrid), evaluation scripts for all five tasks, and prompt templates.

F.2 AUTHOR STATEMENT

We confirm that all data sources used in EPIKG and EPIBENCH are accessed and used under their respective licenses, access controls, and data-use terms. We bear full responsibility for ensuring compliance. All ontology sources (ILAE, MeSH, HPO, OMIM, HGNC, ChEBI, AES 2024) are credited with their original citations in the main paper. The Harvard EEG dataset is used within its credentialed environment under its stated academic use terms Zafar et al. (2025); Sun et al. (2025); restricted record-level content is not redistributed. MedQA-USMLE Jin et al. (2021) and MMLU Professional Medicine Hendrycks et al. (2020) are used under their respective open licenses. PMC papers used in T5 are accessed via the PubMed Central Open Access Subset under CC-BY or CC0 licenses.

F.3 HOSTING, LICENSING, AND MAINTENANCE

Hosting. EPIKG and EPIBENCH will be hosted on Hugging Face Datasets and GitHub respectively. Both platforms provide version control, allowing users to track changes and ensure reproducibility across experiments.

Licensing. Author-created documentation and eligible metadata are released under the **CC BY 4.0** license, and source code is released under its accompanying software license. Third-party ontologies, guidelines, articles, and credentialed clinical records retain their original licenses and

access restrictions and are not relicensed by this project. The release includes a field-level provenance and license manifest; non-redistributable content is omitted or represented only by source identifiers.

Maintenance. We commit to maintaining both resources for a minimum of three years following publication. Planned updates include: (1) annual EPIKG refresh incorporating new epilepsy literature and updated clinical guidelines; (2) expansion of EPIBENCH T3 with new CPIC guideline releases; and (3) community contributions via GitHub pull requests for additional task instances and evaluation metrics.

F.4 ACCESS AND REPRODUCIBILITY

Users can access EPIKG and load it directly via the Hugging Face Datasets library:

```
from datasets import load_dataset
epikg = load_dataset("[https://anonymous.4open.science/r/EpiVerse-BF8E/]/epikg",
                     split="full")
```

The EPIBENCH evaluation pipeline can be run as follows:

```
git clone [https://anonymous.4open.science/r/EpiVerse-BF8E/]/epibench
cd epibench
pip install -r requirements.txt
python evaluate.py --task t1_mcq \
    --model gpt-4o \
    --retriever hybrid \
    --kg_path data/epikg/graph.pkl
```

Full environment specifications and hyperparameter settings are provided in `configs/` and documented in Appendix D.3.

F.5 EPIKG STATISTICS

Table 9 provides a detailed breakdown of EPIKG entity and relation statistics, complementing the summary in Appendix B.4.

Table 9: Detailed EPIKG statistics. Cross-layer triplets cover all pairwise layer combinations; the densest connections are between L1 Syndrome and L4 Treatment (3,217 triplets) and between L3 Gene and L1 Syndrome (2,845 triplets).

Statistic	**Value**	**(%)**
Total nodes	24,324	100.0
Total edges	32,009	100.0
Rule-based triplets	9,670	30.2
LLM-based triplets	22,339	69.8
Cross-layer triplets	14,576	45.5
Within-layer triplets	17,433	54.5
Median paper count $\mathcal{P}$	3	—
IQR paper count	1–8	—
Triplets with $\mathcal{P} \geq 10$	4,612	14.4

G PROMPTS

System prompt (all tasks).

```
You are an expert epileptologist with deep knowledge
of epilepsy syndromes, antiseizure medications,
pharmacogenomics, EEG interpretation, and epilepsy research.
Answer the following question based on the provided clinical
context and knowledge graph evidence.  Think step by step
and ground your answer in the evidence provided.
```

T1 Clinical Decision Accuracy MCQ input prompt.

```
Context:  {retrieved_kg_paths}
Question:  {question}
Options:  A) {opt_a} B) {opt_b} C) {opt_c} D) {opt_d}
Answer with the option letter and a brief justification.
```

T1 Clinical Decision Accuracy Open-ended QA input prompt.

```
Context:  {retrieved_kg_paths}
Question:  {question}
Provide a detailed answer grounded in the provided evidence.
Cite specific entities or relations from the knowledge graph
context where relevant.
```

T2 Clinical Report Generation input prompt.

```
Context:  {retrieved_kg_paths}
Patient history:  {patient_history}
EEG description:  {eeg_description}
Band power:  {band_power_stats}
Spike rate:  {spike_rate_stats}
Generate a clinical impression for this EEG report.  Your
impression should include:  (1) identification of any
epileptiform activity or abnormal patterns, (2) syndrome
or diagnosis consistent with the findings, and (3) relevant
clinical recommendations.  Ground your impression in the
provided knowledge graph evidence linking EEG patterns to
syndromes and treatments.
```

T3 Biomarker-Driven Precision Medicine input prompt.

```
Context:  {retrieved_kg_paths}
Patient:  {genetic_variant}, {phenotype}
Select the most appropriate ASM from the following options
and justify your selection based on the genetic evidence and
clinical guidelines.
Options:  A) {opt_a} B) {opt_b} C) {opt_c} D) {opt_d}
```

T4 Treatment Recommendation input prompt.

```
Context:  {retrieved_kg_paths}
Clinical scenario:  {clinical_scenario}
Select the most appropriate treatment option from the
following choices.  Consider guideline concordance, drug
safety, and potential contraindications based on the
provided knowledge graph evidence.
Options:  A) {opt_a} B) {opt_b} C) {opt_c} D) {opt_d}
Answer with the option letter and justify your
selection with reference to clinical guidelines and any
contraindication evidence.
```

T5 Deep Research Planning input prompt.

```
Context: {retrieved_kg_paths}
Paper: {paper_abstract}
Generate a structured research plan covering: (1) a focused
research question, (2) a study design rationale, and (3)
required data sources. Ground your plan in the provided
knowledge graph evidence and identify at least one knowledge
gap beyond the paper's scope.
```

H Additional Experimental Results

MMLU Treatment Recommendation. Table 10 reports full results on the MMLU Professional Medicine subset, complementing the MedQA-USMLE results in the main text.

Table 10: T4 Treatment Recommendation: full MMLU Professional Medicine results. DFS: Drug Safety Score; GC: Guideline Concordance; KGEC: KG Evidence Coverage. Results are mean ± std; Δ: relative improvement (%).

Model	RAG	Acc	Δ	DFS	GC	KGEC
GPT-4o	–	74.0±1.8	–	0.35±0.03	70.0±2.0	0.66±0.02
	w/	**83.0±1.5**	+12.2	0.44±0.02	78.0±1.7	0.72±0.02
Claude S4	–	71.0±1.9	–	0.39±0.03	67.0±2.1	0.68±0.02
	w/	80.0±1.6	+12.7	**0.52±0.02**	**81.0±1.6**	**0.76±0.01**
Gemini	–	67.0±2.0	–	0.36±0.03	62.0±2.3	0.62±0.03
	w/	76.0±1.7	+13.4	0.43±0.02	74.0±1.9	0.71±0.02
Llama	–	56.0±2.3	–	0.30±0.04	53.0±2.5	0.63±0.03
	w/	68.0±1.9	+21.4	0.42±0.03	66.0±2.1	0.69±0.02
Qwen	–	54.0±2.4	–	0.31±0.04	51.0±2.6	0.61±0.03
	w/	65.0±2.0	+20.4	0.39±0.03	63.0±2.2	0.66±0.02
Mistral	–	50.0±2.6	–	0.29±0.04	47.0±2.8	0.58±0.03
	w/	61.0±2.1	+22.0	0.36±0.03	59.0±2.3	0.64±0.02
Avg. Δ		**+17.2**		**+28.1**	**+28.4**	**+13.2**

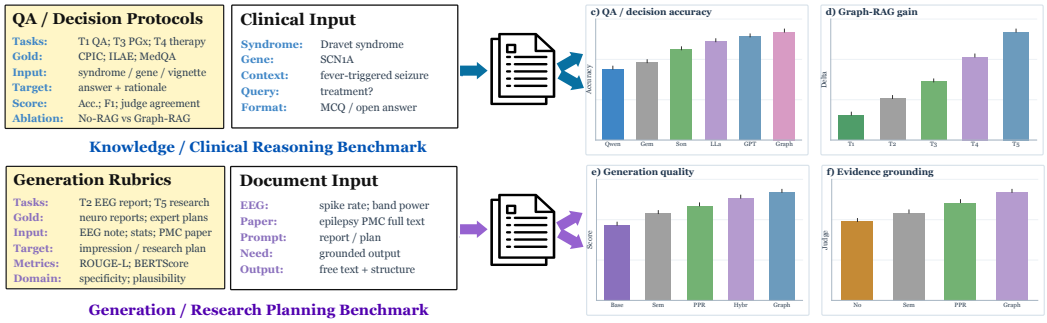


Figure 7: Overview of the **EPIBENCH**.

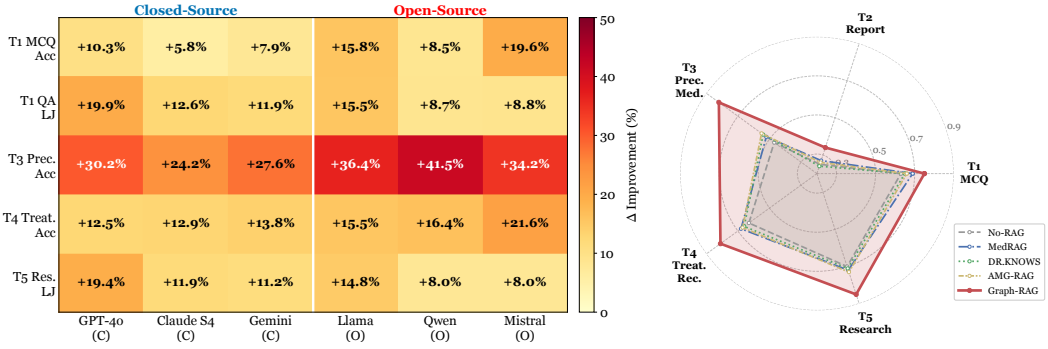


Figure 8: Overview of **EPIBENCH** Result.

I SCOPE AND FUTURE WORK

Language and ontology coverage. EPIKG is constructed from English-language ontologies and literature, limiting its coverage of epilepsy syndromes and gene–disease associations that are primarily documented in non-English sources. Rare syndromes with fewer than five supporting papers receive lower representation in the extracted relation set, as the LLM-based extraction pipeline requires sufficient co-occurrence evidence to produce reliable triplets. Extending EPIKG to multilingual sources and rare disease registries is a natural direction for future work.

LLM-as-Judge reliability. For T5 research planning, LLM-as-Judge produces 133 of 163 gold standards; human experts provide the complementary subset. The judge prompt is designed to minimise positional bias and hallucination, while fluent-output bias remains a reliability risk. Expanding the expert-annotated subset and measuring inter-rater reliability between human and LLM judges is an important direction for strengthening the T5 evaluation.

J NOTATION AND EXPLANATION

Graph definition and accounting. We use $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \phi)$, where $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$ and ϕ stores evidence spans, source identifiers, and dates. Five task-facing clinical layers (Syndrome, Diagnostic, Gene, Treatment, Outcome) are distinguished from three auxiliary categories (Protein, Anatomy, Other). The 1,370 normalized predicates are grouped into six semantic families. The seven source files contain 16,325 raw seed rows; cross-source merging removes 1,413 duplicates, leaving 14,912 unique seeds. Adding 9,412 validated literature-discovered entities yields 24,324 final entities (11,683 clinical and 12,641 auxiliary). UMLS is used for alignment and is omitted from the seven-source count.

Expert graph audit. Two clinically trained annotators reviewed 800 stratified triplets. A triplet was counted as correct when its entities, predicate, direction, qualifiers, and complete claim were supported by the cited span. Of 800, 709 were correct (88.6%; bootstrap 95% CI 86.2–90.7; Cohen's $\kappa = 0.84$). The 91 errors comprise wrong predicate or overstated causality (34), normalization (23), direction or strength (19), and insufficient support (15). Exhaustive annotation of 120 held-out passages gives recall 812/1,042=77.9% (95% CI 75.3–80.4).

Medical-label audit. All 180 draft T3 items were reviewed: 21 lost qualifiers and were corrected; 29 ambiguous or unsupported items were removed or replaced, leaving 151. A *TSC2* variant alone carries no carbamazepine contraindication edge. Everolimus is represented as an adjunctive mTOR option for eligible tuberous-sclerosis-associated seizures. *SCN1A* edges distinguish association, variant-specific causality, and syndrome-level drug cautions.

Split isolation. T1 uses a 200-item source-disjoint development set and an 800-item post-snapshot test set; retriever settings are frozen after development. This prevents direct graph/test overlap; model-pretraining exposure remains outside experimental control. T2 is split by patient and time; gold impressions, encounter duplicates, and near-duplicate reports are removed from retrieval. T5 removes the query paper and its direct evidence under a task-specific isolation rule; T1 uses temporal isolation.

Matched-context controls. All T3 controls use the same source snapshot, prompt, decoder, and 2,048-token budget. Accuracies for Base/Flat/Graph/Oracle/No-guideline/Shuffled are: GPT-4o 53.0/60.3/68.9/79.5/64.8/65.2; Claude S4 66.2/74.8/82.1/89.2/78.4/79.0; Gemini 58.3/67.5/74.2/84.0/70.1/70.8; Llama 43.7/52.3/60.3/72.0/56.2/57.1; Qwen 41.1/50.3/58.3/70.5/54.0/55.0; and Mistral 37.7/45.0/51.0/65.8/47.6/48.5. Flat evidence helps every model, while typed structure adds 6.0–9.0 points. Edge shuffling and guideline removal erase much of that increment; the oracle exposes remaining retrieval headroom.

Uncertainty and evaluation. Primary accuracy comparisons use paired bootstrap intervals and two-sided McNemar tests with Holm correction. Temperature-zero MCQ uncertainty is item-bootstrap variability; generation rows use run SD over three temperature-0.3 seeds. Experts reviewed 200 T1 answers ($\kappa = 0.76$), all 151 T3 labels ($\kappa = 0.88$), 120 T4 answers ($\kappa = 0.81$), and 30 T5 plans ($\kappa = 0.72$). The remaining 133 T5 cases use LLM-as-Judge evaluations; expert gold labels cover the expert-reviewed subset.

Metric interpretation and intended use. KG Evidence Coverage measures whether retrieved entities appear in an output; relation and conclusion correctness require separate assessment. The NLI hallucination rate captures contradictory claims, while unsupported non-contradictory claims require complementary review. Drug Safety Score and Guideline Concordance serve as benchmark metrics; safety certification requires dedicated clinical validation. T2 accepts report descriptions, metadata, and summary statistics; waveform analysis lies outside its input scope. EPIGRAPH is intended for research and evidence navigation. Clinical diagnosis and medication selection require source verification and qualified clinician judgment.

K DISCUSSION

Why an epilepsy-specific reasoning framework is necessary. Epilepsy reasoning joins syndrome classification, EEG interpretation, genetic mechanisms, pharmacogenomics, treatment guidance, contraindications, and outcomes in a single decision process. The evidence for these decisions is distributed across literature, ontologies, guidelines, and clinical records. The experimental results show that model scale alone leaves substantial gaps across this chain. Graph-free performance places T3 accuracy at 38–66%, T1 MCQ accuracy at 51–69%, and EEG impression generation at METEOR values of 0.23–0.29. These gaps occur across proprietary and open-source models, including systems with strong general medical capabilities. EpiGraph addresses the underlying evidence-integration problem by coupling a domain-specific graph with a benchmark that measures five forms of reasoning over the same clinical structure. Evaluation centered on recall omits EEG synthesis, gene–drug reasoning, contraindication handling, and research-plan construction. A traceable evidence layer connects each gain to its supporting clinical structure and makes the benchmark explanatory.

Scale, structure, and evidence quality jointly define the resource. EpiKG converts 48,166 papers and seven clinical resources into 24,324 entities and 32,009 triplets. Its 14,576 cross-layer triplets span 45.5% of the graph and encode the gene–syndrome–treatment–outcome paths required by the benchmark. The graph combines 9,670 rule-derived triplets with 22,339 literature-derived triplets, while paper counts preserve evidence frequency: the median support is three papers, and 4,612 triplets have support from at least ten publications. The Appendix audit adds a direct quality measurement to these scale statistics. Two clinically trained annotators judged 709 of 800 stratified triplets correct, yielding 88.6% precision, a 95% bootstrap interval of 86.2–90.7, and Cohen's $\kappa = 0.84$; exhaustive annotation of 120 held-out passages yielded 77.9% recall. The error decomposition further identifies concrete improvement targets in predicate choice, causal strength, normalization, direction, and evidential support. The medical-label audit corrected 21 qualifier errors and removed or replaced 29 ambiguous items from 180 draft T3 questions, producing the final set of 151. These measurements turn graph quality into an auditable research object with explicit error categories and version-level targets.

The five tasks reveal complementary forms of knowledge dependence. On T1, Graph-RAG improves MCQ accuracy by 11.3 percentage points on average across six models. Reasoning accuracy rises by 31.9%, compared with 14.6% for ROUGE-L and 7.6% for BERTScore, showing that retrieved paths contribute most strongly to the clinical reasoning chain. GPT-4o reaches 4.33/5.0 on open-ended QA, and the comparison with MedRAG at 62% and AMG-RAG at 66.3% links the advantage to epilepsy-specific curation and relational retrieval. The larger gains for Mistral and Llama further show how external structure compensates for weaker parametric coverage.

T2 exposes a different regime. Graph augmentation improves all four locally deployed report-generation models, with MedGemma rising from 0.26 ± 0.13 to 0.34 ± 0.17 METEOR, a 30.8% relative gain. The low absolute scores across every model preserve the difficulty of neurologist-style EEG synthesis even after retrieval. The qualitative examples connect the metric gain to more precise waveform descriptions, abnormal slowing patterns, and clinically aligned impressions. This result gives future report-generation work a grounded baseline and a clear target: combine EEG-specific representation learning with evidence paths while retaining clinician-authored evaluation inside the credentialed environment.

T3 delivers the largest gains because pharmacogenomic decisions depend directly on multi-hop relations. Claude Sonnet 4 reaches 82% accuracy, while Qwen and Llama improve by 42% and 36%. The no-KG Mistral result of 38% lies close to the 25% four-option baseline, illustrating the scarcity of gene–drug knowledge in general model parameters. Matched-context controls sharpen this interpretation: flat evidence improves every model, typed graph structure adds another 6.0–9.0 points, and guideline removal or edge shuffling erases much of that increment. The oracle results, ranging from 65.8% to 89.2%, quantify the remaining retrieval headroom. The gain arises from selecting and preserving clinically meaningful paths, with the TSC2/mTOR case study showing how graph reranking promotes the two supporting papers above keyword-matched distractors.

T4 connects answer accuracy with clinically relevant behavior. On MedQA-USMLE, Graph-RAG raises accuracy by 15.6%, Drug Safety Score by 12.4%, and Guideline Concordance by 14.1%. The Appendix MMLU results strengthen this pattern: average gains reach 17.2% for accuracy, 28.1% for Drug Safety Score, 28.4% for Guideline Concordance, and 13.2% for KG Evidence Coverage. Safety

and guideline gains exceed the accuracy gain, showing that a correct option captures one component of treatment reasoning. The paired growth of evidence coverage and contraindication avoidance links retrieved clinical structure to safer benchmark behavior and motivates future evaluation that reports accuracy, evidence use, and safety together.

T5 extends the same graph beyond immediate clinical questions into scientific planning. Graph-RAG improves LLM-as-Judge scores by 12.2% on average; GPT-4o reaches 4.25/5.0 with a 19.4% gain, and Llama reaches 3.87, leaving a gap of 0.38. This convergence shows that structured evidence paths help models formulate research questions, connect study rationales to prior evidence, and identify knowledge gaps. The 30 expert-annotated papers and 133 LLM-judged papers define two transparent evaluation strata. The expert subset anchors interpretation, and the scalable stratum supports broader model comparison. Expanding expert coverage and inter-rater analysis will convert this task into an increasingly strong test bed for evidence-grounded scientific reasoning.

Retriever analyses explain the source of the gains. PPR-PCST exceeds semantic retrieval by 2.6–7.7 percentage points on T1 and 7.7–8.4 points on T3, while the hybrid retriever adds another 1.2–2.4 points. Sensitivity analysis selects 30 nodes and four hops; on T3, increasing depth from two to four hops raises GPT-4o accuracy from 58.7% to 69.0%. These results connect performance directly to controlled subgraph size and path depth. They also explain why precision medicine gains exceed gains on simpler recall tasks: gene–syndrome–pathway–treatment reasoning requires several typed transitions. The running-time study gives Graph-RAG the best average rank, 1.5, across the five task forms, pairing the accuracy gains with competitive inference efficiency. Together, the ablations, matched-context controls, edge perturbations, oracle retrieval, and runtime comparisons provide a mechanistic account of when graph retrieval helps and where engineering effort yields the largest return.

The Appendix converts the benchmark into a reproducible experimental foundation. The construction protocol records ontology sources, normalization stages, relation templates, evidence counts, and graph statistics. The benchmark protocol documents 1,000 T1 MCQs, 5,199 T1 open-ended questions, more than 40,000 credentialed T2 instances, 151 T3 questions, 200 MedQA and 272 MMLU T4 questions, and 163 T5 papers. Temporal, patient-level, and source-level isolation rules protect the evaluation splits. Human review covers 200 T1 answers, all 151 T3 labels, 120 T4 answers, and 30 T5 plans, with reported agreement values. The prompt suite, metric definitions, hyperparameters, statistical tests, provenance fields, access rules, and licensing manifest provide the components needed to reproduce the experiments and extend them under consistent controls. This level of documentation supports direct comparison across new models, retrievers, graph versions, and evaluation metrics.

A concrete foundation for future research. EpiGraph supports several research programs from a shared starting point. Multilingual literature and regional guidelines can expand geographic coverage; rare-disease registries can strengthen low-frequency syndromes; additional sites can test EEG generalization; larger expert panels can extend T5 gold standards and subgroup evaluation. The graph audit supplies error categories for targeted extraction models, calibrated edge confidence, and clinician-prioritized review. The oracle gap supplies a measurable objective for retriever learning, query decomposition, and adaptive path expansion. The cross-task results support retrieval-augmented fine-tuning, memory systems, tool-using agents, and continual guideline updates, with each method evaluated across recall, generation, precision medicine, safety, and research planning. Versioned evidence, correction channels, and source-specific access controls preserve traceability as these extensions grow.

The combined evidence supports a central insight: progress in epilepsy language models depends on the organization, retrieval, and evaluation of clinical knowledge as much as on model capacity. EpiKG supplies the structured evidence layer, EpiBench exposes task-specific capability gaps, and the audit and ablation framework identifies the mechanisms behind improvement. This integrated design gives future work a solid empirical foundation, a set of clinically meaningful targets, and a path from new evidence to measurable reasoning gains.