KG-QUEST: KNOWLEDGE GRAPH–ENHANCED QUESTION ANSWERING AND REASONING IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) achieve strong results on medical and open-domain QA but remain limited by static retrieval and parametric memory, hindering adaptation to evolving ontologies and multi-hop reasoning. We present KG-QUEST, a framework for Knowledge Graph–Enhanced QA that grounds questions in Entity–Attribute–Value (EAV) and Entity–Relation–Entity (ERE) triples and dynamically discovers an answer-specific knowledge graph during inference. A query graph is softly matched to a global biomedical KG and expanded via hop-limited frontier search with predicate weights, synonym/inverse alignment, and negation-aware pruning to form a minimal, high-support subgraph. Phase I (KG generation) fine-tunes LLaMA 3.1 (8B) with ensemble refinement to produce ontology-aligned triples; answers are then selected by dual grounding—scoring KG paths (and optional text evidence) with hop decay and abstention—yielding explicit evidence chains. On *MedQA (USMLE)* and *MMLU medical subsets*, KG-QUEST establishes new state-of-the-art results (*93.7%* and *92.0%* accuracy, respectively), surpassing GPT-4 and Med-PaLM 2 while maintaining verifiability. Beyond medical QA, KG-QUEST shows how LLMs can not only retrieve but also construct and navigate structured knowledge graphs for complex reasoning.

1 INTRODUCTION

Large Language Models (LLMs) have achieved strong results on open-domain and medical QA, yet they remain constrained by static parametric memory and brittle retrieval-augmented generation (RAG) pipelines. In high-stakes healthcare settings, precise alignment with evolving ontologies (e.g., FHIR, UMLS, SNOMED) and sensitivity to patient context are essential; current systems still suffer from hallucinations, stale knowledge, and limited ability to *adapt representations during inference*. Hybrid KG–text approaches such as GraphRAG (Edge et al., 2024), Chain-of-Knowledge (Li et al., 2024), and ToG-2 (Ma et al., 2025) highlight the promise of coupling retrieval with knowledge graphs (KGs), but often rely on prompt heuristics, Wiki-centric coverage, or fragile entity linking. In healthcare-specific pipelines, KARE (Jiang et al., 2025) shows that EHR-anchored KG construction with community retrieval improves clinical prediction on MIMIC (Johnson et al., 2016; 2023), while KG-SFT (Chen et al., 2025) boosts multilingual medical QA by injecting KG-grounded explanations into supervised fine-tuning—at the cost of curated resources and training overhead. Meanwhile, interactive benchmarks like AGENTCLINIC (Schmidgall et al., 2024) move beyond static exam-style QA (Jin et al., 2019; 2021; Singhal et al., 2023a; Hendrycks et al., 2020) toward sequential, multimodal, and bias-sensitive evaluation, underscoring the need for *dynamic* reasoning under incomplete information.

We introduce *KG-QUEST* (*Knowledge Graph–Enhanced Question Answering and Reasoning in LLMs*), a framework that integrates LLMs with an *answer-specific, dynamically discovered knowledge graph* built on-the-fly from the question. The key idea is to represent questions with *Entity–Attribute–Value (EAV)* and *Entity–Relation–Entity (ERE)* triples, aligning natural language to ontology-grounded structure. From these triples we construct a *query graph* that is softly matched to a global biomedical KG; we then perform hop-limited frontier expansion with predicate-specific weights, synonym/inverse handling, and negation-aware pruning to discover the minimal, high-support subgraph sufficient to answer the question. This subgraph is scored with *dual grounding* (KG paths + optional text evidence), uncertainty-aware hop decay, and an abstention rule, producing answers accompanied by explicit evidence chains.

Phase I (KG generation) uses a fine-tuned LLaMA 3.1 (8B) model as the triple extractor. We fine-tune it to produce ontology-aligned EAV/ERE triples, canonical entity mentions (synonyms/inverses), and negation-aware predicates. Multiple extraction variants are combined via ensemble refinement to obtain calibrated reliabilities, which seed the dynamic discovery process.

Concretely, KG-QUEST comprises: (i) an *ensemble-refined seed* produced by the fine-tuned LLaMA 3.1 (8B) extractor and a triple-scoring model to yield calibrated EAV/ERE edges aligned to the ontology; (ii) a *query-graph–guided dynamic discovery* stage that matches nodes/relations, expands neighborhoods with hop decay and boundary detection, and constructs a compact answer-specific KG; and (iii) an *answer selection* module that aggregates path evidence with dual (KG+text) grounding and returns both the choice and a verbalized justification. This design departs from static clustering and prompt-only reasoning: the *question itself* becomes a structural constraint that shapes the evidence graph used for answering.

Contributions. (1) A *query-graph–guided dynamic KG* procedure that discovers, scores, and *bounds* an answer-specific subgraph via hop-limited frontier expansion, predicate weighting, synonym/inverse alignment, and negation-aware pruning. (2) A *dual-grounded scoring* objective that composes ER-weighted EAV/ERE paths and optionally fuses textual evidence, with uncertainty-aware hop decay and abstention for safety. (3) *Empirical gains* on standard QA benchmarks *(e.g., MedQA, PubMedQA, MMLU)* (Jin et al., 2021; 2019; Hendrycks et al., 2020), with transparent evidence chains and improved robustness under incomplete graphs and noisy linking.

By unifying ontology-grounded representation with *dynamic* KG discovery and dual grounding—and by fine-tuning LLaMA 3.1 (8B) explicitly for Phase I KG generation—KG-QUEST advances a paradigm in which LLMs not only retrieve facts but also *construct and navigate* the structured knowledge needed for clinically meaningful reasoning.

2 Related Work

Static KGE vs. LLM-driven KG discovery. Classical KGE models—TransE, DistMult, ComplEx, RotatE—learn fixed embeddings and triple scorers on a *static*, closed-world schema Bordes et al. (2013); Yang et al. (2014); Trouillon et al. (2016); Sun et al. (2019). They capture algebraic relation patterns efficiently, but struggle with *open-world* additions (unseen entities/relations) and offer limited evidence traceability. In contrast, LLM-driven extraction discovers *new* EAV/ERE facts from unstructured corpora and aligns them to ontologies, enabling schema-flexible KG growth with natural-language rationales Pan et al. (2024); Chen et al. (2025). *KG-QUEST* bridges these lines by combining ontology-aligned triples with *query-graph–guided, dynamic* graph discovery that operates in an open-world setting while preserving verifiable evidence paths.

LLMs for medical QA. Foundation LLMs (e.g., GPT-4, PaLM/Flan-PaLM) achieve strong results on MedQA/PubMedQA, especially with instruction tuning and safety-aware decoding Chowdhery et al. (2023); Achiam et al. (2023). Med-PaLM and Med-PaLM 2 further improve reasoning via ensemble refinement and chain-of-retrieval (CoR) Singhal et al. (2023b; 2025). *KG-QUEST* is complementary: instead of relying solely on generator-side advances, it *structures the reasoning substrate* as an answer-specific KG discovered from the question, enabling explicit multi-hop paths, uncertainty handling (hop decay, negation), and abstention.

Hybrid KG–text retrieval. Graph-augmented RAG links corpus elements into graphs for multi-hop reasoning (e.g., GraphRAG, Chain-of-Knowledge, ToG-2) Edge et al. (2024); Li et al. (2024); Ma et al. (2025), and Self-RAG learns *when* to retrieve Asai et al. (2024). These systems interleave retrieval with graph exploration but often lack a tight mechanism to *constrain* graph growth to the question. *KG-QUEST* performs *query-graph matching with hop-limited frontier expansion* and explicit *boundary discovery* of a minimal, answer-sufficient subgraph, then scores it with dual grounding (KG+text), improving precision under noisy linking and incomplete KGs.

Healthcare KGs for reasoning. Clinical systems such as KARE assemble UMLS- and PubMed-based subgraphs and score patient-centric communities Jiang et al. (2025); Bodenreider (2004); Canese & Weis (2013); Johnson et al. (2016; 2023). While effective, they rely on curated resources and largely *static* communities. *KG-QUEST* instead builds a *question-specific* dynamic KG via query alignment and neighborhood discovery, emphasizing just-enough structure for the current question and producing auditable evidence chains.

KG-enhanced fine-tuning. KG-SFT augments supervised fine-tuning data with KG-grounded explanations, improving multilingual medical QA Chen et al. (2025). Unlike this *training-time* augmentation, *KG-QUEST* focuses on *inference-time* dynamic KG construction and scoring, reducing dependence on curated expansions while retaining verifiability and controllable uncertainty.

Positioning. In summary, *KG-QUEST* unifies ontology-grounded EAV/ERE representation with *query-guided, dynamic* KG discovery and dual grounding. This departs from static KGE and complements retrieval-augmented LLMs by using the *question as a structural constraint* that shapes the evidence graph, yielding precise multi-hop reasoning with explicit uncertainty and abstention.

3 METHOD

We introduce *KG-QUEST*, a framework for medical QA that extends retrieval-augmented generation (RAG) with structured reasoning over *Entity–Attribute–Value (EAV)* and *Entity–Relation–Entity (ERE)* triples. KG-QUEST integrates LLMs with an *answer-specific, dynamically discovered knowledge graph*, built on-the-fly by matching a query graph to a global biomedical KG and expanding through scored neighborhoods. Unlike prior pipelines relying on static clustering or fixed retrieval, KG-QUEST explicitly *discovers the boundary* of a minimal, high-support subgraph sufficient to answer the question. The framework is formalized in Algorithm 1.

3.1 REPRESENTING QUESTIONS AS EAV/ERE TRIPLES AND A QUERY GRAPH

Each natural-language question q is mapped to triples

$$q \;\mapsto\; \mathcal{T}(q) = \{t_i\}_{i=1}^{N}, \quad t_i \in \{\langle e,a,v\rangle,\ \langle e,r,e'\rangle\}. \tag{1}$$

EAV encodes descriptive properties (e.g., *Patient–Symptom–Chest pain*); *ERE* captures relational links (e.g., *Chest pain–indicates–MI*). We then form a *query graph* $\mathcal{Q} = (\mathcal{V}_q, \mathcal{R}_q, \mathcal{E}_q)$ by wiring the triples from $\mathcal{T}(q)$; nodes/edges inherit lexical labels, ontology hints, and type constraints (e.g., UMLS/SNOMED).

3.2 PHASE I: SEED GRAPH VIA ENSEMBLE REFINEMENT (ER)

We obtain m diverse triple sets $\{\mathcal{T}^{(j)}\}_{j=1}^{m}$ using prompt/seed/tool variants, aggregate to $\mathcal{T}_{\text{pool}}$, and estimate a calibrated consensus score $\tilde{s}_{\text{ER}}(t)$ for each triple. Entities/relations are aligned to ontology concepts, yielding a seed subgraph

$$\mathcal{G}_0 \;=\; \text{KGG}(\mathcal{T}_{\text{pool}}) \;\cup\; \text{OntologyAlign}(\mathcal{T}_{\text{pool}}, \mathcal{O}). \tag{2}$$

A triple-scoring model f_θ provides $s_{\text{KG}}(t)$; we fuse with ER via $s^\star(t) = \lambda \tilde{s}_{\text{ER}}(t) + (1-\lambda) s_{\text{KG}}(t)$ and prune under threshold τ to obtain $\mathcal{G}_1$.

3.3 PHASE II: DYNAMIC KG DISCOVERY VIA QUERY-GRAPH MATCHING

Node/edge matching. We create soft alignments from $\mathcal{Q}$ to the global KG $\mathcal{G} = (\mathcal{V}, \mathcal{R}, \mathcal{E})$:

$$\phi(v_q) \in \mathcal{V}, \qquad \psi(r_q) \in \mathcal{R},$$

selected by a match score

$$M(v_q \!\rightarrow\! u) \;=\; \underbrace{\text{lex}(v_q, u)}_{\text{string/synonym}} + \underbrace{\text{sem}(v_q, u)}_{\text{type/embedding}} + \underbrace{\text{ont}(v_q, u)}_{\text{ontology prior}},$$

and analogously for $M(r_q \!\rightarrow\! r)$, with inverse-relations allowed. Synonym expansion Σ (e.g., *palatine process*↔*palatine shelf*) prevents surface-form mismatch.

Query-guided frontier expansion. Starting from the matched anchors $\Phi_0 = \{\phi(v_q)\}$, we maintain a frontier $\mathcal{F}_t$ and a discovered subgraph $\mathcal{G}_t$ (initialized with $\mathcal{G}_1$ and anchor nodes). At step t we pop the highest-priority frontier node u and expand to neighbors $u' \in \mathcal{N}(u)$ if they improve alignment to some pending query edge $(v_q \xrightarrow{r_q} v'_q) \in \mathcal{E}_q$. Each candidate edge $(u \xrightarrow{r} u')$ receives

$$\text{EdgeScore} \;=\; w(r) \cdot \alpha^{h(u)} \cdot M(v_q \to u) \cdot M(r_q \to r) \cdot M(v'_q \to u') \cdot \mathbb{I}_{\text{no-neg}}, \tag{3}$$

where $h(u)$ is hop distance from anchors, $w(r)$ encodes predicate specificity (`causes` > `associated_with` > `found_in`), $\alpha \in (0, 1]$ applies hop decay, and $\mathbb{I}_{\text{no-neg}}$ nulls edges hitting negated predicates (e.g., `not_fusion_time`). Approved edges/nodes are added to $\mathcal{G}_t$ and new neighbors join the priority queue.

Boundary discovery and stopping. We define the boundary $\partial\mathcal{G}_t$ as active frontier nodes. Expansion stops when any of the following holds:

$$\text{(i) } \Delta U_t < \varepsilon, \quad \text{(ii) } \max\text{-hop} \geq H_{\max}, \quad \text{(iii) all query edges are matched above } \tau_m, \tag{4}$$

where ΔU_t is marginal utility (increase in total matched-query support), $H_{\max} \leq 3$, and τ_m is a match threshold. The result is a compact, answer-specific dynamic graph $\widehat{\mathcal{G}} = \mathcal{G}_t$ and a mapping between $\mathcal{Q}$ and $\widehat{\mathcal{G}}$.

Optional retrieval check. When the KG lacks direct support, we issue structured lookups from missing query edges to a corpus and extract candidate triples for verification. Retrieved triples are fused using a combined score:

$$\text{TotalScore} \;=\; \lambda_{\text{kg}} \cdot \text{EdgeScore} + \lambda_{\text{text}} \cdot s_{\text{text}}, \tag{5}$$

where s_{text} is the textual retrieval score and $\lambda_{\text{kg}}, \lambda_{\text{text}}$ balance KG-based and text-based evidence. A small retrieval budget ensures efficiency.

3.4 Phase III: Answer Selection over the Discovered Dynamic KG

Let $\mathcal{O} = \{o_1, \ldots, o_4\}$ be the answer options. For each o_k, we treat its node(s) (plus synonyms) as targets in $\widehat{\mathcal{G}}$ and score all paths connecting anchors to o_k. For a path $P = (v_0 \xrightarrow{r_1} \ldots \xrightarrow{r_L} v_L)$ inside $\widehat{\mathcal{G}}$,

$$\text{score}(P) = \Big(\prod_{i=1}^{L} w(r_i)\Big) \cdot \alpha^L \cdot \text{sim}(q, P) \cdot \mathbb{I}_{\text{no-neg}}(P), \tag{6}$$

and the option score is $S(o_k) = \sum_{P \in \mathcal{P}(o_k)} \text{score}(P)$. We add a small clinical-context term $R_{\text{ctx}}(o_k)$ (templates for ethics/statistics/trial design) and compute

$$S_{\text{total}}(o_k) = \lambda S(o_k) + (1 - \lambda) R_{\text{ctx}}(o_k), \qquad \hat{o} = \arg\max_{o_k} S_{\text{total}}(o_k), \tag{7}$$

with abstention if $\max_k S_{\text{total}}(o_k) < \tau$. KG-QUEST returns $\hat{o}$ and a verbalized justification from the highest-scoring path(s).

Why dynamic discovery? Replacing static clustering with query-graph–guided expansion (i) concentrates computation on question-relevant neighborhoods, (ii) discovers just-enough structure to satisfy all query edges, and (iii) produces a smaller, auditable subgraph that transfers better across unseen layouts.

3.5 End-to-End Loop

$$q \Rightarrow \mathcal{T}_{\text{pool}} \Rightarrow \mathcal{G}_0 \Rightarrow \mathcal{G}_1 \Rightarrow \text{Query-Graph Matching} \Rightarrow \widehat{\mathcal{G}} \Rightarrow \hat{o}. \tag{8}$$

Complexity. If b is the average branching factor and $H_{\max} \leq 3$, expansion is $O(b^{H_{\max}})$ per anchor (small in medical KGs). Scoring is linear in retrieved paths. Synonym expansion and negation pruning reduce the effective frontier.

Algorithm 1 KG-QUEST with Query-Graph Matching and Dynamic Boundary Discovery

```
Require: Question $q$, LLM $\mathcal{L}$, ontology $\mathcal{O}$, global KG $\mathcal{G}$, triple scorer $f_\theta$, hop decay $\alpha$, thresholds $(\tau, \tau_m)$, budgets $(H_{\max}, B)$
Ensure: Answer $\hat{o}$, discovered dynamic KG $\widehat{\mathcal{G}}$, evidence $\mathcal{E}$
Phase I: ER seed. Extract $\{\mathcal{T}^{(j)}\}$; aggregate $\mathcal{T}_{\text{pool}}$ with $\tilde{s}_{\text{ER}}$. Build $\mathcal{G}_0$, fuse with $s_{\text{KG}}$ to get $\mathcal{G}_1$.
Build query graph $\mathcal{Q}$ from $\mathcal{T}(q)$ with synonym set $\Sigma$ and predicate map $\mathcal{M}_{\text{pred}}$.
Phase II: Dynamic discovery.
Compute anchor matches $\Phi_0$ and initialize frontier $\mathcal{F}_0$; set $\mathcal{G}_0 \subseteq \mathcal{G}_1$ as seed of $\mathcal{G}_t$.
while stopping criteria not met do
    Pop $u \in \mathcal{F}_t$; for neighbors $u' \in \mathcal{N}(u)$ compute Eq. (3); add approved edges to $\mathcal{G}_t$; push new frontier nodes.
    Optionally retrieve & verify missing query edges with budget $B$; integrate verified triples.
end while
$\widehat{\mathcal{G}} \leftarrow \mathcal{G}_t$.
Phase III: Answering. Score options via Eqs. (6)–(7); return $\hat{o}$ and justification.
```

4 Representation Learning

At the core of *KG-QUEST* is a multi-view learner that jointly embeds: (i) ontology-aligned EAV/ERE triples, (ii) question semantics, and (iii) evidence signals from *dual grounding* (KG structure + text). Unlike static pipelines, representation learning in KG-QUEST is tightly coupled to the *dynamic knowledge-graph discovery* process in Sec. 3.3: a query graph derived from the question guides node/edge matching, hop-limited neighborhood expansion, and boundary detection of an *answer-specific* subgraph. We detail (a) ER-weighted triple embeddings; (b) path composition and question alignment with dual grounding on the discovered subgraph; and (c) how *query-graph matching + boundary discovery* act as representation-learning operators that replace static clustering.

4.1 Embedding EAV/ERE Triples with Ensemble Refinement

Given m extraction variants, we pool triples $\mathcal{T}_{\text{pool}} = \bigcup_{j=1}^{m} \mathcal{T}^{(j)}$, compute an ensemble-refined consensus $\tilde{s}_{\text{ER}}(t) \in [0,1]$ (majority/Borda/learned combiner), and a KG model score $s_{\text{KG}}(t) = f_\theta(h, r, u)$. We then define a calibrated reliability

$$w(t) = \sigma\big(\lambda\, \tilde{s}_{\text{ER}}(t) + (1-\lambda)\, s_{\text{KG}}(t)\big), \quad \lambda \in [0,1],$$

with σ logistic.

For **EAV** triples $t{=}\langle e, a, v\rangle$ and **ERE** triples $t{=}\langle e, r, e'\rangle$, we embed and calibrate as

$$t_{\text{EAV}} = f_\phi([e; a; v]), \qquad t_{\text{ERE}} = g_\phi(e, r, e'), \qquad \widetilde{t} = w(t) \cdot t,$$

where $\widetilde{t}$ (ER-calibrated) serves as the primitive for path and subgraph composition.

4.2 Question–Graph Alignment over the Discovered Dynamic KG

Let $\widehat{\mathcal{G}}$ denote the *dynamic* subgraph produced by query-graph matching and frontier expansion (Sec. 3.3). We encode the question q as $q \in \mathbb{R}^d$. Consider a path $p = (t_1, \ldots, t_L)$ in $\widehat{\mathcal{G}}$ with ER-calibrated triples $\{\widetilde{t}_\ell\}$. We compose a question-aware path embedding with attention that also injects *query-match priors* from the discovery stage:

$$\alpha_\ell \;\propto\; \underbrace{w(t_\ell)}_{\text{ER reliability}} \; \underbrace{m(t_\ell)}_{\text{query-match prior}} \; \exp\big(\eta \, \cos(q, \widetilde{t}_\ell)\big), \qquad p = \sum_{\ell=1}^{L} \alpha_\ell \, \widetilde{t}_\ell. \tag{9}$$

Here $m(t)$ summarizes the structural priors used during expansion:

$$m(t) = w(r) \cdot \alpha^{\text{hop}(t)} \cdot \underbrace{M_{\text{node}}(t)\, M_{\text{edge}}(t)}_{\text{query-graph alignment}},$$

where $w(r)$ is a predicate specificity weight (`causes> associated_with > found_in`), $\alpha \in (0,1]$ is hop decay, and $M_{\text{node}}, M_{\text{edge}}$ are the soft node/edge match scores from Sec. 3.3.

Dual grounding (KG + text). If optional text evidence $\mathcal{E}(p)$ was fetched for missing query edges, we encode it as $e(p)$ and score

$$G(q,p) = g_{\text{KG}}(q,p) + \beta\, g_{\text{text}}(q, e(p)) - \gamma\, u(p), \tag{10}$$

where $u(p)$ is an uncertainty penalty (e.g., attention entropy or ER variance), $\beta, \gamma \geq 0$. Answer selection in Phase III maximizes a sum of such path scores that reach each option.

Training objectives (optional). With supervision (gold path p^{+} or answer links),

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(G(q,p^{+}))}{\exp(G(q,p^{+})) + \sum_{p^{-}\in\mathcal{N}} \exp(G(q,p^{-}))}, \tag{11}$$

$$\mathcal{L}_{\text{cons}} = \sum_{t\in p^{+}} \max(0, \tau - s_{\text{KG}}(t)) \;+\; \text{CE-Calib}(\tilde{s}_{\text{ER}}), \tag{12}$$

encouraging dual-grounded alignment and calibrated confidence.

4.3 Query-Graph Matching, Boundary Discovery, and Iterative Refinement

KG-QUEST performs *dynamic representation learning* without static clustering by (i) *query-graph matching*, which restricts candidate nodes/edges to those satisfying the typed constraints of the question, concentrating attention mass and reducing negative sampling variance; (ii) *frontier expansion with hop decay*, which regularizes path length and predicate specificity via $m(t)$; and (iii) *boundary discovery*, which halts expansion when marginal utility falls below a threshold, yielding a compact, answer-specific subgraph $\widehat{\mathcal{G}}$ that serves as an inductive bias for Eq. 9–10. Empirically, this improves sample efficiency and stabilizes learning relative to global, static neighborhoods.

At iteration t, KG-QUEST selects the best path $p^{\star} = \arg\max_{p\subset\widehat{\mathcal{G}}} G(q,p)$, identifies unmatched query edges along $p^{\star}$, and (if needed) retrieves verifiable text to propose $\Delta\mathcal{T}$. New triples are ER-scored and integrated; embeddings are updated by weighting each triple with its reliability score and composing

$$p_{t+1} \leftarrow \text{Compose}\big(\{\widetilde{t}\} \cup \Delta\widetilde{t}; q\big),$$

reducing uncertainty $u(p)$ until convergence or budget exhaustion. In effect, frontier operations (match → expand → stop) become feature operations (filter → weight → compose).

The procedure admits a constrained EM view: the *E-step* (evidence) comprises query-graph matching, frontier expansion, and optional retrieval that propose structure and assign priors $m(t)$; the *M-step* (model) fuses ER/KG reliability scores with dual-grounded scoring (Eq. 10) to re-weight and compose representations for maximum alignment. Thus, KG-QUEST functions as a dynamic, dual-grounded representation learner whose hypothesis space is carved by the question itself, not by static clustering of a global graph.

5 Experimental Results

5.1 Datasets

We evaluate on two complementary benchmarks that together cover clinical reasoning and broad general knowledge. Table 1 compares *MedQA (USMLE)* Jin et al. (2020) and *MMLU* Hendrycks et al. (2021). MedQA emphasizes clinical reasoning via vignette-style multiple-choice questions (MCQs), whereas MMLU spans 57 subjects (STEM, humanities, and social sciences) with standardized 4-option MCQs. Both report top-1 accuracy (MMLU often under few-shot prompts).

Beyond the aggregate MMLU score, we also evaluate on six *MMLU medical subsets* to obtain fine-grained medical performance. Table 2 lists the datasets used in our experiments: the *core* benchmark (MedQA) and the MMLU medical subsets (Clinical Knowledge, Medical Genetics, Anatomy, Professional Medicine, College Biology, College Medicine), with sizes ranging from 100 to 272 items. This configuration provides a balanced mix of large-scale clinical evaluation (MedQA, 1,273 items) and targeted medical domains (e.g., Medical Genetics, 100 items).

Table 1: Comparison of two QA datasets used in biomedical and general-domain evaluation.

Property	**MedQA (USMLE)** Jin et al. (2020)	**MMLU** Hendrycks et al. (2021)
Domain	Clinical medicine (USMLE-style)	57 subjects (STEM, humanities, social sciences, etc)
Format	Vignette-style MCQs (single best)	4-option MCQ
Input	Clinical vignette text	Standalone question + 4 choices
Scale	Dev: 11,450; Test: 1,273	∼15,908 total
Metric	Top-1 accuracy	Top-1 accuracy (few-shot)
Use	Clinical reasoning	General knowledge and reasoning

Table 2: Evaluation datasets for benchmarking biomedical and clinical knowledge in LLMs. Core benchmarks are listed alongside the MMLU medical subsets.

Cat.	**Dataset**	**Size**	**Description**
Core	MedQA (USMLE)	1,273	USMLE-style MCQs; clinical reasoning; general medical knowledge
MMLU	Clinical Knowledge	265	Applied clinical concepts (MCQs)
	Medical Genetics	100	Inheritance, molecular biology, disease associations
	Anatomy	135	Human anatomy and structural knowledge
	Professional Medicine	272	Advanced medical knowledge; professional practice
	College Biology	144	Undergraduate biology; medical foundations
	College Medicine	173	Pre-clinical medical knowledge; college-level curriculum

5.2 Knowledge Graph Generation Results

We generated knowledge graphs (KGs) by extracting *Entity–Attribute–Value (EAV)* and *Entity–Relation–Entity (ERE)* triples from the *MMLU medical subsets* and *MedQA*. The pipeline used GPT-5, Gemini Flash 2.0, and Grok for gold-standard extraction, followed by QLoRA fine-tuning of LLaMA 3.1-8B. The fine-tuned extractor (rank r=32, α=16, dropout 0.25; 1.36×10^7 trainable parameters, 0.30% of total) was trained for 20 epochs with AdamW, cosine decay, and dataset-specific learning rates (2×10^{-5} for MMLU, 2×10^{-4} for MedQA), producing reliable triples aggregated into structured KGs.

Table 3 shows cross-entropy training and validation losses: most MMLU subsets converge to low validation loss (≤ 0.7), while MedQA (0.97) is moderately higher, reflecting the complexity of USMLE-style vignettes. Table 4 summarizes structural statistics: MedQA yields the largest graph (36k triples; 22k nodes), with *Professional Medicine* the largest MMLU subset (11.7k triples) and *Medical Genetics* the smallest (2.4k triples). Across datasets, graphs are connected, moderately sparse, and well-suited for traversal and path-based reasoning.

Summary. The pipeline consistently produces large, connected, and moderately sparse KGs: MedQA and Professional Medicine yield tens of thousands of triples, while smaller subsets still form coherent graphs—providing a robust substrate for reasoning and explainable QA.

5.3 MMLU Results

Figure 1 presents performance across **Accuracy, Precision (macro), Recall (macro), and F1 (macro)** on six MMLU medical subsets. Several observations emerge. First, *Medical Genetics* and *College Biology* show the strongest overall results, with all metrics exceeding 0.95, indicating both high correctness and balanced precision–recall performance. Second, *Anatomy (KG)* also achieves consistently high scores (∼0.93), suggesting that structured knowledge grounding particularly benefits ontology-heavy domains. In contrast, *College Medicine* lags behind (∼0.86 across metrics), reflecting greater complexity and variability in pre-clinical medical curricula. Finally, across all subsets, the closeness of accuracy, precision, recall, and F1 demonstrates that the model not only achieves high accuracy but also avoids bias toward particular answer classes, yielding stable performance under different evaluation perspectives. Together, these results confirm that **KG-QUEST** delivers robust, domain-aligned improvements across diverse medical reasoning tasks.

Table 5 summarizes representative accuracies on the original MMLU benchmark and its six medical subsets. On the overall benchmark, closed-source and open-source LLMs cluster in the mid-to-high

Table 3: Cross-entropy training and validation losses.

Dataset	Train Loss	Val. Loss
MMLU–Anatomy	0.43	0.51
MMLU–Clinical Knowledge	0.52	0.66
MMLU–Medical Genetics	0.57	0.63
MMLU–College Biology	0.61	0.78
MMLU–College Medicine	0.51	0.57
MMLU–Professional Medicine	0.68	0.70
MedQA	0.74	0.97

Table 4: Graph statistics of generated KGs.

Dataset / Subset	Triples	Entities	Attributes	Values	Nodes	Edges
MMLU–Anatomy	3,474	1,157	299	1,843	2,313	3,439
MMLU–Clinical Knowledge	5,399	1,845	442	2,921	3,690	5,399
MMLU–Medical Genetics	2,413	810	234	1,222	1,620	2,373
MMLU–College Biology	3,704	1,287	334	1,927	2,573	3,688
MMLU–College Medicine	4,839	1,650	458	2,519	3,301	4,839
MMLU–Professional Medicine	11,701	3,867	736	6,495	7,741	11,701
MedQA	35,955	11,065	2,067	19,800	22,142	35,955

80s: *GPT-4* reports 86.4% OpenAI (2023), *GPT-4o mini* reaches 82.0% Tong (2024), and *Llama-3 70B* achieves 88.6% Paul (2024). This saturation motivates evaluation on the medical subsets, which remain more discriminative.

Across the six MMLU medical subsets, **KG-Quest (ours)** attains a macro average of **92.0%**, outperforming the strongest prior average (90.5%) by nearly +1.5 points. KG-Quest sets new state of the art on three subsets—*Anatomy* (+8.1 points), *Medical Genetics* (+1.0), and *College Medicine* (+2.9)—and is competitive on the remaining three. Relative to GPT-4 (5-shot), KG-Quest is higher on 5/6 subsets, with especially large gains in structure-heavy domains such as Anatomy and Medical Genetics. These results demonstrate that *explicit KG grounding combined with sequence modeling* improves reasoning fidelity beyond prompt-only large language models.

Table 5: Accuracies (%) on MMLU medical subsets. **KG-Quest (Ours)** compared against prior baselines: GPT-4 and GPT-4-base (5-shot) OpenAI (2023), Med-PaLM 2 best and ER variants Nori et al. (2024), and Flan-PaLM Nori et al. (2024). Bold indicates the best per row.

Dataset	KG-Quest (Ours)	GPT-4 (5-shot)	GPT-4-base (5-shot)	Med-PaLM 2 (best)	Med-PaLM 2 (ER)	Flan-PaLM (best)
Clinical Knowledge	87.9	86.4	**88.7**	**88.7**	**88.7**	80.4
Medical Genetics	**98.0**	92.0	97.0	92.0	92.0	75.0
Anatomy	**93.3**	80.0	85.2	84.4	84.4	63.7
Professional Medicine	91.0	93.8	93.8	**95.2**	92.3	83.8
College Biology	95.8	95.1	**97.2**	95.8	95.8	88.9
College Medicine	**86.1**	76.9	80.9	83.2	83.2	76.3
MMLU Average	**92.0**	87.4	90.5	89.9	89.4	78.0

5.4 MedQA Results

Table 6 presents representative results on MedQA (USMLE), alongside prior state-of-the-art and baselines. Results are sorted from lowest to highest accuracy. Several trends emerge. First, early baselines underperform: *Flan-PaLM (best)* achieves only 67.6%, while *GPT-4 (5-shot)* improves to 81.4%. Second, domain adaptation and explicit reasoning push performance higher: *Med-PaLM 2 (ER)* reaches 85.4%, and *Med-PaLM 2 (best)* slightly improves to **86.5%**, just above *GPT-4-base (5-shot)* at 86.1%. Third, agent-based augmentation remains less effective: AgentClinic records 56.1%, far below tuned LLMs.

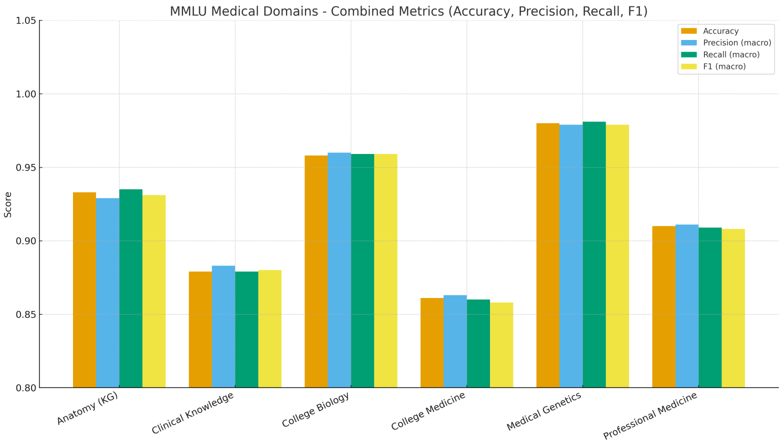

Figure 1: Performance on **MMLU medical domains** across combined metrics. Bars show Accuracy, Precision (macro), Recall (macro), and F1 (macro) for six subsets: Anatomy, Clinical Knowledge, College Biology, College Medicine, Medical Genetics, and Professional Medicine. The results highlight that KG-QUEST achieves consistent improvements across multiple evaluation criteria.

KG-QUEST (ours) delivers a substantial advance with a new state-of-the-art: macro accuracy of **93.7%**, macro F1 of **93.7**, and balanced per-class performance. In particular, KG-QUEST maintains high recall across all answer classes (ranging from 0.925 to 0.949), with the strongest performance on option D (F1 = 0.953). This demonstrates that explicit KG grounding not only improves overall accuracy but also reduces class-level variance compared to baselines.

Table 6: Representative results on **MedQA (USMLE)** benchmark, sorted from lowest to highest accuracy. **KG-QUEST (Ours)** achieves the strongest performance with macro metrics reported.

Model / Method	**Accuracy (%)**	**Notes**
KG-SFT Chen et al. (2025)	41.8	3-hop commonsense
AgentClinic (Schmidgall et al., 2024)	56.1	agent tools (Notebook memory)
Flan-PaLM (best) Nori et al. (2024)	67.6	early generation baseline
GPT-4 (5-shot) OpenAI (2023)	81.4	general-purpose baseline
Med-PaLM 2 (ER) Nori et al. (2024)	85.4	explicit reasoning variant
GPT-4-base (5-shot) Nori et al. (2024)	86.1	reference setting
Med-PaLM 2 (best) Nori et al. (2024)	86.5	domain-tuned, instruction-optimized
KG-QUEST (Ours)	**93.7**	Macro: Prec.=93.7, Rec.=93.8, F1=93.7

Summary. KG-QUEST establishes a new SOTA on MedQA with **93.7%** macro accuracy and balanced per-class precision/recall, substantially outperforming Med-PaLM 2 and GPT-4 baselines. This highlights the effectiveness of KG-grounded reasoning in complex clinical QA tasks.

6 Conclusion

We introduced *KG-QUEST*, a framework that extends LLM-based QA with dynamically constructed, answer-specific knowledge graphs over EAV/ERE triples. By grounding inference in structured relations rather than static retrieval or parametric memory, KG-QUEST establishes new state-of-the-art performance: **92.0%** average accuracy on MMLU medical subsets and **93.7%** on MedQA (USMLE), substantially outperforming GPT-4 and Med-PaLM 2 baselines. These gains are especially pronounced in ontology-heavy domains such as Anatomy and Medical Genetics, where explicit KG grounding reduces hallucination and improves interpretability. Beyond accuracy, the pipeline produces large, connected, domain-consistent graphs that support transparent reasoning and robust error analysis. Looking forward, we will extend KG-QUEST to multimodal signals, long-context agentic retrieval, and real-world clinical decision support tasks beyond exam-style QA.

ETHICS STATEMENT

We adhered to the ICLR Code of Ethics. This work uses only publicly available datasets (MMLU, MedQA) that contain no personally identifiable information and are licensed for research use. Potential risks include over-reliance on automated clinical decision support; to mitigate this, our framework incorporates abstention and explicit evidence chains for interpretability. Fairness is considered through macro metrics across answer classes, though no demographic attributes are present in the datasets. No human subjects or patient data were collected, so IRB approval was not required. We declare no conflicts of interest or external sponsorship that could influence this work.

REPRODUCIBILITY STATEMENT

We provide anonymized code and experiment scripts in the supplementary materials, including dataset preprocessing, model configurations, and training schedules. Data splits, architectures, hyperparameters, and extended ablation results are fully documented in the supplementary to ensure transparency. All experiments were run with fixed random seeds, and environment files are included to replicate the reported results. If the paper is accepted, we will release the full codebase and data processing pipelines on GitHub to support community use and further research.

LLM USAGE STATEMENT

Large language models (LLMs), specifically OpenAI's ChatGPT, were used to assist in paper preparation. Their role was limited to language refinement, LaTeX formatting, and generating alternative phrasings of author-written content. All scientific ideas, experimental design, analysis, and claims were conceived, implemented, and verified by the authors. LLM outputs were carefully reviewed and edited for accuracy. No part of the research methodology, data analysis, or results relies on unverifiable LLM generation. The authors take full responsibility for the content of this paper.

REFERENCES

Joshua Achiam et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

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

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, 2013.

Kathi Canese and Sarah Weis. Pubmed: The bibliographic database. In *The NCBI Handbook*. National Center for Biotechnology Information (US), Bethesda, MD, 2nd edition, 2013. URL https://www.ncbi.nlm.nih.gov/books/NBK153385/.

Hanzhu Chen, Xu Shen, Jie Wang, Zehao Wang, Qitan Lv, Junjie He, Rong Wu, Feng Wu, and Jieping Ye. Knowledge graph finetuning enhances knowledge manipulation in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Aakanksha Chowdhery et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24:1–113, 2023.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.

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*, 2021. URL https://arxiv.org/abs/2009.03300. NeurIPS 2021 Datasets and Benchmarks Track.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

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. *arXiv preprint arXiv:2009.13081*, 2020. URL https://arxiv.org/abs/2009.13081.

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 *EMNLP-IJCNLP 2019*, pp. 2567–2577, 2019. doi: 10.18653/v1/D19-1259.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*, 2024.

Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, Cehao Yang, Jiaxin Mao, and Jian Guo. Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Harsha Nori, Joshua M. Mayer, Nicholas King, Daniel Carignan, Eric Horvitz, Harlan M. Krumholz, and et al. Evaluation of med-palm 2 on medical question answering. *Nature Medicine*, 30: 1032–1041, 2024. doi: 10.1038/s41591-024-03423-7.

OpenAI. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. URL `https://arxiv.org/abs/2303.08774`.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Katie Paul. Meta releases newest ai model Llama 3 and ai assistant. *Reuters*, April 2024. News report.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *CoRR*, 2024.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, et al. Towards expert-level medical question answering with large language models. *Nature Medicine*, 2023a. Early online release prior to print, superseded by Singhal et al. (2025).

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.

Karan Singhal et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023b.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*, 2019.

Anna Tong. Openai unveils cheaper small ai model GPT-4o mini. *Reuters*, July 2024. News report.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pp. 2071–2080, 2016.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.