DUAL-KG: Dynamic & Static Knowledge Graph Fusion for Biomedical QA

Anonymous ACL submission

Abstract

We present **DUAL-KG**, a hybrid biomedical question answering framework that targets three persistent failure modes of LLM-based BioQA: *knowledge staleness*, *unsupported generation*, and *opaque evidence attribution*. DUAL-KG decomposes complex biomedical questions into clinically meaningful PICO-aligned sub-questions, constructs a query-specific *Dynamic Knowledge Graph* from recent literature, and complements it with a curated *Static Knowledge Graph* to improve coverage and disambiguation. Across PubMedQA, BioASQ 10b, and MedHop, DUAL-KG yields consistent gains over strong text-only baselines and substantially improves traceability: evidence-path completeness increases from 52.3% to 83.5%, and mechanical verifiability increases from 18.6% to 73.4%. However, unsupported high-confidence claims remain non-trivial and performance degrades on difficult multi-step reasoning tasks, highlighting open challenges in confidence calibration and error accumulation across retrieval, extraction, and fusion. Overall, DUAL-KG provides a practical evidence organization and verification layer toward more auditable biomedical QA systems.

1 Introduction

Large language models (LLMs) have become a central paradigm for biomedical question answering (BioQA), enabling natural-language access to tasks such as literature triage, evidence aggregation, and hypothesis exploration (Lee et al., 2020; Godwin et al., 2025; He et al., 2025; Jin et al., 2023; Lopez et al., 2025; Low et al., 2025). Rather than updating model weights, most systems rely on prompt- and pipeline-level design, using PubMed—now exceeding 38 million abstracts—as an external memory (Li et al., 2022b). Despite this scale, reliability remains a critical concern: hallucination, knowledge staleness, and opaque evidence attribution persist in medical domains (Pham and Vo, 2024; Seo et al., 2024).

Early benchmarks such as BioASQ (Tsatsaronis et al., 2015; Nentidis et al., 2024), PubMedQA (Jin et al., 2019), and MEDIQA (Abacha et al., 2019) demonstrated the feasibility of LLM-based BioQA, but also exposed systematic limitations, including reliance on small fixed document sets, weak multi-hop reasoning, and optional evidence provenance. Retrieval-augmented methods incorporate recent literature (Luo et al., 2022b), yet remain sensitive to retrieval errors and context-window constraints (Zhang et al., 2025; Jin et al., 2024). Knowledge-graph–augmented approaches improve logical consistency, but inherit the slow update cycle of static graphs (Feng et al., 2025b). Recent tri-modal systems combine LLMs, retrieval, and graphs, but typically treat each query monolithically, leaving conflicts between newly published evidence and curated knowledge unresolved (Zhao et al., 2025; Matsumoto et al., 2024).

We propose DUAL-KG, a dual-memory BioQA framework that explicitly structures external knowledge and intermediate reasoning. DUAL-KG integrates (1) **Dynamic Knowledge Graphs (DKGs)**, constructed on-the-fly from query-relevant literature to capture recency; (2) **Static Knowledge Graphs (SKGs)**, derived from expert-curated ontologies to provide stable background knowledge; and (3) **PICO-based step-wise decomposition**, which breaks complex queries into clinically meaningful sub-questions (Schardt et al., 2007). We formalize a *DKG–SKG fusion task* that constructs minimal evidence graphs supporting `yes|no|maybe` assessments under explicit uncertainty. We evaluate DUAL-KG against strong RAG-based pipelines on PubMedQA (Jin et al., 2019), BioASQ-10b (Krithara et al., 2021), and MedHop (Welbl et al., 2018), and introduce a lightweight baseline, **KG-RAG-Lite**, for controlled comparison.

Contributions.

- We propose a dual-memory BioQA architecture that jointly conditions an LLM on dynamic and static knowledge graphs, improving evidence coverage and traceability.
- We introduce a conflict-aware fusion mechanism that aligns heterogeneous entities and aggregates competing evidence under explicit uncertainty.
- We demonstrate consistent gains over strong RAG baselines on three public BioQA benchmarks, while analyzing remaining limitations in confidence calibration and multi-hop reasoning.

2 Related Works

2.1 Knowledge Graphs and Retrieval-Augmented Generation

LLM-based biomedical systems exhibit strong linguistic fluency but remain prone to hallucination, knowledge staleness, and opaque reasoning (Lee et al., 2020; Huang et al., 2019; Li et al., 2022a; Hossain et al., 2023). Retrieval-augmented generation (RAG) mitigates these issues by conditioning generation on retrieved evidence (Jin et al., 2023; Li et al., 2025; Amugongo et al., 2025; Zhao et al., 2024), but inherits retriever limitations such as missed documents, context overflow, and unresolved disagreement among sources (Farahani and Johansson, 2024; Chowdhury et al., 2025).

Knowledge graphs (KGs) support structured reasoning and explicit provenance, yet are costly to construct and slow to update, limiting their coverage of emerging biomedical knowledge (Rosenbloom et al., 2006; Alshahrani et al., 2022; Nicholson and Greene, 2020; Morris et al., 2023). Recent work increasingly integrates LLMs, RAG, and KGs within unified pipelines (Jeong et al., 2024; Gargari and Habibi, 2025), but typically treats textual and graph knowledge as implicitly consistent context, leaving conflicts to be resolved implicitly during generation.

2.2 Hybrid BioQA Systems

Early hybrid systems focused on post hoc verification against static KGs, improving logical consistency but remaining coverage-limited (Soman et al., 2024). Subsequent approaches introduced deeper integration, including retrieval graphs for long-range reasoning (GraphRAG, Med-GraphRAG), domain-specific dynamic extraction (DALK) (Li et al., 2024), and structured generate–verify pipelines emphasizing transparency (Zhan et al., 2024; Matsumoto et al., 2024). Other systems incorporate anatomical maps, clinical codes, or ontologies for specialized BioQA settings (Zhao et al., 2025; Feng et al., 2025a; Wang et al., 2025).

Despite these advances, dynamic updates remain coarse-grained, conflicts between new evidence and curated knowledge are rarely modeled explicitly, and early-stage errors can propagate through multi-module pipelines (Toro et al., 2024; Lopez et al., 2025). In contrast, *DUAL-KG* treats conflicts between dynamic, literature-derived knowledge and static, expert-curated relations as first-class reasoning objects. By constructing query-specific minimal graphs and fusing competing evidence under explicit uncertainty, DUAL-KG emphasizes traceability and verifiability over autonomous resolution.

3 DUAL-KG: A Dual-Memory Framework for Biomedical QA

DUAL-KG equips a large language model (LLM) with a dual external memory to address persistent challenges in biomedical QA, including hallucination, knowledge staleness, and limited evidence traceability (Niu et al., 2023; Feng et al., 2025c). Rather than performing autonomous clinical decision-making, DUAL-KG functions as an evidence organization and verification layer that supports expert judgment. The framework combines an expert-curated **Static Knowledge Graph (SKG)** for stable background knowledge (Stear et al., 2024; Su et al., 2023) with a query-specific **Dynamic Knowledge Graph (DKG)** constructed from recent biomedical literature to improve temporal coverage (Bonner et al., 2022; Bang et al., 2023). An overview is shown in Figure 1, with key design motivations summarized in Table 1.

3.1 Architectural Pipeline

Given a query Q, DUAL-KG produces an evidence-backed response through the following stages.

Structured query decomposition. The query is decomposed into PICO-aligned sub-questions $\{q_i\}$ (Schardt et al., 2007), organizing reasoning into clinically meaningful units and reducing retrieval ambiguity. Redundant or low-utility sub-questions are filtered, and remaining ones are typed to guide downstream processing.

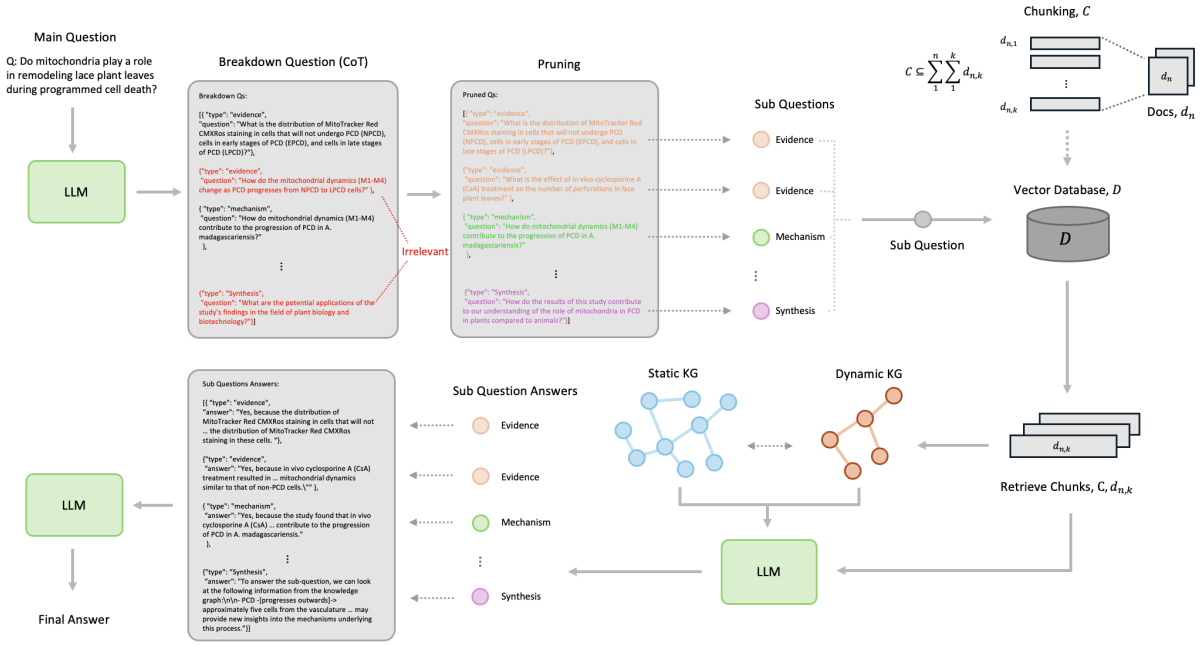

Figure 1: DUAL-KG architecture. The pipeline decomposes a query into structured sub-questions, retrieves evidence from static and dynamic knowledge sources, constructs and fuses query-specific graphs, and synthesizes an evidence-backed answer with an explicit provenance trail.

Table 1: Motivating design choices behind DUAL-KG.

Challenge	Why vanilla LLMs fall short	DUAL-KG design choice
Hallucination	Purely parametric generation with weak grounding	Evidence-scoped generation with explicit citations to snippets and KG triples
Staleness	Knowledge frozen at pre-training cut-off	Query-specific DKG construction from time-filtered literature retrieval
Opacity	Implicit reasoning without inspectable structure	Explicit evidence paths via typed KG edges and linked citations

Dual knowledge construction. For each q_i, evidence is collected from two complementary sources. A dynamic graph $\mathcal{K}_i^{\text{dyn}}$ is constructed by retrieving relevant literature and extracting relational triples with attached provenance (Luo et al., 2022a; Li et al., 2022b). In parallel, a local static subgraph $\mathcal{K}_i^{\text{stat}}$ is retrieved from an expert-curated KG via limited-hop expansion (Livingston et al., 2015; Ye et al., 2025).

Graph fusion and conflict handling. Dynamic and static graphs are fused through canonical entity alignment using biomedical identifiers, with approximate matching used only for candidate generation (Bodenreider, 2004). Conflicting claims between sources are preserved with provenance rather than collapsed. When applicable, preference annotations are added while retaining alternative edges, enabling comparison rather than forced resolution.

Evidence-based synthesis. Each fused graph $\mathcal{F}_i$ and its supporting passages are provided to an instruction-tuned LLM, which produces a provisional answer, confidence label, and explicit citations (Wang et al., 2023; Gao et al., 2024; Shi et al., 2025). An evidence-aware confidence policy restricts strong confidence to answers supported by both literature snippets and KG triples.

Aggregation and output. Sub-question answers are aggregated via consistency-weighted voting based on evidence availability and agreement with the fused graph. The system returns a final answer together with a complete evidence trail, including cited passages and a rendered evidence graph to support auditing (Xinna Lin et al., 2024).

3.2 Experimental Configuration

Table 2 summarizes the hyper-parameters used in our experiments.

Discrete retrieval and information-extraction prompt choices are lightly tuned using a policy-gradient (REINFORCE-style) objective (Williams, 1992) on a held-out development split. We use BioASQ dev *MAP* only as a task-level *tuning proxy*

Table 2: Key hyper-parameters for the DUAL-KG framework.

Hyper-parameter	Value
Main Components	
LLM Backbone	Mistral-7B-Instruct (8k context)
Retriever (Dense)	BioLinkBERT-base
Information Extraction (IE)	GPT-3.5-Turbo (baseline)
Pipeline Configuration	
Retrieval Top-k Passages	10 (3–6 for demos)
KG k-hop Expansion	2 (3 for MedHop)
Entity Linking Threshold (τ_{fuzzy})	70
LLM Temperature	0.0
Max Triples to LLM	20
Ablation Flags (Defaults)	
Use Sub-Questioning	`True`
Use Dynamic KG	`True`
Use Static KG	`True`

for prompt selection (reward), while all *reported* BioASQ results in §4 use ROUGE-L F1 to match standard benchmark practice. This lightweight tuning yields an additional 1.2 pp improvement without updating any LLM parameters.

On PubMedQA, a full inference run averages approximately 5 s (retrieval 2 s, IE 1 s, generation 2 s) under our hardware setup. Ablation experiments show that removing either the Dynamic KG or Static KG consistently reduces performance, supporting the dual-memory design.

4 Experiments

4.1 Evaluation Benchmarks and Metrics

We evaluate **DUAL–KG** on three public biomedical QA benchmarks of increasing difficulty: PubMedQA, a yes/no/maybe classification task (Jin et al., 2019); BioASQ 10b, a free-form biomedical QA task (Krithara et al., 2021); and MedHop, a multi-hop drug–drug interaction reasoning benchmark (Welbl et al., 2018).

We report Accuracy for PubMedQA and MedHop, and ROUGE-L F1 for BioASQ. To evaluate the framework beyond end-task scores, we additionally track auditability- and deployment-oriented metrics that operate at the *sub-question* level (unless stated otherwise):

- **Operational Hallucination:** The fraction of sub-answers that are expressed with *strong* confidence but cite neither a valid literature snippet ($\leq$25 words) nor a supporting KG triple.
- **Explainability and Traceability:** Evidence-path completeness, mechanical verifiability, and step traceability, which quantify whether each intermediate decision can be inspected through explicit provenance.
- **Efficiency:** Latency (s/problem), throughput (problems/hour), and cost ($/problem) computed from Together billing records under a fixed inference stack.

4.2 Implementation Details

All experiments are conducted on a fixed hardware configuration (Ryzen 5 5600X CPU, 32 GB RAM, RTX 3060 Ti GPU) and use an identical retrieval index consisting of 33M PubMed text chunks indexed with FAISS IVFPQ. We issue up to eight sub-queries concurrently via API key rotation to stabilize end-to-end latency under rate limits.

We evaluate six open-weight LLMs served through the Together inference endpoint: *Llama 3.3 70B, Llama 3.1 70B, Llama 4 Maverick 17B, Llama 3.1 8B, Qwen 2.5 72B,* and *Qwen 2 72B*. All models are run with temperature set to zero, and all other decoding parameters follow provider defaults.

To ensure comparability across models, we fix the retrieval index, the sub-questioning prompts, and the information extraction module, varying only the answering LLM. Costs are computed from Together billing records and include both prompt and completion tokens. Because provider pricing and serving conditions may change, absolute dollar values should be interpreted as point-in-time estimates; relative trends across models are the main focus.

5 Main Results

We evaluate Dual-KG on PubMedQA, BioASQ, and MedHop using six open-weight LLMs under a unified inference and evaluation protocol (Table 3). Our analysis focuses on task performance, hallucination and confidence behavior, traceability, and efficiency. Unless stated otherwise, metrics are computed at the sub-question level and averaged per dataset.

Overall task performance. Table 4 summarizes dataset-specific metrics and the Overall Score. Llama 3.3 70B achieves the best overall performance (0.578), followed closely by Qwen 2.5 72B (0.568). Across models, PubMedQA is consistently easier than BioASQ and MedHop, reflecting the difficulty of free-form synthesis and multi-step reasoning. Performance increases with model size,

Table 3: Operational metrics used to assess hallucination, explainability, and efficiency in DUAL–KG. Unless noted otherwise, metrics are computed per sub-question and then averaged across questions within each benchmark.

Metric Category	Definition and Computation
Hallucination	
Overall Hallucination Rate	$\frac{\#\text{ hallucinated sub-answers}}{\#\text{ total sub-answers}} \times 100\%$. A sub-answer is hallucinated if it is labeled *strong* confidence but cites neither (i) a valid supporting snippet nor (ii) any KG triple.
Confidence Mismatch	Percentage of sub-answers whose confidence label disagrees with correctness. We report mismatch rates for *strong* and *medium* separately when relevant.
KG–Answer Mismatch	Rate at which the predicted label contradicts the polarity implied by its cited KG evidence after entity canonicalization (e.g., a `yes` label supported only by contradictory KG relations).
Poor Citation	Sub-answers with zero citations or invalid citations (e.g., no retrievable source, snippet exceeds the 25-word cap, or snippet does not appear in the retrieved context).
Phantom Knowledge	Fraction referencing entities or relations not present in the retrieved sources or the fused graph $\mathcal{F}_i$ after canonicalization.
Knowledge Contradiction	Cases where mutually exclusive claims are produced within the same question thread (e.g., both `yes` and `no` asserted for an identical sub-question intent).
Explainability	
Evidence Path Completeness	Fraction of structured reasoning steps that return at least one linked support (KG edge or literature snippet).
Mechanical Verifiability	Share of generated claims whose cited entities and relations can be matched to the fused graph via automated look-ups after canonicalization.
Reasoning Step Traceability	Average fraction of reasoning steps that contain an explicit citation and a recoverable provenance pointer (PMID for text; canonical IDs for KG).
Entity Linking Accuracy	Percentage of extracted entity mentions that map to the correct canonical ID (e.g., DrugBank or UMLS). We estimate this metric on an annotated sample of entity mentions (details in Appendix **??**).
Efficiency	
Latency (s / problem)	Mean wall-clock time to fully answer one question (including retrieval, IE, and generation).
Cost ($ / problem)	Average API charge per question computed from Together billing, including both prompt and completion tokens.
Throughput (problems/hour)	$3600/\text{Latency}$, representing the expected number of questions solvable per hour under the same stack.

but gains among 70B-scale models are modest under a fixed retrieval and extraction stack, indicating diminishing returns from scale alone.

Metric aggregation. Hallucination and explainability metrics are computed per sub-question. Evidence-path completeness measures whether each reasoning step is supported by at least one KG edge or literature snippet, while mechanical verifiability measures whether cited entities and relations can be matched to the fused graph after canonicalization.

5.1 Traceability and Explainability

Dual-KG substantially improves auditability relative to a text-only citation baseline (Table 5). By grounding intermediate decisions in structured triples linked to canonical biomedical identifiers,

Table 4: Performance comparison by dataset. The Overall Score is a weighted average of dataset-specific metrics as defined in §4.1.

Model	PubMedQA (Accuracy)	BioASQ (ROUGE-L F1)	MedHop (Accuracy)	Overall Score
Llama 3.3 70B Turbo	**0.863**	**0.438**	**0.434**	**0.578**
Qwen2.5 72B Turbo	0.855	0.429	0.421	0.568
Llama 3.1 70B Turbo	0.841	0.425	0.398	0.555
Llama 4 Maverick	0.847	0.412	0.412	0.557
Qwen 2 72B	0.833	0.408	0.387	0.543
Llama 3.1 8B Turbo	0.782	0.367	0.324	0.491

Figure 2: Example evidence trail generated by Dual-KG. The pipeline decomposes a complex query into sub-questions and grounds each intermediate conclusion in knowledge-graph relations linked to canonical IDs, supporting auditability.

Dual-KG improves evidence-path completeness and mechanical verifiability. This representation also makes failures more diagnosable, allowing errors to be localized to retrieval, entity alignment, or conflicts between dynamic and static sources.

Figure 2 illustrates a representative evidence trail produced by Dual-KG, where a complex query is decomposed into typed sub-questions and each sub-answer is grounded in explicit knowledge-graph relations.

5.2 Hallucination and Evidence-Aware Confidence

Table 6 reports operational hallucination statistics computed at the sub-question level. Following Table 3, we operationalize hallucination as *high-confidence unsupported claims*: a sub-answer labeled *strong* confidence that cites neither a supporting literature snippet nor a knowledge graph triple. Across models, the dominant failure mode is confidence mismatch, indicating that calibration remains a first-order limitation even under evidence-augmented pipelines. On average, only 68.7% of sub-answers labeled *strong* are correct, which motivates stricter evidence-aware confidence policies and abstention behavior.

Despite residual errors, incorporating knowledge graphs provides meaningful mitigation: Dual-KG reduces overall hallucination rates by approximately 31% relative to sub-queries that do not use a KG (17.1% vs. 24.8%). Remaining errors are not uniformly distributed: larger models tend to reduce phantom entities and internal contradictions, while smaller models exhibit higher rates of poor citation and KG–answer mismatches, suggesting that model capacity and explicit evidence constraints are complementary.

5.3 Calibration

We evaluate confidence calibration using Expected Calibration Error (ECE), computed over sub-question predictions with $M=10$ confidence bins. In addition to standard calibration, we apply an evidence-aware policy that down-weights confidence when answers lack both KG and literature support. Reliability diagrams and ECE (Figure 3) show that explicit evidence constraints improve, but do not fully resolve, overconfidence.

5.4 Efficiency Trade-offs

Table 7 reports latency, cost, and throughput. Llama 3.3 70B achieves the highest performance but incurs the highest cost ($0.0281 per problem). In contrast, Llama 3.1 8B offers the best throughput and lowest cost ($0.0039), while still benefiting from explicit evidence grounding. These results indicate that Dual-KG supports multiple deployment regimes, ranging from high-accuracy decision support to large-scale literature triage.

Table 5: Explainability and traceability comparison against a text-only citation baseline.

Explainability Dimension	DUAL-KG Score	Text Baseline Score	Improvement (%)
Evidence Path Completeness	83.5%	52.3%	+59.7%
Mechanical Verifiability	73.4%	18.6%	+294.6%
Reasoning Step Traceability	78.1%	44.5%	+75.5%
Entity Linking Accuracy	79.6%	31.2%	+155.1%

Table 6: Hallucination analysis by model: percentage of sub-queries exhibiting different operational hallucination types.

Model	Overall Hallucination	Confidence Mismatch	KG–Answer Mismatch	Poor Citation	Phantom Knowledge	Knowledge Contradiction
Llama 3.3 70B	**16.3%**	**4.7%**	**3.8%**	**4.2%**	**2.1%**	1.5%
Qwen2.5 72B	16.9%	5.1%	4.1%	4.3%	2.3%	1.1%
Llama 3.1 70B	17.5%	5.4%	4.3%	4.6%	2.4%	0.8%
Llama 4 Maverick	18.7%	6.2%	4.8%	4.9%	**2.1%**	0.7%
Qwen 2 72B	19.2%	6.1%	5.2%	5.1%	2.2%	**0.6%**
Llama 3.1 8B	33.4%	10.9%	8.7%	7.3%	4.1%	2.4%

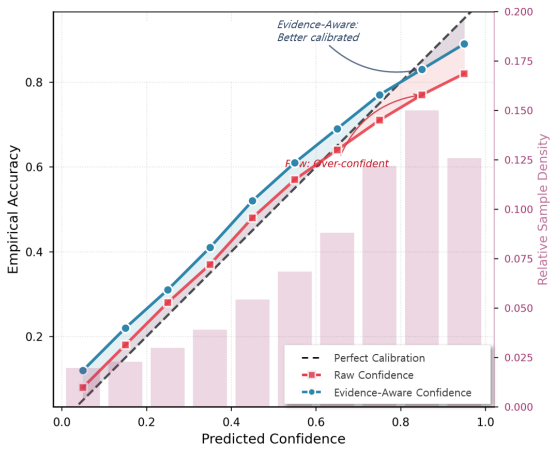

Figure 3: Reliability diagram and ECE for evidence-aware confidence. Each point corresponds to a confidence bin (10 equal-width bins), showing the empirical accuracy of sub-question answers versus the mean predicted confidence in that bin. The diagonal indicates perfect calibration; deviations indicate over- or under-confidence. We report ECE alongside the diagram to summarize calibration error.

5.5 Ablation Study: Component Contributions

To quantify the contribution of major components within Dual-KG, we conduct an ablation study using Llama 3.3 70B. Table 8 shows that removing any component degrades performance, indicating that Dual-KG benefits from the interaction of dynamic evidence extraction, static background grounding, and graph fusion rather than a single dominant module. Removing Dynamic KG produces the largest degradation, emphasizing the importance of incorporating up-to-date literature evidence. Removing Static KG also causes a clear drop, consistent with the role of curated background knowledge in recall and entity disambiguation. Finally, removing global graph fusion yields a smaller but consistent decrease, suggesting complementary benefits from integrating dynamic and static evidence into a unified representation.

6 Discussion

Dual-KG addresses a core limitation of biomedical QA: producing answers that are current, evidence-grounded, and auditable under rapidly evolving medical knowledge. Across three benchmarks, combining query-specific Dynamic KGs with curated Static KGs consistently improves evidence traceability and mechanical verifiability, often more reliably than raw accuracy.

The primary benefit of Dual-KG lies in auditability. Structured triples linked to canonical identifiers allow users to inspect why an answer is produced, rather than accepting opaque generations, which is essential in biomedical settings. However, unsupported high-confidence claims persist. Only 68.7% of strong-confidence sub-answers are cor-

Table 7: Efficiency and performance–cost trade-off analysis across models.

Model	Overall Score	Avg Time (s) / Problem	Avg Cost ($) / Problem	Throughput (problems/hour)
Llama 3.3 70B	**0.578**	26.7	0.0281	134
Qwen2.5 72B	0.568	24.6	0.0089	146
Llama 3.1 70B	0.555	25.1	0.0265	143
Llama 4 Maverick	0.557	22.4	0.0241	161
Qwen 2 72B	0.543	23.9	0.0260	151
Llama 3.1 8B	0.491	**16.8**	**0.0039**	**214**

Table 8: Component-wise contribution analysis via ablation study (Llama 3.3 70B baseline). Performance drop is measured against the Full System.

Setting	PubMedQA	BioASQ	MedHop	Avg	Performance	
					Drop	Rank
Full System	**0.863**	**0.438**	**0.434**	**0.578**	–	–
No Dynamic KG	0.798	0.389	0.376	0.521	**-9.9%**	1st
No Static KG	0.821	0.407	0.398	0.542	**-6.2%**	2nd
No Global Graph	0.847	0.425	0.423	0.565	**-2.2%**	3rd

rect, indicating that confidence calibration remains a first-order limitation despite explicit grounding.

Ablation results confirm that both Dynamic and Static KGs contribute substantially, with additional gains from global fusion. Yet biomedical evidence is frequently contradictory or context-dependent. Collapsing such conflicts into a single narrative risks overconfidence; instead, competing claims and supporting evidence should be surfaced explicitly.

Performance on multi-hop tasks such as BioASQ and MedHop remains limited, reflecting error accumulation across retrieval, extraction, entity linking, and fusion. Future systems should therefore emphasize verification-driven inference, requiring explicit evidence support for each intermediate claim and enabling abstention when evidence is insufficient.

Overall, Dual-KG improves transparency and verifiability in biomedical QA, but calibration, conflict handling, and robust multi-step reasoning remain critical challenges for safe deployment.

7 Conclusion

We presented **Dual-KG**, a dual-memory biomedical QA framework that integrates Dynamic and Static Knowledge Graphs to improve reliability, currency, and auditability. Across three benchmarks, Dual-KG outperforms text-only baselines and delivers substantial gains in evidence traceability and mechanical verifiability, often exceeding the benefits of increased model scale.

By grounding intermediate decisions in explicit, inspectable triples, Dual-KG enables users to trace and verify model outputs, a prerequisite for high-stakes biomedical use. Smaller models benefit particularly from this explicit grounding, enabling more cost-efficient yet transparent deployments.

Key challenges remain. Unsupported high-confidence claims persist, confidence calibration is limited, and performance degrades on complex multi-step reasoning tasks. Addressing these issues through calibration-aware inference, explicit conflict reporting, and verification-driven reasoning is essential for future progress.

In sum, Dual-KG provides a principled foundation for more auditable biomedical QA, prioritizing evidence and accountability alongside accuracy.

8 Limitations

While Dual-KG demonstrates clear benefits in evidence grounding, traceability, and explainability, several limitations remain.

Residual hallucination and confidence calibration. Despite explicit grounding in knowledge graphs, hallucinations are not fully eliminated. In particular, confidence–evidence mismatch persists, with a substantial fraction of answers expressed

with strong confidence despite insufficient or conflicting evidence. This reflects a broader limitation of current large language models, whose confidence estimates are not reliably calibrated to evidence availability. As a result, Dual-KG should be viewed as a mitigation strategy rather than a complete solution to hallucination, and further work on explicit uncertainty modeling and abstention mechanisms is required.

Sensitivity to upstream errors. Dual-KG relies on a multi-stage pipeline including retrieval, information extraction, entity linking, and graph fusion. Errors introduced at early stages can propagate and compound in later steps, particularly for complex queries. Although structured representations make such errors more diagnosable, they do not prevent error accumulation, which remains a fundamental challenge for modular biomedical QA systems.

Limited support for deep multi-hop reasoning. Performance degrades on benchmarks requiring long reasoning chains, such as MedHop, indicating that the framework remains limited in handling deep multi-hop biomedical inference. This limitation arises from both incomplete evidence coverage and the difficulty of maintaining consistent intermediate reasoning across multiple steps. Future work should explore step-wise verification and intermediate validation to reduce cascading failures.

Dependence on knowledge graph quality and coverage. The effectiveness of Dual-KG depends on the quality of both Dynamic and Static Knowledge Graphs. Dynamic KGs may reflect noisy or preliminary findings from recent literature, while Static KGs may lag behind emerging discoveries or encode outdated assumptions. Although Dual-KG surfaces conflicts rather than silently collapsing them, it does not guarantee that the resolved output reflects clinical consensus.

Clinical applicability and evaluation scope. Our evaluation is limited to public biomedical QA benchmarks and does not constitute clinical validation. While Dual-KG improves transparency and auditability, it should not be interpreted as a ready-to-deploy clinical decision support system. Rigorous prospective studies and domain-expert evaluation are necessary before real-world medical use.

Overall, these limitations highlight that Dual-KG is a step toward more reliable and interpretable biomedical QA, but not a complete solution. Addressing calibration, robustness, and clinical validation remains essential for future research.

References

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379.

Mona Alshahrani, Abdullah Almansour, Asma Alkhaldi, Maha A Thafar, Mahmut Uludag, Magbubah Essack, and Robert Hoehndorf. 2022. Combining biomedical knowledge graphs and text to improve predictions for drug-target interactions and drug-indications. *PeerJ*, 10:e13061.

Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877.

Dongmin Bang, Sangsoo Lim, Sangseon Lee, and Sun Kim. 2023. Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers. *Nature Communications*, 14(1):3570.

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

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

Mohita Chowdhury, Yajie Vera He, Aisling Higham, and Ernest Lim. 2025. Astrid–an automated and scalable triad for the evaluation of rag-based clinical question answering systems. *arXiv preprint arXiv:2501.08208*.

Mehrdad Farahani and Richard Johansson. 2024. Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models. *arXiv preprint arXiv:2410.05162*.

Hui Feng, Yuntzu Yin, Emiliano Reynares, and Jay Nanavati. 2025a. Ontologyrag: Better and faster biomedical code mapping with retrieval-augmented generation (rag) leveraging ontology knowledge graphs and large language models. *arXiv preprint arXiv:2502.18992*.

Y. Feng and 1 others. 2025b. Knowledge graph–based thought: a knowledge graph–enhanced llm framework for pan-cancer question answering. *Gigascience*, 14:giae082.

Yichun Feng, Jiawei Wang, Ruikun He, Lu Zhou, and Yixue Li. 2025c. A retrieval-augmented knowledge mining method with deep thinking llms for biomedical research and clinical support. *arXiv preprint arXiv:2503.23029*.

Yichen Gao, Licheng Zong, and Yu Li. 2024. Enhancing biomedical question answering with parameter-efficient fine-tuning and hierarchical retrieval augmented generation. *CLEF Working Notes*.

Omid Kohandel Gargari and Gholamreza Habibi. 2025. Enhancing medical ai with retrieval-augmented generation: A mini narrative review. *DIGITAL HEALTH*, 11:20552076251337177.

Ryan C Godwin, Avery Tung, Dan E Berkowitz, and Ryan L Melvin. 2025. Transforming physiology and healthcare through foundation models. *Physiology*, 40(3):233–246.

Jiawei He, Boya Zhang, Hossein Rouhizadeh, Yingjian Chen, Rui Yang, Jin Lu, Xudong Chen, Nan Liu, Irene Li, and Douglas Teodoro. 2025. Retrieval-augmented generation in biomedicine: A survey of technologies, datasets, and clinical applications. *arXiv preprint arXiv:2505.01146*.

Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R Pisani, and Kathryn Turner. 2023. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Computers in biology and medicine*, 155:106649.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. 2024. Long-context llms meet rag: Overcoming challenges for long inputs in rag. *arXiv preprint arXiv:2410.05983*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

A. Krithara, A. Nentidis, . , M. Krallinger, and A. B. d. Miranda. 2021. Bioasq at clef2021: large-scale biomedical semantic indexing and question answering. *Lecture Notes in Computer Science*, pages 624–630.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Dawei Li, Shu Yang, Zhen Tan, Jae Young Baik, Sukwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding, and 1 others. 2024. Dalk: Dynamic co-augmentation of llms and kg to answer alzheimer's disease questions with scientific literature. *arXiv preprint arXiv:2405.04819*.

Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlalı, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, and 1 others. 2022a. Neural natural language processing for unstructured data in electronic health records: a review. *Computer Science Review*, 46:100511.

Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. 2025. Biomedrag: A retrieval augmented large language model for biomedicine. *Journal of Biomedical Informatics*, 162:104769.

Peng-Hsuan Li, Ting-Fu Chen, Jheng-Ying Yu, Shang-Hung Shih, Chan-Hung Su, Yin-Hung Lin, Huai-Kuang Tsai, Hsueh-Fen Juan, Chien-Yu Chen, and Jia-Hsin Huang. 2022b. pubmedkb: an interactive web server for exploring biomedical entity relations in the biomedical literature. *Nucleic Acids Research*, 50(W1):W616–W622.

Kevin M Livingston, Michael Bada, William A Baumgartner, and Lawrence E Hunter. 2015. Kabob: ontology-based semantic integration of biomedical databases. *BMC bioinformatics*, 16:1–21.

Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P Ma, April S Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, and 1 others. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45.

Yen Sia Low, Michael L Jackson, Rebecca J Hyde, Robert E Brown, Neil M Sanghavi, Julian D Baldwin, C William Pike, Jananee Muralidharan, Gavin Hui, Natasha Alexander, Hadeel Hassan, Rahul V Nene, Morgan Pike, Courtney J Pokrzywa, Shivam Vedak, Adam Paul Yan, Dong han Yao, Amy R Zipursky, Christina Dinh, and 8 others. 2025. Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems. *DIGITAL HEALTH*, 11:20552076251348850.

Man Luo, Arindam Mitra, Tejas Gokhale, and Chitta Baral. 2022a. Improving biomedical information retrieval with neural retrievers. In *Proceedings of AAAI*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022b. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Nicholas Matsumoto, Jay Moran, Hyunjun Choi, Miguel E Hernandez, Mythreye Venkatesan, Paul Wang, and Jason H Moore. 2024. Kragen: a knowledge graph-enhanced rag framework for biomedical problem solving using large language models. *Bioinformatics*, 40(6):btae353.

John H Morris, Karthik Soman, Rabia E Akbas, Xiaoyuan Zhou, Brett Smith, Elaine C Meng, Conrad C Huang, Gabriel Cerono, Gundolf Schenk, Angela Rizk-Jackson, and 1 others. 2023. The scalable precision medicine open knowledge engine (spoke): a massive knowledge graph of biomedical information. *Bioinformatics*, 39(2):btad080.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Salvador Lima-López, Eulàlia Farré-Maduell, Martin Krallinger, Natalia Loukachevitch, Vera Davydova, Elena Tutubalina, and Georgios Paliouras. 2024. Overview of bioasq 2024: the twelfth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 3–27. Springer.

David N Nicholson and Casey S Greene. 2020. Constructing knowledge graphs and their biomedical applications. *Computational and structural biotechnology journal*, 18:1414–1428.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396*.

Duy Khoa Pham and Bao Quoc Vo. 2024. Towards reliable medical question answering: Techniques and challenges in mitigating hallucinations in language models. *arXiv preprint arXiv:2408.13808*.

S. T. Rosenbloom, R. A. Miller, K. B. Johnson, P. L. Elkin, and S. H. Brown. 2006. Interface terminologies: Facilitating direct entry of clinical data into electronic health record systems. *Journal of the American Medical Informatics Association*, 13(3):277–288.

Connie Schardt, Martha B Adams, Thomas Owens, Sheri Keitz, and Paul Fontelo. 2007. Utilization of the pico framework to improve searching pubmed for clinical questions. *BMC medical informatics and decision making*, 7:1–6.

Jean Seo, Jongwon Lim, Dongjun Jang, and Hyopil Shin. 2024. Dahl: Domain-specific automated hallucination evaluation of long-form text through a benchmark dataset in biomedicine. *arXiv preprint arXiv:2411.09255*.

Yucheng Shi, Shaochen Xu, Tianze Yang, Zhengliang Liu, Tianming Liu, Xiang Li, and Ninghao Liu. 2025. Mkrag: Medical knowledge retrieval augmented generation for medical question answering. In *AMIA Annual Symposium Proceedings*, volume 2024, page 1011.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, and 1 others. 2024. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40(9):btae560.

Benjamin J. Stear, Taha Mohseni Ahooyi, J. Alan Simmons, Charles Kollar, Lance Hartman, Katherine Beigel, Aditya Lahiri, Shubha Vasisht, Tiffany J. Callahan, Christopher M. Nemarich, Jonathan C. Silverstein, Deanne M. Taylor, and et al. 2024. Petagraph: A large-scale unifying knowledge graph framework for integrating biomolecular and biomedical data. *Scientific Data*, 11:1338.

Chang Su, Yu Hou, Manqi Zhou, Suraj Rajendran, Jacqueline RMA Maasch, Zehra Abedi, Haotan Zhang, Zilong Bai, Anthony Cuturrufo, Winston Guo, and 1 others. 2023. Biomedical discovery through the integrative biomedical knowledge hub (ibkh). *Iscience*, 26(4).

Sabrina Toro, Anna V Anagnostopoulos, Susan M Bello, Kai Blumberg, Rhiannon Cameron, Leigh Carmody, Alexander D Diehl, Damion M Dooley, William D Duncan, Petra Fey, and 1 others. 2024. Dynamic retrieval augmented generation of ontologies using artificial intelligence (dragon-ai). *Journal of Biomedical Semantics*, 15(1):19.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28.

FeiLong Wang, Donghui Shi, Jose Aguilar, Xinyi Cui, Jinsong Jiang, Longjian Shen, and Mengya Li. 2025. Llm-kgmqa: large language model-augmented multi-hop question-answering system based on knowledge graph in medical field. *Knowledge and Information Systems*, pages 1–43.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2023. Augmenting black-box llms with medical textbooks for biomedical question answering (published in findings of emnlp 2024). *arXiv preprint arXiv:2309.02233*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256.

Ziri Chen Xinna Lin, Zirui Zhao, Yutong Zhang, Siyuan Wang, Zichang Liu, Kaicheng Li, and Kaixuan Yu. 2024. Biokgbench: A knowledge graph checking benchmark of ai agent for biomedical science. *Preprint*, arXiv:2407.00466.

Qing Ye, Yundian Zeng, Linlong Jiang, Yu Kang, Peichen Pan, Jiming Chen, Yafeng Deng, Haitao Zhao, Shibo He, Tingjun Hou, and 1 others. 2025. A knowledge-guided graph learning approach bridging phenotype-and target-based drug discovery. *Advanced Science*, 12(16):2412402.

Buchao Zhan, Anqi Li, Xin Yang, Dongmei He, Yucong Duan, and Shankai Yan. 2024. Rarok: Retrieval-augmented reasoning on knowledge for medical question answering. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2837–2843. IEEE.

Gongbo Zhang, Zihan Xu, Qiao Jin, Fangyi Chen, Yilu Fang, Yi Liu, Justin F Rousseau, Ziyang Xu, Zhiyong Lu, Chunhua Weng, and 1 others. 2025. Leveraging long context in retrieval augmented language models for medical question answering. *npj Digital Medicine*, 8(1):239.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pages 4442–4457.

Appendices

Prompt 1: General PICO-based Sub-query Generation

System: `You have the following PICO roadmap for answering the main question: {framework_json}. For each step in that roadmap, write concrete sub-questions which is needed to answer the main question. Do not write more than 15 sub-questions. Each sub-question must cite at least one context keyword. Assign each sub-question one of these types: "evidence", "mechanism", "consistency", "limitation", "synthesis". Your output **must** start with '[' and end with ']'. Return only the raw JSON array of objects with keys "type" and "question".`

Prompt 2: PubMedQA-specific Sub-query Generation

System: `This is a biomedical research question requiring evidence-based analysis: "{main_question}". Generate EXACTLY 4 sub-questions following this medical evidence framework: 1. Background Analysis, 2. Evidence Evaluation, 3. Methodology Assessment, 4. Clinical Synthesis. Return only the raw JSON array with keys "type" and "question".`

Prompt 3: MedHop-specific Sub-query Generation

System: `You have the following roadmap for analyzing drug interactions: Identify Primary molecule, Determine Interaction mechanism, Evaluate Candidate molecules, Assess Observable Effect. This is a multi-hop biomedical reasoning question. Write concrete sub-questions for each step. Each sub-question must cite at least one context keyword (e.g. drug codes like "DB01234"). Return only the raw JSON array with keys "type" and "question".`

Prompt 4: Sub-query Retry

System: `Your previous response was not valid JSON. Please output **ONLY** the JSON array, without any additional text. Array must start with [ and ends with ]. Elements are in the form of {"type":"...", "question":"..."}. Do not write more than 15 sub questions.`

Prompt 5: Sub-query Pruning

System: `Main question: {main_question}. Here are candidate sub-questions: {subq_json}. Context: {context}. Return a JSON array containing only the sub-question objects that are valid and plausibly contribute to answering the main question. Output **only** the JSON array.`

Prompt 6: General Sub-question Answering

System: `You are a biomedical evidence analyst. If context clearly SUPPORTS the sub-question, label is "yes". If it clearly REFUTES it, label is "no". If`

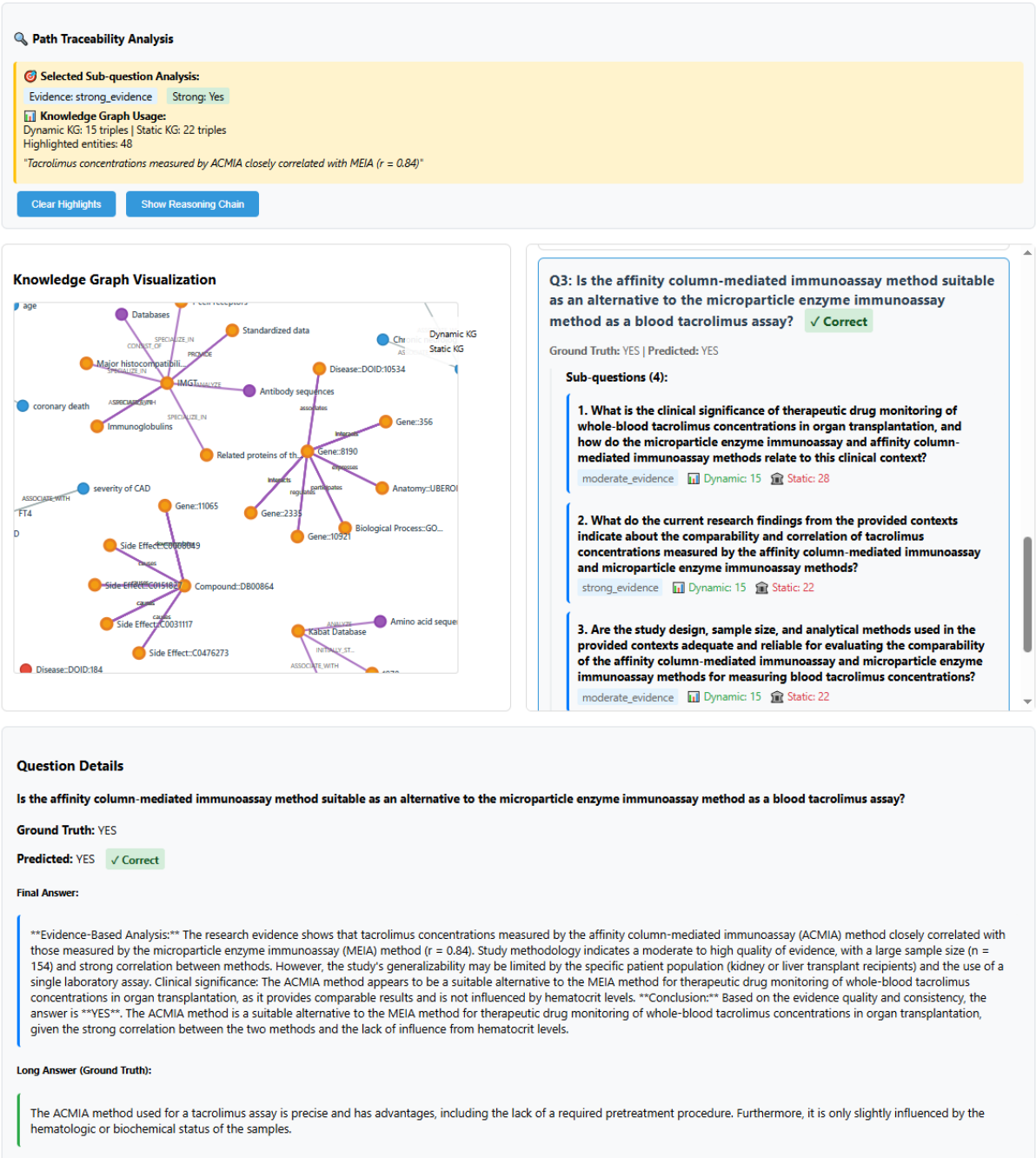

Figure 4: Dual-KG framework user interface for biomedical question answering. The interface displays path traceability analysis with both dynamic (15 triples) and static (22 triples) knowledge graphs, sub-question decomposition with evidence typing, and final answer synthesis. The example shows a tacrolimus assay comparison question with correct prediction (Ground Truth: YES, Predicted: YES).

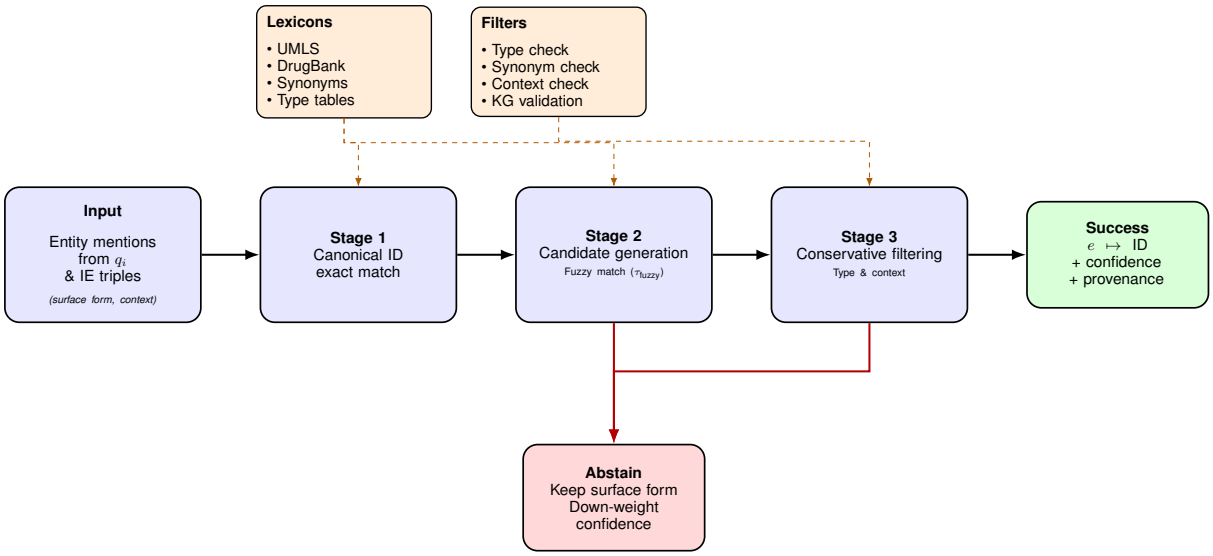

Figure 5: **Entity linking and canonicalization in DUAL-KG.** Mentions from sub-questions and IE triples are mapped to canonical biomedical IDs via staged resolution: exact ID matching, conservative fuzzy candidate generation, and type/context filtering. Ambiguous or failed links trigger abstention and confidence down-weighting to reduce error propagation into graph fusion.

evidence is absent or ambiguous, label is "unclear". Select ONE snippet of ≤25 words from context. Return one JSON object: {"label": "yes|no|unclear", "quote": "≤25 words exact context snippet", "strong": true|false}

Prompt 7: KG-based Sub-question Answering

System: You are a biomedical evidence analyst. Return exactly ONE valid JSON object: {"label": "yes|no|unclear|no information retrieved", "quote": "≤25-word snippet from SOURCE A", "strong": true|false, "kg_used": [{"subj":"...", "rel":"...", "obj":"..."}]}. Use EXACTLY "rel" as the key name for relations. kg_used must contain ONLY the top 3-5 most relevant triples.

Prompt 8: Final Answer Generation (PubMedQA)

System: Main question: "{main_question}". Here are relevant evidence snippets: {evidence_snippets}. Based on the evidence analysis, the computed answer is: {computed_answer}. Your answer MUST start with exactly one word: {computed_answer}. Then write 1-2 short sentences justifying this answer.

Prompt 9: Final Answer Generation (BioASQ)

System: Main question: "{main_question}". Here are relevant evidence snippets from biomedical literature: {evidence_snippets}. Provide a comprehensive and factual answer. Write 2-4 complete sentences that directly answer the question. Include specific details, mechanisms, or findings from the evidence. Use appropriate biomedical terminology.

Prompt 10: Final Answer Generation (MedHop)

System: Main question: "{main_question}". Here are relevant evidence snippets about drug interactions: {evidence_snippets}. Available answer choices: {answer_choices}. This is a drug interaction question requiring identification of a specific drug code. Choose ONE drug code from the available answer choices. End with the exact drug code in the format: DB##### (exactly 7 characters).