ReGraM: Region-First Knowledge Graph Reasoning for Medical Question Answering

Anonymous ACL submission

Abstract

Recent studies in medical question answering (Medical QA) have actively explored the integration of large language models (LLMs) with biomedical knowledge graphs (KGs) to improve factual accuracy. However, most existing approaches still rely on traversing the entire KG or performing large-scale retrieval, which introduces substantial noise and leads to unstable multi-hop reasoning. We argue that the core challenge lies not in expanding access to knowledge, but in identifying and reasoning over the appropriate subset of evidence for each query. **ReGraM** is a region-first knowledge graph reasoning framework that addresses this challenge by constructing a query-aligned subgraph and performing stepwise reasoning constrained to this localized region under multiple evidence-aware modes. By focusing inference on only the most relevant portion of the KG, ReGraM departs from the assumption that all relations are equally useful—an assumption that rarely holds in domain-specific medical settings. Experiments on seven medical QA benchmarks demonstrate that ReGraM consistently outperforms a strong baseline (KGARevion), achieving an 8.04% absolute accuracy gain on MCQ, a 4.50% gain on SAQ, and a 42.9% reduction in hallucination rate. Ablation and qualitative analyses further show that aligning region construction with hop-wise reasoning is the primary driver of these improvements. Overall, our results highlight region-first KG reasoning as an effective paradigm for improving factual accuracy and consistency in medical QA.

1 Introduction

Medical question answering (Medical QA) increasingly requires multi-hop reasoning that connects symptoms, mechanisms, diagnoses, treatments, and molecular interactions (Xie et al., 2024; Singhal et al., 2025; Wu et al., 2025).

Despite this growing demand, most existing LLM-based Medical QA systems still rely on shallow, retrieval-based reasoning. Even methods designed for multi-hop inference often lack explicit structural constraints, causing reasoning chains to drift across unrelated biomedical concepts. As a result, clinical evaluations frequently report medical hallucinations—plausible but unsupported statements that pose serious safety risks in real-world applications (Saab et al., 2024; Williams et al., 2024). To address these issues, recent work has integrated large language models with biomedical KGs through KG-based QA, KG-RAG, and graph-constrained prompting (Zeng et al., 2025; Sui et al., 2025; Shi et al., 2025; Long et al., 2024). Methods such as KGARevion further apply post-hoc verification of LLM-generated triplets against a KG (Jeong et al., 2024). However, most KG-based approaches still treat the KG as a flat search space spanning heterogeneous biomedical subdomains and relation types (Chandak et al., 2023). Because typical queries require only a small, task-relevant subset of the KG, global traversal introduces substantial retrieval noise that accumulates over multi-hop reasoning and destabilizes inference (Su et al., 2024; Jeong et al., 2024; Wu et al., 2025).

We propose a different design principle: rather than exploring the entire KG prior to reasoning, we first construct a semantically aligned subgraph tailored to the query and constrain all subsequent reasoning within that region. Unlike post-hoc verification or soft context filtering, this region is treated as a hard structural boundary throughout multi-hop inference, preventing semantically irrelevant relations from intervening once reasoning begins. We instantiate this principle in a framework called **ReGraM**, which performs query typing, region selection, evidence-aware reasoning, and verification entirely within a localized KG region. This design enforces early restriction of the reasoning space and promotes factual consistency by structurally excluding irrelevant relations.

We evaluate ReGraM on PrimeKG using the

same entity set and relation schema as KGARevion to ensure a controlled comparison. To improve region construction, we refine the relation description map used for retrieval and prompting, without modifying the underlying KG.

Although evaluated in the medical domain, the region-first principle generalizes to other structured domains with high relational complexity. Our main contributions are as follows:

1. We propose **ReGraM**, a region-first KG reasoning framework that constructs a query-aligned subgraph and performs reasoning via multi-hop decomposition within it. This design structurally constrains the reasoning space and mitigates semantic drift by excluding irrelevant relations.
2. We demonstrate substantial empirical gains: an 8.04% improvement in MCQ accuracy, a 4.50% improvement in SAQ accuracy, a 32–48% reduction in hallucination rates, and a 46% reduction in inference latency across seven medical QA benchmarks and multiple hallucination testbeds.
3. We conduct ablation and error analyses showing that region selection and multi-hop decomposition are critical to factual consistency, and we validate the contribution of each module to overall system stability.

2 Related Work

2.1 LLM-based Reasoning and Multi-hop Inference

Medical question answering (MedQA) requires multi-hop reasoning that connects diverse biomedical concepts in a stepwise manner, which is challenging for single-step retrieval-based approaches. To address this challenge, a wide range of LLM-based reasoning methods have been proposed, including Chain-of-Thought and its variants, ReAct, Reflexion, SelfCheck, and STaR (Wei et al., 2022; Yao et al., 2022; Shinn et al., 2023; Manakul et al., 2023; Zelikman et al., 2024). However, these methods typically rely on the LLM's internal knowledge without external structural constraints, often leading to reasoning drift, cumulative errors, and factual hallucinations in complex biomedical scenarios (Bang et al., 2023; Huang et al., 2025).

In particular, many MedQA tasks require connecting multiple entities across relational chains. LLMs alone struggle to maintain consistency across these hops, highlighting the need to incorporate external structured knowledge—such as knowledge graphs (KGs)—to provide relational grounding and guide multi-hop inference (Chen et al., 2019; Luo et al., 2023; Sakarvadia, 2024).

2.2 KG-based Reasoning for Medical QA

To improve factual grounding, a variety of models have explored knowledge graph-based QA in the biomedical domain, including QAGNN (Yasunaga et al., 2021), JointLK (Sun et al., 2022), KagNet (Lin et al., 2019), MedReason (Wu et al., 2025), and ontology-guided KGQA (Liu et al., 2025). While effective in incorporating KG structure, these methods typically operate over the entire graph without explicitly tailoring the reasoning space to the query context.

In large-scale biomedical KGs such as PrimeKG, full-graph traversal introduces irrelevant relations, propagates retrieval noise, and fragments evidence due to sparsity (Zhai et al., 2024; Mavi et al., 2023; Chandak et al., 2023). Some recent methods address this issue by incorporating verification mechanisms (e.g., KGARevion (Su et al., 2024)) or by organizing retrieval contexts using graph structure (e.g., GraphRAG (Xiao et al., 2025) and MedGraphRAG (Wu et al., 2025)). However, these approaches often treat KG retrieval as a preprocessing step, leaving the subsequent reasoning process unconstrained beyond the retrieved context. As a result, retrieval noise and semantically irrelevant relations can still influence downstream reasoning, especially in multi-hop and generative settings. This issue is particularly pronounced in generative QA, where early retrieval errors can propagate and compound across reasoning steps.

In contrast, our work imposes an explicit structural boundary *before* reasoning begins. Rather than applying KG pruning post hoc or relying on soft context filtering, we define a query-specific subgraph as the primary reasoning space, ensuring that multi-hop inference remains structurally grounded throughout the entire process.

3 Method

Figure 1 presents an overview of ReGraM. ReGraM is a region-first knowledge graph reasoning framework that constructs a query-aligned subgraph and performs multi-hop reasoning under a hard region constraint.

This section describes the overall framework and its core components. Additional implementation details, prompt templates, and qualitative examples

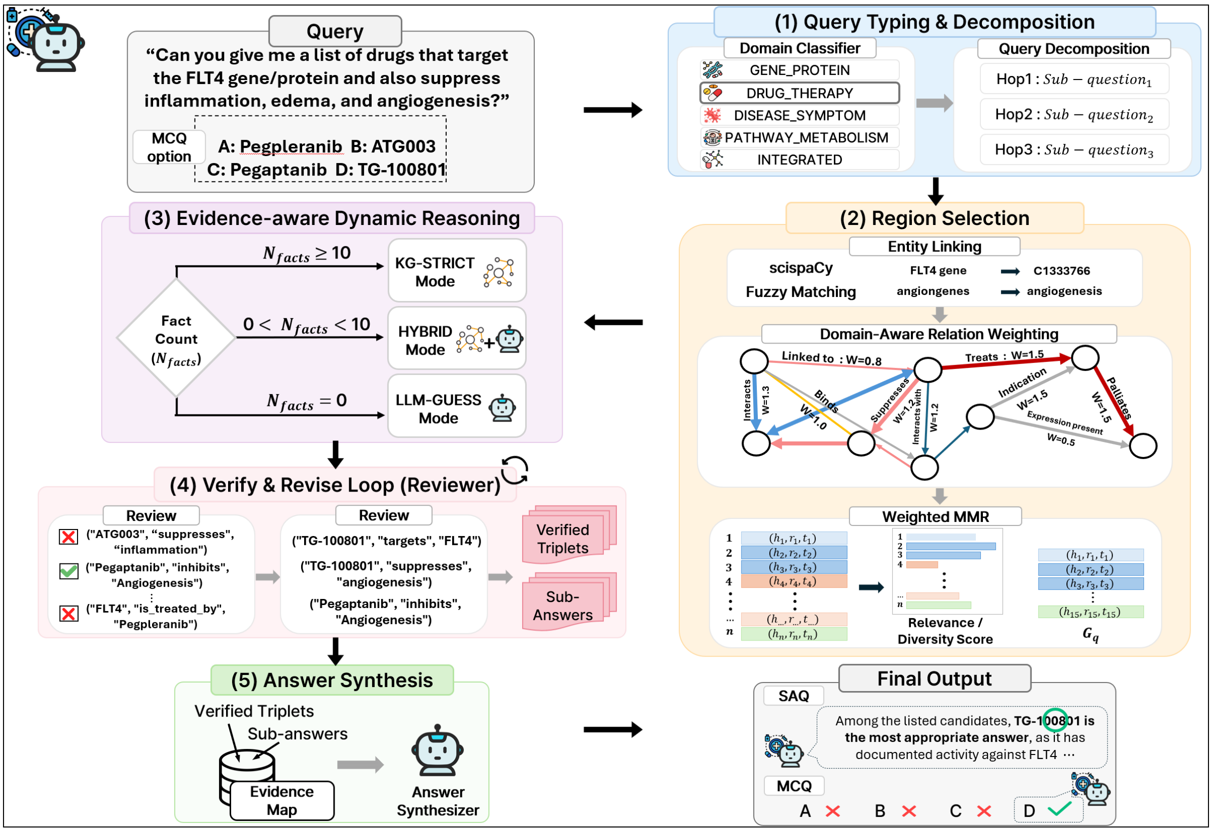

Figure 1: Overview of the ReGraM framework. Given a medical query, ReGraM constructs a query-aligned local knowledge graph region and performs reasoning exclusively within the selected region to generate the final answer.

are provided in the appendix.

3.1 Query Typing and Decomposition

ReGraM begins by analyzing the structure of the input query to guide subsequent region construction. Given an input query Q, the model first classifies it into one of five high-level medical domains (`GENE/PROTEIN`, `DRUG/THERAPY`, `DISEASE/SYMPTOM`, `PATHWAY/METABOLISM`, or `INTEGRATED`) using an LLM-based classification prompt. The predicted domain serves as a semantic prior for adjusting relation importance during KG region selection.

The query is then decomposed into up to three sub-questions using a one-shot CoT prompt, aligning the reasoning process with the underlying multi-hop structure (Appendix A and Figure 3).

3.2 Region Selection

For each sub-question q_i, ReGraM constructs a localized KG region rather than retrieving from the full graph. Medical entities are extracted using the UMLS linker in SciSpaCy (Neumann et al., 2019; Bodenreider, 2004) and expanded into an enriched entity set E_{exp} via UMLS synonyms, custom alias dictionaries curated from PrimeKG node attributes, and fuzzy string matching (similarity ≥ 90) with RapidFuzz (?).

Candidate triplets T_{cand} connected to E_{exp} are collected from the KG and reweighted using domain-specific relation weights $w_r \in \{0.5, 0.8, 1.0, 1.2, 1.5\}$, which encode the relevance of each relation type to the predicted query domain. To prioritize both semantic relevance and evidence diversity, triplets are re-ranked using a weighted Maximal Marginal Relevance (MMR) criterion (Carbonell and Goldstein, 1998):

$$\begin{aligned}\mathrm{MMR}(t) = {} & \lambda\,\mathrm{Sim}(q_i,t)\,w_r \\ & - (1-\lambda)\max_{t'\in S}\mathrm{Sim}(t,t')\end{aligned} \quad (1)$$

Here, $\mathrm{Sim}(\cdot,\cdot)$ denotes the cosine similarity between Sentence-BERT embeddings (`all-MiniLM-L6-v2`) of the sub-question q_i and the textual representation of a triplet t (Reimers and Gurevych, 2019). The second term penalizes redundancy by discouraging selection of triplets that are semantically similar to previously selected ones, where S denotes the set of triplets already chosen for the current region. The hyperparameter $\lambda = 0.7$ balances relevance and diversity during region construction.

The top-K ranked triplets (with $K = 15$) constitute the query-specific KG region $G_q = (V_q, E_q)$, where E_q is the selected triplet set and V_q contains all head and tail entities appearing in E_q. All subsequent knowledge access operations during reasoning are restricted to this region, formally denoted as $\text{Lookup}(q_i) \subseteq G_q$. This restriction enforces region-constrained reasoning by preventing the model from accessing relations outside the query-aligned subgraph (see Appendix B).

ReGraM does not modify the underlying KG; performance gains stem from structurally bounding inference within a query-aligned subgraph. This design enables the model to more effectively interpret and utilize the semantic structure already encoded in PrimeKG, for example by distinguishing clinically relevant relations from generic associations during region construction.

3.3 Evidence-aware Dynamic Reasoning

ReGraM adapts its reasoning behavior based on the amount of retrieved evidence. We unify the notion of hop count across modules by fixing the maximum number of reasoning steps to $H = 3$, implemented as up to three decomposed sub-questions. No additional graph-walk or in-region traversal depth is introduced beyond these steps; all reasoning is organized within this fixed hop budget.

Let N_{facts} denote the number of region triplets available for a given sub-question. Based on N_{facts}, the model dynamically switches among three reasoning modes. Importantly, even when hypothesis generation is permitted, reasoning remains constrained. Across all modes, ReGraM operates under a closed-world assumption: entities must belong to V_q, and relations must be selected from a predefined schema set $\mathcal{R}$.

KG-STRICT relies solely on region evidence when sufficient facts are available, while **LLM-GUESS** falls back to parametric knowledge when no region evidence exists. **HYBRID** performs constrained hypothesis completion using only in-region entities and schema-allowed relations, followed by Reviewer verification, enabling limited generalization when KG evidence is partial but non-empty. All reasoning steps are executed strictly within the selected region G_q, with the total number of reasoning steps fixed to $H = 3$ for stability.

3.4 Verify & Revise Loop

In **HYBRID** mode, ReGraM employs a verification mechanism to control hallucinated or inconsistent hypotheses. Generated triplets are evaluated by a LoRA-tuned Llama-3.1-8B binary classifier, referred to as the **Reviewer** (Hu et al., 2022; Grattafiori et al., 2024). The Reviewer assigns a score s_θ to each triplet and accepts it as valid if $s_\theta \geq 0.5$.

Triplets rejected by the Reviewer are not discarded immediately; instead, they are passed to a revision prompt that attempts to correct semantic or schema-level inconsistencies. The revised triplets are then re-evaluated by the Reviewer. This verification–revision loop runs for up to two iterations, and only Reviewer-approved triplets are incorporated into the reasoning evidence (see Appendix A and Figure 5).

3.5 Answer Synthesis

Throughout reasoning, ReGraM stores intermediate sub-answers and supporting evidence in an **Evidence Map**. Final answer generation is performed using mode-specific synthesis templates: **KG-STRICT** produces concise factual answers grounded solely in region evidence; **HYBRID** generates stepwise explanations based on Reviewer-verified triplets; and **LLM-GUESS** explicitly indicates missing KG evidence while providing a plausible answer.

The **Answer Synthesizer** integrates hop-level outputs into a coherent final response grounded in the accumulated evidence (see Appendix A, Figures 6–9). Taken together, ReGraM's region-first design separates *where* to reason (region construction) from *how* to reason (evidence-aware inference), allowing reasoning behavior to adapt to evidence availability.

4 Experiments

This section analyzes how region-first KG reasoning affects medical QA accuracy, hallucination robustness, and reasoning efficiency. All experiments compare ReGraM with the baseline KGARevion under identical settings. Both methods operate on the same biomedical knowledge graph, PrimeKG, with the same entity set and relation schema.

4.1 Experimental Setup

Both ReGraM and KGARevion are based on `Llama-3.1-8B` and operate on the same biomedical

Task	Metric	Model	Basic	Inter.	Expert	MedQA	Afrimed	MMLU	PubMed	Avg
MCQ	ACC ↑	KGARevion	38.8	35.0	33.8	51.7	49.9	62.8	54.8	46.2
		ReGraM	**47.1**	**45.0**	**45.9**	**54.4**	**56.5**	**65.7**	**64.9**	**54.2**
		Δ	+8.3	+9.9	+12.1	+2.7	+6.6	+2.9	+10.1	+8.0
SAQ	ACC ↑	KGARevion	10.2	13.9	14.8	36.5	26.1	65.7	53.5	31.5
		ReGraM	**11.3**	**13.9**	**20.8**	**51.0**	**47.1**	65.5	44.7	**36.0**
		Δ	+1.1	0.0	+6.0	+14.5	+21.0	-0.3	-8.8	+4.5
	IS ↓	KGARevion	57.5	53.8	57.6	52.9	23.7	41.8	36.0	46.2
		ReGraM	**30.5**	**32.0**	**30.1**	**28.2**	**16.1**	**23.8**	**24.4**	**26.4**
		Δ_{IS}	27.0	40.5	47.7	46.7	32.1	43.1	32.2	42.9

Table 1: Performance of **ReGraM** and KGARevion on seven medical QA benchmarks. ACC denotes accuracy (%), and IS denotes the hallucination inconsistency score evaluated only for SAQ (lower is better). Δ indicates absolute accuracy improvement (percentage points), and Δ_{IS} denotes relative reduction for IS (%), computed as $(\text{IS}_{\text{KGARevion}} - \text{IS}_{\text{ReGraM}})/\text{IS}_{\text{KGARevion}} \times 100$.

knowledge graph, PrimeKG, under identical training and inference settings. No additional knowledge graphs, entities, relations, or edges are introduced in ReGraM.

The key difference lies in how relational knowledge is utilized. ReGraM extends the relation description map originally used in KGARevion to more comprehensively cover the diverse relation types present in PrimeKG. This extension does not alter the knowledge available to the model, but ensures that domain-relevant relations are not systematically underweighted when the reasoning space is explicitly constrained. As a result, region construction more accurately aligns retrieved triplets with the query intent, while preserving a fully controlled KG setting. Performance gains in ReGraM are not attributable to increased knowledge coverage, as the underlying knowledge graph remains unchanged. Instead, improvements stem from semantically interpreting and structurally constraining existing PrimeKG relations during region construction and reasoning.

Traditional KG-QA models such as QAGNN, JointLK, and KagNet assume supervised QA training and differ fundamentally from our generative KG-alignment objective. ReGraM is positioned as a structural extension of KGARevion under identical training and inference settings; accordingly, we focus on a controlled comparison to isolate the effect of region-first bounding. Broader comparisons to KG-RAG or GraphRAG-style baselines are left to future work. Experiments were conducted on seven medical QA benchmarks: MedDDx-Basic, MedDDx-Intermediate, MedDDx-Expert, MedQA, PubMedQA, MMLU-Medical, and AfrimedQA. For readability, we refer to the three MedDDx variants as Basic, Intermediate, and Expert in tables and figures. Overall, region-first reasoning improves stability under the semantic diversity of PrimeKG relations.

4.2 Evaluation Protocols

We evaluate model performance along two axes: **accuracy** and **hallucination robustness**. Standard medical QA benchmarks are assessed under both multiple-choice (MCQ) and short-answer (SAQ) settings, while hallucination robustness is evaluated using the MedHallu and Med-HALT benchmarks.

Standard Medical QA Benchmarks (MCQ & SAQ) In the MCQ setting, each model receives a question with candidate options and outputs an answer choice, and we report exact-match accuracy. In the SAQ setting, the model generates an open-ended answer. Because embedding similarity alone does not reliably capture biomedical correctness, we employ an LLM-as-a-Judge protocol (Badshah and Sajjad, 2025; Chang et al., 2024) using GPT-4o to assess semantic alignment between generated and reference answers. Following common practice, outputs with a similarity score of at least 0.8 are treated as correct, and the same judge and threshold are used for both methods to ensure a controlled comparison. Evaluation prompts are provided in Appendix A and Figure 10.

Hallucination Benchmarks (MedHallu, Med-HALT) MedHallu (Huang et al., 2025) measures factual accuracy, hallucination rate, and reasoning error rate in free-form generation. Med-HALT (Wu et al., 2025), on the other hand, evaluates model stability under adversarial conditions through three sub-tests: Fake Question Test (FQT), None-of-the-Above (NOTA), and False Confidence Test (FCT).

Our LLM-based inconsistency analysis and MedHallu evaluation follow the SAQ (open-ended) setting, while Med-HALT adopts the MCQ (multiple-

Domain	KG Coverage (%)	Mean IS
Drug_Therapy	68.14	5.08
Gene_Protein	60.30	5.97
Disease_Symptom	15.66	7.16
Pathway_Metab.	26.31	5.42
Integrated	83.52	6.83

Table 2: Domain-wise hallucination inconsistency score (IS) vs KG coverage. Higher coverage generally leads to lower IS.

Metric	KGARevion	ReGraM
MH-ACC	10.00	**43.50**
MH-Sim	0.00	**54.70**
MH-Hallu	90.00	**89.00**
MH-RError	100.00	**72.00**

Table 3: MedHallu results under SAQ. MH-ACC (↑): accuracy on a 0–100 scale. MH-Sim (↑): semantic similarity to the ground-truth answer on a 0–100 scale. MH-Hallu (↓) and MH-RError (↓): hallucination and reasoning error rates (%).

choice) discriminative format. Evaluation prompts and parsing rules are provided in Appendix A and Figure 12.

For MedHallu's semantic alignment score (MH-Sim), we use GPT-4o to measure the similarity between the generated answer and the ground-truth answer on a 0–100 scale, and report the average score across samples.

4.3 Main Results on Medical QA Performance

Table 1 summarizes the comparative performance of ReGraM and KGARevion across seven medical QA benchmarks under both multiple-choice (MCQ) and short-answer (SAQ) settings. We report absolute accuracy improvements (Δ, percentage points) as well as relative reductions in inconsistency score (Δ_{IS}, percentage). Overall, ReGraM achieves consistent gains over the baseline, improving MCQ accuracy by 8.04% and SAQ accuracy by 4.50%, with particularly strong improvements on datasets requiring intensive multi-hop reasoning.

ReGraM shows the largest gains on the MedDDx benchmarks (Basic, Intermediate, and Expert), all of which involve compositional reasoning across multiple biomedical entities.

In the SAQ setting, ReGraM yields substantial improvements on MedQA (14.50%) and AfrimedQA (21.04%), both of which involve complex clinical scenarios that demand precise evidence alignment. These results indicate that integrating semantic region selection with dynamic reasoning modes effectively improves open-ended answer generation.

A performance trade-off is observed on PubMedQA (-8.85%), where KG evidence is sparse or weakly aligned with question semantics. This behavior reflects a design trade-off rather than a failure case: in datasets with limited KG coverage, aggressive region restriction and conservative verification can under-retrieve relevant evidence needed for answer generation.

Results in Table 1 are averaged over three runs (mean ± std). Observed improvements consistently exceed baseline variance, supporting the effectiveness of the region-first design in enhancing accuracy and compositional reasoning across diverse medical QA tasks.

4.4 Hallucination Evaluation Results

We evaluate hallucination robustness using IS (SAQ), MedHallu (generative), and Med-HALT (adversarial MCQ).

Dataset-wise Hallucination Reduction (IS) Table 1 shows factual hallucination inconsistency scores (IS) across datasets under the SAQ setting, scored using an LLM-based inconsistency evaluator (see Appendix A). ReGraM reduces the average IS from 46.2 to 26.4, achieving a 42.9% reduction overall. The largest reductions are observed in multi-hop–intensive datasets such as MedDDx-Basic (–47.0%), MedDDx-Intermediate (–40.5%), and MedDDx-Expert (–47.8%), demonstrating that ReGraM is particularly effective where multi-step reasoning is required. This indicates that region bounding reduces drift, and the Reviewer suppresses unsupported steps.

Domain-wise Hallucination Robustness Table 2 reports mean hallucination inconsistency scores (IS) across five biomedical domains under the SAQ setting. Overall, higher KG coverage tends to correspond to lower hallucination rates; for instance, the Drug_Therapy domain exhibits the lowest IS. However, this relationship is not uniform across domains. Notably, the `Integrated` domain shows relatively high IS despite having the highest KG coverage.

We further conduct a finer-grained analysis in Appendix C.7, demonstrating that hallucination robustness also depends on additional structural factors, including relation granularity, prompt design, and evidence redundancy.

MedHallu: Free-form Generative Hallucination Table 3 presents results on the MedHallu benchmark. While KGARevion hallucinates in nearly all generated answers, ReGraM significantly improves factual accuracy, semantic similarity, and

Dataset	Type	ReGraM	w/o Domain Prior	Δ (prior)	w/o Multihop	Δ (multihop)
MedDDx-Expert	MCQ	45.88	39.34	+6.54	41.41	+3.03
MedQA	MCQ	54.38	56.48	−2.10	55.38	−1.00
MMLU-Med	MCQ	65.72	67.77	−2.05	67.95	−2.23
PubMedQA	MCQ	64.86	67.30	−2.44	67.30	−5.01
MedDDx-Expert	SAQ	20.81	17.18	+3.63	18.43	+2.38
MedQA	SAQ	51.03	2.91	+48.12	19.80	+29.28
MMLU-Med	SAQ	65.45	5.33	+60.12	31.86	+33.59
PubMedQA	SAQ	44.65	41.90	+2.75	42.30	+2.35

Table 4: Ablation results for ReGraM (MCQ and SAQ). Domain prior and multi-hop reasoning are both essential for maintaining factual consistency.

Metric	KGARevion	ReGraM
FQT-ACC	22.00	**24.80**
FQT-Hallu	78.00	**76.50**
NOTA-ACC	22.00	**19.67**
NOTA-Hallu	78.00	**80.33**
FCT-ACC	22.00	**25.19**
FCT-Hallu	78.00	**74.81**

Table 5: MedHALT results. ACC (↑): accuracy. Hallu (↓): hallucination rate.

Metric	ReGraM	w/o Reviewer
SAQ Accuracy (avg.) ↑	**36.03**	33.17
MedHallu Accuracy ↑	**43.50**	0.00
MedHallu Similarity ↑	**54.70**	0.00
MedHallu Hallucination ↓	**89.0**	100.0
MedHallu Reasoning Error ↓	**72.0**	100.0
Med-HALT FQT Hallu ↓	**74.81**	97.67
Med-HALT FCT Accuracy ↑	25.19	**33.67**

Table 6: Reviewer ablation. ReGraM vs. *w/o Reviewer* on QA accuracy and hallucination benchmarks. Metrics are computed under our GPT-4o–based automatic evaluation protocol on a 0–100 scale (higher is better for accuracy/similarity; lower is better for hallucination/error rates, ↓).

reasoning reliability. These results demonstrate ReGraM's effectiveness in suppressing hallucinations under free-form generative settings.

Med-HALT: Adversarial Discriminative Hallucination Table 5 presents performance on Med-HALT, which probes hallucination under adversarial conditions through Fake-Question (FQT), None-of-the-Above (NOTA), and False-Confidence Test (FCT). ReGraM outperforms KGARevion in the FCT sub-test (+3.19 ACC), indicating stronger factual discrimination under confidence-sensitive evaluation. However, performance slightly degrades on FQT and NOTA. This can be attributed to ReGraM's region-first design: while it discourages unsupported extrapolation, it also biases the model toward selecting the most plausible answer *within* the retrieved region. As a result, in adversarial settings where the correct response is to abstain (e.g., NOTA), ReGraM may over-commit to weak in-region evidence. Overall, these results highlight a fundamental trade-off between hallucination suppression and abstention under adversarial inputs.

4.5 Efficiency Analysis

ReGraM reduces MCQ latency by 46% via region-first design. SAQ latency slightly increases due to verification. Full results are in Appendix C.5.

4.6 Ablation Studies

To analyze how each structural component contributes to performance, we conducted ablation studies targeting the region selector, multi-hop decomposition, region refinement (MMR), reasoning depth, and their interactions.

Effect of Domain Prior and Multi-hop Reasoning Table 4 compares the impact of removing two core components of the region-first design: the *domain prior* and *multi-hop decomposition*. Both ablation settings retain the region selector but deactivate either the semantic domain prior used during region construction or the hop-by-hop decomposition used in reasoning.

Under the MCQ setting, removing either component results in relatively mild changes. In contrast, under the SAQ setting, both ablations lead to substantial performance degradation. In particular, **w/o Domain Prior** constructs regions without domain-level semantic guidance, which weakens relation reweighting during region selection and allows semantically irrelevant triplets to dominate the selected region. As a result, the evidence presented to the reasoner becomes noisy and internally inconsistent, leading to sharp accuracy drops on MedQA (51.03 → 2.91) and MMLU-Med (65.45 → 5.33).

Similarly, the **w/o Multihop** setting shows pronounced degradation on multi-step reasoning tasks such as MedQA (51.03 → 19.80) and MMLU-Med (65.45 → 31.86), suggesting that single-step reasoning fails to reconstruct compositional dependencies across medical concepts.

Overall, these results indicate that the factual stability and reasoning accuracy of ReGraM are jointly supported by (1) the **domain prior**, which stabilizes region construction by suppressing semantically irrelevant evidence, and (2) the **multi-hop reasoning** mechanism, which enables structured inference over interdependent medical concepts.

Reviewer Ablation: Accuracy vs. Stability Trade-off Table 6 summarizes the effect of removing the Reviewer. SAQ Accuracy (avg.) is averaged over four SAQ benchmarks (MedDDx-Expert, MedQA, MMLU-Med, and PubMedQA). Without the Reviewer, ReGraM experiences catastrophic failure on MedHallu (0% accuracy, 100% hallucination), while performance on Med-HALT becomes inconsistent, hallucination increases, but some discriminative accuracy improves (e.g., FCT accuracy). These results highlight a central insight: the Reviewer enforces conservative generation behavior that reduces hallucination, but may under-answer when KG evidence is limited. Such discriminative accuracy gains come at the cost of substantially increased hallucination rates, underscoring the trade-off between discriminative confidence and factual reliability.

Interaction Between Domain Prior and Multi-hop Reasoning Removing either the domain prior or multi-hop decomposition significantly reduces accuracy, and removing both leads to catastrophic failure, as shown in Appendix C.5.

Region Refinement and Hop-depth Ablation results on MMR and hop-depth variation are provided in Appendix C.2. The region refinement improves SAQ stability by reducing redundancy, and a hop depth of 3 yields optimal factuality-performance trade-offs.

5 Discussion

ReGraM suggests that instability in medical QA is not only a modeling issue but also a *structural* one. In particular, how and when the reasoning space is bounded can be as important as the choice of the underlying language model. Reasoning over a heterogeneous KG without explicit boundaries allows irrelevant relations to intervene, amplifying semantic drift in multi-hop and generative settings. By constructing a query-aligned region *before* reasoning, ReGraM effectively bounds the reasoning space and improves robustness under both MCQ and SAQ settings.

We observe a tension between precision and completeness: strict region bounding may lead to under-answering in low-coverage settings, whereas dense domains benefit more consistently. This trade-off is also reflected in adversarial settings (e.g., NOTA), where strict in-region evidence can bias the model toward a plausible in-region option unless an explicit abstention mechanism is introduced. These limitations and broader evaluation caveats are discussed in Section 7.

6 Conclusion

This work introduced **ReGraM**, a region-first KG reasoning framework for improving stability and factuality in medical QA. ReGraM constructs query-aligned local KG regions and performs reasoning within a bounded region, mitigating inefficiency and semantic drift in heterogeneous biomedical KGs. Across seven medical QA benchmarks, ReGraM improves accuracy over KGARevion by 8.04 points in MCQ and 4.50 points in SAQ, while also reducing hallucinations and improving inference efficiency. Notably, these gains are achieved using a lightweight 8B-scale language model without task-specific fine-tuning, highlighting the practical efficiency and deployability of region-first reasoning in resource-constrained medical settings. ReGraM remains robust under free-form generation and empirically suggests improved scalability by maintaining a bounded reasoning space during inference, although formal guarantees are left for future work. By treating the KG as a structural component that shapes reasoning, ReGraM enables more controllable inference; future work will extend region-first reasoning to better handle answer-absent queries via explicit abstention strategies.

7 Limitations

Despite the strong empirical performance of ReGraM, several limitations remain and warrant careful consideration.

Reliance on automatic evaluation protocols Our evaluation primarily relies on LLM-based automatic judges rather than clinician-centered assessment. While this choice enables scalable and consistent benchmarking across diverse datasets, it cannot fully capture clinical appropriateness, safety, or decision validity in real-world medical contexts. Moreover, the reported inconsistency score (IS) is computed using a fixed scorer prompt with deterministic decoding, and its absolute value may vary depending on the evaluator model or prompt design. Accordingly, IS should be interpreted primarily as a relative comparison under a fixed evaluation protocol. In future work, we plan to incorporate clinician-in-the-loop evaluation and scenario-based expert review to more comprehensively assess clinical reliability (Livingston et al., 2025).

Dependence on relation semantic representations ReGraM relies on PrimeKG as its sole knowledge base and does not introduce new entities or relations. The effectiveness of region selection therefore depends on how well relation semantics are captured by the relation description map. Although we expand and refine this map to better reflect PrimeKG's relational diversity, some fine-grained or context-dependent relations may still be underrepresented. This limitation highlights an open challenge in representing domain semantics when reasoning is explicitly constrained to a localized region. Improving automatic or learned relation semantic representations remains an important direction for future work.

Challenges in evidence aggregation at the decision stage Even when ReGraM successfully accumulates relevant evidence through region selection and multi-hop reasoning, final answer selection can occasionally become misaligned with the supporting evidence (Appendix C.9, Example 2), particularly in discrete option selection scenarios. This observation suggests that robustness in medical QA depends not only on retrieving and verifying high-quality evidence, but also on how such evidence is aggregated, calibrated, and translated into final decisions. Improving evidence-aware aggregation and decision calibration remains an important direction for future research (Wen et al., 2025).

Limited scalability analysis ReGraM empirically reduces inference latency by reasoning over compact, query-aligned regions rather than the full knowledge graph. However, we do not provide a formal asymptotic analysis with respect to KG size. Consequently, observed efficiency gains may vary depending on KG structure, domain distribution, and reasoning depth. A more rigorous theoretical and system-level analysis of scalability is left for future work (Chen et al., 2025).

8 Ethics Statement

This work aims to improve the factual consistency of medical question answering systems by constraining reasoning over a biomedical knowledge graph. All experiments were conducted on publicly available benchmark datasets using automatic evaluation protocols, and no private or patient-specific data were used.

The generated outputs are not intended for clinical use without human oversight, and the system should not be deployed in real-world diagnostic settings without expert review, as medical inaccuracies may still occur.

An anonymized version of the code and additional materials is available at `https://anonymous.4open.science/r/ReGraM-6B41/`, and will be fully released upon acceptance.

References

Sher Badshah and Hassan Sajjad. 2025. Reference-guided verdict: Llms-as-judges in automatic evaluation of free-form qa. In *Proceedings of the 9th Widening NLP Workshop*, pages 251–267.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

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

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

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

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Jialin Chen, Houyu Zhang, Seongjun Yun, Alejandro Mottini, Rex Ying, Xiang Song, Vassilis N Ioannidis, Zheng Li, Qingjun Cui, and Yale University Amazon. 2025. Gril: Knowledge graph retrieval-integrated learning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16306–16319.

Jifan Chen, Shih-ting Lin, and Greg Durrett. 2019. Multi-hop question answering via reasoning chains. *arXiv preprint arXiv:1910.02610*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

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.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Runxuan Liu, Bei Luo, Jiaqi Li, Baoxin Wang, Ming Liu, Dayong Wu, Shijin Wang, and Bing Qin. 2025. Ontology-guided reverse thinking makes large language models stronger on knowledge graph question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15269–15284, Vienna, Austria. Association for Computational Linguistics.

Leah Livingston, Amber Featherstone-Uwague, Amanda Barry, Kenneth Barretto, Tara Morey, Drahomira Herrmannova, and Venkatesh Avula. 2025. Reproducible generative artificial intelligence evaluation for health care: a clinician-in-the-loop approach. *JAMIA Open*, 8(3):ooaf054.

Cui Long, Yongbin Liu, Chunping Ouyang, and Ying Yu. 2024. Bailicai: A domain-optimized retrieval-augmented generation framework for medical applications. *arXiv preprint arXiv:2407.21055*.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Vaibhav Mavi, Abulhair Saparov, and Chen Zhao. 2023. Retrieval-augmented chain-of-thought in semi-structured domains. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 178–191, Singapore. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Mansi Sakarvadia. 2024. Towards interpreting language models: A case study in multi-hop reasoning. *arXiv preprint arXiv:2411.05037*.

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.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, and Marinka Zitnik. 2024. Kgarevion: an ai agent for knowledge-intensive biomedical qa. *arXiv preprint arXiv:2410.04660*.

Yuan Sui, Yufei He, Nian Liu, Xiaoxin He, Kun Wang, and Bryan Hooi. 2025. Fidelis: Faithful reasoning in large language models for knowledge graph question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8315–8330.

Yueqing Sun, Qi Shi, Le Qi, and Yu Zhang. 2022. JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5049–5060, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556.

Christopher YK Williams, Brenda Y Miao, Aaron E Kornblith, and Atul J Butte. 2024. Evaluating the use of large language models to provide clinical recommendations in the emergency department. *Nature communications*, 15(1):8236.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. Medical graph rag: evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467.

Yilin Xiao, Junnan Dong, Chuang Zhou, Su Dong, Qianwen Zhang, Di Yin, Xing Sun, and Xiao Huang. 2025. Graphrag-bench: Challenging domain-specific reasoning for evaluating graph retrieval-augmented generation. *arXiv preprint arXiv:2506.02404*.

Yunfei Xie, Juncheng Wu, Haoqin Tu, Siwei Yang, Bingchen Zhao, Yongshuo Zong, Qiao Jin, Cihang Xie, and Yuyin Zhou. 2024. A preliminary study of o1 in medicine: Are we closer to an ai doctor? *arXiv preprint arXiv:2409.15277*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D Goodman. 2024. Star: Self-taught reasoner bootstrapping reasoning with reasoning. In *Proc. the 36th International Conference on Neural Information Processing Systems*, volume 1126.

Zefan Zeng, Qing Cheng, Xingchen Hu, Yan Zhuang, Xinwang Liu, Kunlun He, and Zhong Liu. 2025. Kosel: Knowledge subgraph enhanced large language model for medical question answering. *Knowledge-Based Systems*, 309:112837.

Weihe Zhai, Arkaitz Zubiaga, Bingquan Liu, Cheng-Jie Sun, and Yalong Zhao. 2024. Towards faithful knowledge graph explanation through deep alignment in commonsense question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18920–18930.

A Prompt Templates Used in ReGraM

ReGraM performs prompt-based reasoning across multiple stages, including query decomposition, region selection, reasoning, answer generation, and verification. This appendix visualizes the main prompts used at each stage of execution. Each prompt is designed according to its functional role, and curly brackets {} denote dynamically filled input slots at runtime.

A.1 Reasoning Prompts

Prompts for query decomposition, KG interaction, and hop-level synthesis.

Prompt 1. Query Domain Classification Prompt

System: You are an expert biomedical query classifier. Your task is to choose the single most relevant category for the given query from the list below. The query may span multiple domains; select the category that best represents the main thrust of the question.
Category Options:

- `"GENE_PROTEIN"`: Questions about genes, proteins, interactions, expression.
- `"DRUG_THERAPY"`: Questions about drugs, therapies, treatments, side effects, indications.
- `"DISEASE_SYMPTOM"`: Questions about diseases, symptoms, phenotypes, diagnoses.
- `"PATHWAY_METABOLISM"`: Questions about pathways, metabolism, or processes.
- `"INTEGRATED"`: Complex questions spanning multiple categories.

Query: {user_question}
Output (JSON only):

```
«JSON_START»
{"category": "DRUG_THERAPY"}
«JSON_END»
```

Figure 2: Query Domain Classification Prompt. Classifies a question into one of five high-level biomedical domains (GENE_PROTEIN, DRUG_THERAPY, etc.). The predicted category is later used for relation weighting and KG region selection

Prompt 2. Multi-hop CoT Decomposition Prompt

System: You are a world-class medical reasoning engine. Your task is to decompose a complex medical question into up to three logical, answerable sub-questions (hops). First, provide your reasoning process (**Analysis**), then output the sub-questions as JSON.
Example:

```
User Query: "What drug that targets the FLT4 gene also suppresses inflammation?"

Analysis:
1. The primary entity is "FLT4".
2. The secondary condition is "inflammation".
3. The task is to find a drug connecting them.

Output (JSON):
{"hops": ["Hop 1: ...", "Hop 2: ...", "Hop 3: ..."]}
```

Task: User Query: «<{user_question}>» Your Response: Provide both "Analysis" and "Final JSON".

Figure 3: Multi-hop CoT Decomposition Prompt. Decomposes a query into up to three logical sub-questions (hops) using a one-shot Chain-of-Thought prompt. Serves as the core step for constructing ReGraM's reasoning path

Prompt 3. Hypothetical Triplet Generation Prompt (HYBRID Mode)

System: You are a cautious biomedical reasoning expert. Your goal is to generate logical triplets to answer the given sub-question.
Sub-question: {hop_question}
Known Facts (from KG, may be empty): {known_facts}
Rules:

1. Use only the information in "Known Facts".
2. If insufficient, generate general, high-level plausible triplets.
3. Do not invent new entities or overly specific relations.
4. All generated triplets must use only entities present in V_q and relations from the allowed schema set R.

Output (JSON only):

```
«JSON_START»
{"Triplets": [["entity_A", "has_function", "function_X"], ...]}
«JSON_END»
```

Figure 4: Hypothetical Triplet Generation Prompt (HYBRID Mode). Used when insufficient KG evidence is retrieved. The LLM conservatively generates triplets using only allowed entities and relations while avoiding new entity creation

Prompt 4. Triplet Revision Prompt (Reviewer Loop)

System: You are a fact-checker verifying the factuality and usefulness of the given triplet.
Constraints:

- Relation MUST be one of: {allowed_relations}.
- Use only entities present in the input.

Triplet: {t}
Question (context): {q}
Output (JSON only):

```
«JSON_START»
{"Revised_Triplets": [["Head", "relation", "Tail"]]}
«JSON_END»
```

Figure 5: Triplet Revision Prompt (Reviewer Loop). Used when a generated triplet is judged as false by the Reviewer. Encourages correcting schema-violating relations and removing unnecessary hallucinated content

Prompt 5. Hop Answer Synthesis Prompt — KG_STRICT Mode

System: You are an expert medical reasoner. Based **only** on the “Verified Facts,” provide a direct, concise answer to the sub-question.
Sub-question: {hop_question}
Verified Facts: {verified_triplets}
Concise Answer:

Figure 6: Hop Answer Synthesis Prompt — KG_STRICT Mode. Used when $\geq$ 10 KG facts are available; the LLM generates answers solely from verified evidence without using its internal knowledge

Prompt 6. Hop Answer Synthesis Prompt — HYBRID Mode

System: You are a knowledgeable analyst. Derive a final answer by reasoning step-by-step using the verified facts.
Sub-question: {hop_question}
Verified Facts: {verified_triplets}
Reasoning Steps:

```
1. Fact: ...
2. Fact: ...
3. Connection: ...
```

Final Answer (1 sentence):

Figure 7: Hop Answer Synthesis Prompt — HYBRID Mode. Activated when 1–9 KG facts exist; the LLM performs stepwise reasoning using partial evidence to derive a final answer

Prompt 7. Hop Answer Synthesis Prompt — LLM_GUESS Mode

System: You are an expert biomedical researcher. No specific facts were found in the KG. Based on general knowledge, provide a plausible answer beginning with “Based on general knowledge,”.
Sub-question: {hop_question}
Plausible Answer:

Figure 8: Hop Answer Synthesis Prompt — LLM_GUESS Mode. Applied when no relevant KG facts are retrieved; the LLM relies on parametric knowledge and explicitly expresses uncertainty in its answer

Prompt 8. Final Answer Synthesis Prompt

System: You are a lead researcher writing a final, conclusive answer for a user. Synthesize the reasoning contained in the Evidence Map to answer the Original Question.
Original Question: {original_query}
Evidence Map: {evidence_map}
Final Conclusive Answer:

Figure 9: Final Answer Synthesis Prompt. Integrates hop-level reasoning results from the Evidence Map into a coherent final answer, ensuring logical consistency across hops

A.2 Evaluation Prompts

Prompts for automatic evaluation, hallucination detection, and inconsistency scoring.

Prompt 9. SAQ Judge Prompt (LLM-as-a-Judge)

System: You are an expert evaluator grading a submitted answer. Compare the “Ground Truth Answer” with the “Model’s Answer.”
Task:

1. Assess correctness.
2. Provide a similarity score (0.0–1.0).
3. Output JSON including:
   - `"reasoning"`
   - `"similarity_score"`
   - `"is_correct"` (true if score $\geq$ 0.8)

Question: {query}
Ground Truth Answer: {ground_truth_answer}
Model’s Answer: {model_response}

Figure 10: SAQ Judge Prompt (LLM-as-a-Judge). Evaluates semantic alignment between a model-generated free-form answer and the ground truth using GPT-4o; responses with similarity $\geq$ 0.8 are considered correct

Prompt 10. Hallucination Judge Prompt (MedHallu)

System: You are an expert medical fact-checker. Determine whether the model’s answer contains hallucination (unsupported claim, contradiction, fabrication) given the knowledge context.
Question: {query}
Knowledge Context: {reference_context}
Model’s Answer: {model_answer}
Output (JSON only):

```
- supported / unsupported /
contradicted
```

Figure 11: Hallucination Judge Prompt (MedHallu). Used in MedHallu evaluation to determine whether a model response contains hallucination (e.g., unsupported claims or fabricated facts) given reference context

Prompt 11. Inconsistency Scorer Prompt (IS)

System: You are an evaluator that measures factual inconsistency (hallucination) between two answers to the same medical question.
Task: Given a **Question**, a **Reference Answer** (ground truth), and a **Model Answer**, output a JSON object with:

- `"is"`: a number in [0,1], where 1 indicates maximally inconsistent / hallucinated and 0 indicates fully consistent.
- `"rationale"`: a brief explanation (1–2 sentences).

Constraints:

- Output **JSON only**.
- Use deterministic evaluation (temperature=0) in our setup.

Question: {query}
Reference Answer: {ground_truth}
Model Answer: {model_answer}

Figure 12: Inconsistency Scorer Prompt (IS). Computes a factual inconsistency score in [0,1] between the model answer and the reference answer under SAQ. We use deterministic decoding (temperature=0) and report dataset-level averages (Table 1)

B Implementation Details

To enhance reproducibility, this appendix details the implementation settings of ReGraM, including model configurations, inference hyperparameters, KG preprocessing pipeline, relation weighting schema, and full KG statistics.

B.1 Model and Inference Settings

Backbone Language Model The core reasoning modules of ReGraM are based on `Llama-3.1-8B-Instruct` (query decomposition, triplet generation, hop-by-hop reasoning, synthesis). We applied 4-bit quantization (NF4, via `bitsandbytes`) and used `torch.bfloat16` precision.

- **Model:** `meta-llama/Meta-Llama-3.1-8B-Instruct`
- **Quantization:** 4-bit NF4
- **Precision:** `torch.bfloat16`

Reviewer Model (Binary Verifier) LoRA-tuned `Llama-3.1-8B-Instruct`; entity/relation embeddings initialized via RotatE on PrimeKG.

- **Base:** `meta-llama/Meta-Llama-3.1-8B-Instruct`
- **Embedding Init:** RotatE (PrimeKG)
- **LoRA:** rank $r = 64$, $\alpha = 16$, modules = `q_proj`, `v_proj`, dropout=0.1

Inference Settings

- Temperature = 0.3, Top-p = 0.9
- Max tokens: Classification = 256, Reasoning = 512, Synthesis = 1024

Evaluation Phase:

- Model: GPT-4o (OpenAI), Temperature = 0.0

Expanded Relation Description Map ReGraM extends the relation description map used in KGARevion by providing richer natural-language descriptions, domain annotations, and domain-aware weights for PrimeKG relation types. This expansion is designed to improve region-aware retrieval accuracy rather than to introduce new knowledge. All mappings operate strictly over existing PrimeKG relations.

B.2 Knowledge Graph Preprocessing and Retrieval

This subsection details the implementation of the **Region Selection** module (Step (2) in Figure 1), which constructs a query-aligned KG region for each sub-question. As shown in Figure 1, region selection consists of (i) entity linking and expansion, (ii) domain-aware relation weighting, and (iii) weighted MMR re-ranking to select Top-K triplets that form the localized region G_q.
6

Entity Linking and Expansion (Figure 1-(2), "Entity Linking") To convert natural language queries into structured, KG-aligned representations, we adopt a hybrid entity linking pipeline that combines symbolic linking with fuzzy matching and synonym expansion. This step corresponds to the "Entity Linking" component in the Region Selection block of Figure 1.

- **Primary linker:** `scispaCy` using the UMLS Knowledge Base linker (`en_core_sci_sm`).
- **Fuzzy Matching:** `rapidfuzz` (Levenshtein ratio ≥ 90) to recover surface-form variations.
- **Synonym Expansion:** A domain-specific alias dictionary (`SYNONYM_MAP`) curated from PrimeKG node attributes (e.g., "Tylenol" → "Acetaminophen").

Domain-Specific Relation Weighting (Figure 1-(2), "Domain-Aware Relation Weighting") ReGraM assigns domain-specific weights w_r to relation types during region construction. This corresponds to the "Domain-Aware Relation Weighting" component in Figure 1-(2), and it biases region selection toward relations that are semantically relevant to the predicted query domain. Table 7 summarizes the full weighting scheme.

Weighted MMR Re-ranking and Region Formation (Figure 1-(2), "Weighted MMR") Given candidate triplets connected to the expanded entity set, we re-rank them to balance semantic relevance and diversity using weighted Maximal Marginal Relevance (MMR). This step corresponds to the "Weighted MMR" component in Figure 1-(2) and selects Top-K triplets to form the localized region G_q.

- **Embedding:** `sentence-transformers/all-MiniLM-L6-v2` (384-d)
- **MMR diversity factor:** $\lambda = 0.7$

Relation	**Gene_Protein**	**Drug_Therapy**	**Disease_Symptom**	**Pathway_Metabolism**	**Integrated**
Interacts_with	1.5	0.8	0.6	1.0	1.2
Targets	0.8	1.5	0.8	1.0	1.3
Treats	0.6	1.5	1.2	0.8	1.1
Causes	0.5	0.7	1.5	0.8	1.0
Expressed_in	1.3	0.7	0.5	1.0	1.1
Associated_with	1.0	1.0	1.3	0.9	1.2
Regulates	1.4	0.8	0.7	1.5	1.3
Occurs_in	0.9	0.8	1.0	1.2	1.1

Table 7: Relation weighting matrix used in KG region construction Weights w_r are applied during region selection based on the predicted domain (rows: relations, columns: query domains). Higher weights prioritize semantically aligned relations during retrieval

Dataset	N
AfrimedQA (MCQ)	3,910
AfrimedQA (SAQ)	1,236
MedQA (USMLE)	1,273
MedDDx	1,769
MedHallu	1,000
PubMedQA	1,000
MedHalt (FCT)	300
MedHalt (NOTA)	300
MedHalt (Fake)	300

Table 8: Number of evaluation instances used in our experiments.

Dataset	**Type**	**ReGraM**	**w/o MMR**	**Δ**
MedDDx-Expert	MCQ	44.44	43.06	$+1.38$
MedQA	MCQ	54.38	54.80	-0.42
MMLU-Med	MCQ	65.72	65.66	$+0.06$
MedDDx-Expert	SAQ	20.81	27.60	-6.79
MedQA	SAQ	51.03	25.30	$+25.73$
MMLU-Med	SAQ	65.45	42.30	$+23.15$

Table 9: Effect of removing MMR on MCQ and SAQ accuracy MMR reduces redundancy in KG retrieval, especially in SAQ, where excessive overlap between triplets amplifies reasoning drift

- **Region size:** $K = 15$

Overall, the above pipeline operationalizes Step (2) in Figure 1 by transforming domain-aware, query-linked triplet candidates into a compact, query-aligned region used as the reasoning space in subsequent modules.

B.3 Dataset Statistics

Table 8 reports the number of evaluation instances used for each dataset in our experiments. For large-scale datasets, we follow prior work and report results on a randomly sampled subset.

	H=1	**H=3**	**H=5**	**H=7**	**H=10**
MCQ	33.82	**44.44**	42.33	41.67	39.33
SAQ	14.84	**20.81**	20.33	21.00	18.00

Table 10: Impact of reasoning hop-depth on MCQ and SAQ performance (MedDDx-Expert) The best results are obtained at Hop 3, while performance drops at larger hop depths due to over-expansion and noise accumulation.

C Additional Experiments

This appendix presents additional ablation results and qualitative reasoning examples that complement the main paper. These findings further illustrate ReGraM's reasoning stability, structural effects, and observed failure modes, providing insights for future analyses.

C.1 Effect of Removing MMR (Extended Results)

ReGraM employs weighted MMR-based triplet reranking to reduce redundancy within the KG region and to construct query-centered, refined subgraphs. Table 9 reports extended results showing the performance change when MMR is removed. In the MCQ setting, MMR removal has minimal impact, while in SAQ, redundant KG candidates cause increased reasoning drift and reduced consistency.

C.2 Number of reasoning steps(H) Analysis

Reasoning hop depth is a key factor for evaluating how multi-hop decomposition affects the accuracy and stability of KG reasoning. Table 10 presents results for MedDDx-Expert, measuring accuracy at hop depths of 1, 3, 5, 7, and 10. Performance peaks at Hop 3, while deeper paths cause noise propagation and increased hallucination frequency.

Categorization	Multihop	ACC	Δ
O	O	51.03	+0.00
X	O	2.91	−46.17
O	X	19.80	−29.28
X	X	12.17	−36.91

Table 11: Ablation of region selector and multi-hop decomposition on MedQA (SAQ). Disabling either module leads to severe degradation, and removing both causes the largest performance drop.

C.3 Interaction Between Region Selector and Multi-hop Decomposition

ReGraM's region-first architecture achieves maximum stability when semantic region selection and stepwise reasoning alignment operate complementarily. To evaluate their interaction, we conduct an ablation on MedQA (SAQ) by selectively disabling each component.

Table 11 presents the result. Removing either module causes a significant drop in accuracy, while removing both leads to a catastrophic performance collapse—confirming their complementary role in region-first reasoning.

C.4 Effect of Reviewer Module (Extended Results)

Group	Dataset / Setting	ReGraM	w/o Reviewer
SAQ Acc	MedDDx-Expert	20.81	27.67
	MedQA	51.03	24.33
	MMLU-Med	65.45	41.67
	PubMedQA	44.65	39.00
MedHallu	ACC / Sim	43.50 / 54.70	0.00 / 0.00
	Hallu / RError	0.8900 / 0.7200	1.0000 / 1.0000
Med-HALT	FQT (ACC / Hallu)	0.0233 / 0.7481	0.0233 / 0.9767
	FCT (ACC / Hallu)	0.2519 / 0.7481	0.3367 / 0.6833
	NOTA (ACC / Hallu)	0.1967 / 0.8033	0.2100 / 0.7867

Table 12: Reviewer ablation across QA and hallucination benchmarks SAQ Acc. reports exact-match accuracy. MedHallu reports ACC, semantic similarity (Sim; 0–100), hallucination rate (Hallu), and reasoning error rate (RError). Med-HALT reports ACC and hallucination rate (lower is better)

To assess the contribution of the Reviewer module in hallucination suppression and factual consistency, we conducted ablation experiments where the module was removed from ReGraM. Table 12 summarizes performance across QA and hallucination benchmarks.

While MCQ results show mixed changes—with slight accuracy increase in MedQA (+2.29) and minor drops elsewhere—the impact in SAQ settings is substantial. Removing the Reviewer causes severe degradation in MedQA (−26.70), MMLU-Med (−23.78), and PubMedQA (−5.65). Conversely, MedDDx-Expert (SAQ) shows a surprising accuracy increase, suggesting that conservative verification can introduce a trade-off between factual precision and answer completeness when KG evidence is sufficiently dense.

To further evaluate the Reviewer's effectiveness, we ablated it from ReGraM and evaluated the model on MedHallu and Med-HALT hallucination benchmarks. As shown in Table 12, the model without the Reviewer fails catastrophically on MedHallu (ACC=0, hallucination=100%), underscoring the role of verification as a supporting mechanism for enhancing hallucination robustness—particularly when KG evidence is sparse or ambiguous. While not the primary driver of factual accuracy, the verification layer contributes to stabilizing generation under uncertain contexts within the region-first structure.

C.5 Efficiency Analysis (Extended Results)

Setting	KGARevion (s/item)	ReGraM (s/item)
MCQ	1.85	**1.00**
SAQ	3.26	**3.90**

Table 13: Average inference time per item under MCQ and SAQ settings

C.6 Interaction Between Domain Prior and Multi-hop Reasoning

Domain Prior	Multihop	ACC	Δ
O	O	51.03	0.00
X	O	2.91	-48.12
O	X	19.80	-31.23
X	X	12.17	-38.86

Table 14: Interaction between domain prior and multi-hop reasoning on MedQA (SAQ). Removing either component significantly reduces performance, and removing both causes the most severe degradation

C.7 Domain-wise Hallucination Trends

To further investigate the relationship between KG coverage and hallucination robustness, we report in Table 15 the domain-wise inconsistency scores (IS) for each dataset under the SAQ setting. While domains such as `Drug_Therapy` and `Gene_Protein` (with relatively high KG coverage) generally exhibit lower hallucination rates, this pattern does not hold uniformly—e.g., the `Integrated` domain has the highest KG coverage but a relatively high mean IS. This suggests that although KG coverage may contribute to consistency, it is not the sole determining factor; factors such as evidence redundancy, relation granularity, or prompt design may also play important roles.

Dataset	Drug_Therapy	Gene_Protein	Disease_Symptom	Pathway_Metab	Integrated
MedDDx-B	0.0887	0.0455	0.0769	0.1429	0.0784
MedDDx-I	0.0370	0.0679	0.0520	0.0400	0.0000
MedDDx-E	0.0000	0.1198	0.0803	0.0526	0.0347
MedQA	0.1014	0.0833	0.0628	0.0000	0.2500
AfrimedQA	0.0806	–	0.0899	–	0.0294
MMLU	0.0167	0.0417	0.0883	0.0750	0.0800
PubMedQA	0.0310	0.0000	0.0509	0.0147	0.0054
KG Coverage (%)	68.14	60.30	15.66	26.31	83.52

Table 15: Domain-wise hallucination IS vs KG coverage (SAQ). Higher KG coverage correlates with lower hallucination rates, while sparsely represented domains exhibit greater instability.

C.8 Dataset-wise Hallucination Trends

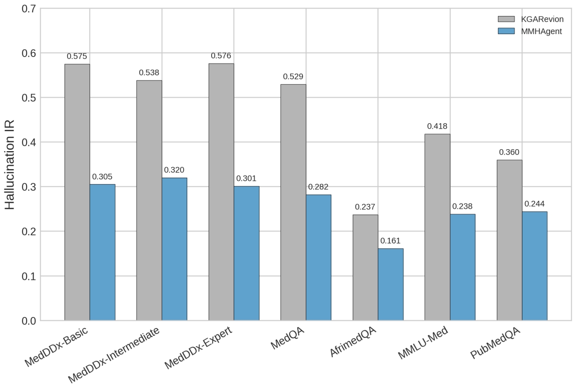

Figure 13: Reduction in hallucination inconsistency score (IS) across datasets under SAQ setting. ReGraM shows consistent improvement across all benchmarks

C.9 Qualitative Examples (Strengths and a Key Bottleneck)

We present two step-by-step examples to illustrate (i) when ReGraM succeeds end-to-end in both MCQ and SAQ, and (ii) a representative bottleneck case where multi-hop evidence is correctly accumulated but the final decision step becomes inconsistent with the evidence.

Example 1 (index=50): End-to-end success in both MCQ and SAQ.

Query. What are the diseases associated with the NGLY1 gene, and what clinical outcomes does this link entail?
Options. (A) ALG1-CDG (B) lipoyl transferase 1 deficiency (C) NGLY1-deficiency (D) aminoacylase 1 deficiency. **Gold: C**

Step 1: Query typing and decomposition. *Domain:* `GENE_PROTEIN`. *Hop decomposition:* single hop (fallback run).

Step 2: Region selection (Top verified triplets). *Verified KG triplets (subset):* (1) `(NGLY1-deficiency, phenotype present, Developmental regression)`; (2) `(NGLY1-deficiency, phenotype present, Cerebral atrophy)`; (3) `(NGLY1-deficiency, phenotype present, Alacrima)`.

Step 3: Hop-level sub-answer. *Sub-answer$_1$:* NGLY1 is associated with **NGLY1-deficiency**, with neurological/systemic phenotypes such as developmental regression and cerebral atrophy.

Step 4: Final outputs. *MCQ:* Pred **C** (correct). *SAQ:* Mapped to option **C** and judged correct under the SAQ protocol.
Takeaway. ReGraM can solve the same clinical-genetics query end-to-end across both MCQ and SAQ by grounding outcomes in a compact, query-aligned region.

Example 2 (index=49): Evidence is correctly accumulated, but the final decision is inconsistent.

Query. What are potential diagnoses for chronic cystitis symptoms with reduced bladder capacity and frequent urination?
Options (MCQ). (A) cystitis cystica (B) chronic cystitis (C) cystitis (D) interstitial cystitis. **Gold: D**

Step 0: Hop decomposition (Adaptive_MultiHop).

- Hop 1: candidate diagnoses for chronic cystitis symptoms
- Hop 2: diagnoses linked to reduced capacity and frequent urination
- Hop 3: final diagnosis matching the combined profile

Step 1: Region evidence (verified triplets; subset).

- `(interstitial cystitis, phenotype present, Functional abnormality of the bladder)`
- `(interstitial cystitis, phenotype present, Pollakisuria)`
- `(interstitial cystitis, phenotype present, Nocturia)`
- `(interstitial cystitis, phenotype present, Urinary bladder inflammation)`

Step 2: Hop-level sub-answers (compressed).

- Sub-answer$_1$: Candidate diagnoses include **interstitial cystitis** and related cystitis subtypes.
- Sub-answer$_2$: Evidence links **interstitial cystitis** to frequent urination and bladder functional abnormality (reduced capacity).
- Sub-answer$_3$: The combined symptom profile is most consistent with **interstitial cystitis** (Option D).

Final output (MCQ). Predicted option: **B** (incorrect; Gold D).
Takeaway. Region-first multi-hop reasoning successfully concentrates evidence toward the correct diagnosis

(D) at the hop level, but the final discrete decision can still become inconsistent with the accumulated evidence, motivating stronger evidence-aware aggregation and calibrated selection.

D Region-Constrained Multi-Hop Reasoning Algorithm

Algorithm 1: Region-Constrained Multi-Hop Reasoning in ReGraM

```
Input. User query $Q$, knowledge graph $\mathcal{G}$, language model $\mathcal{M}$
Output. Final answer $A$, evidence map $\mathcal{E}$

Phase 1: Query Processing
  • $D \leftarrow \textsc{DomainClassification}(Q)$
  • $\{q_1, \ldots, q_k\} \leftarrow \textsc{QueryDecomposition}(Q)$ $(k \leq 3)$
  • $C \leftarrow \emptyset$        (context: hop-level Q/A memory)
  • $\mathcal{E} \leftarrow \emptyset$        (evidence map)
Phase 2: Region Construction & Hop-wise Reasoning (for each sub-question $q_i$)
  • $E_{\text{sci}} \leftarrow \textsc{EntityLink}_{\textsc{UMLS}}(q_i)$
  • $E_{\text{fuzzy}} \leftarrow \textsc{FuzzyMatch}(q_i, \mathcal{G})$
  • $E_{\text{exp}} \leftarrow \textsc{SynonymExpand}(E_{\text{sci}} \cup E_{\text{fuzzy}})$
  • $T_{\text{cand}} \leftarrow \textsc{ExtractTriplets}(E_{\text{exp}}, \mathcal{G})$
  • $E_q \leftarrow \textsc{Weighted}\text{MMR}(T_{\text{cand}}, D, K = 15)$    (Top-$K$ triplets)
  • $V_q \leftarrow \{ v \mid v$ appears in a head/tail of some $t \in E_q \}$
  • $G_q \leftarrow (V_q, E_q)$        (query-aligned reasoning region)
  • $F_i \leftarrow \textsc{Textify}(E_q)$        (facts shown to $\mathcal{M}$)
  • $N_{\text{facts}} \leftarrow |E_q|$    and determine mode:
      – If $N_{\text{facts}} \geq 10$, then KG-STRICT
      – If $0 < N_{\text{facts}} < 10$, then HYBRID
      – Else, LLM-GUESS
  • $(a_i, T_i) \leftarrow \textsc{Reason}(q_i, F_i, C, \text{mode};\ G_q)$
  • Hard constraint: all KG lookups / entity expansion inside REASON satisfy
        $\textsc{Lookup}(\cdot) \subseteq G_q$ and generated entities must be in $V_q$.
  • If mode = HYBRID: $(a_i, T_i) \leftarrow \textsc{ReviewerVerify}(q_i, a_i, T_i;\ V_q, \mathcal{R})$    ($\mathcal{R}$: allowed relations)
  • $C \leftarrow C \cup \{(q_i, a_i)\}$,    $\mathcal{E} \leftarrow \mathcal{E} \cup T_i$
Phase 3: Final Synthesis
  • $A \leftarrow \textsc{Synthesize}(Q, C, \mathcal{E})$
  • return $A, \mathcal{E}$
```