Beyond Lexical Similarity: Evaluating Faithfulness in LLM-Based Medical Question Reformulation

Md Rabiul Hasan, Aleka Melese Ayalew, Mourad Oussalah
Center for Machine Vision and Signal Processing, University of Oulu, 90014, Oulu, Finland
{MdRabiul.Hasan, Aleka.Ayalew, Mourad.Oussalah}@oulu.fi

Abstract

Medical query rewriting transforms verbose consumer health questions into concise clinical queries, a critical step in health information retrieval. Large language models (LLMs) perform well on this task by standard metrics, yet high ROUGE or BERTScore does not guarantee preservation of clinical content. To address this issue, we introduce *MedFaith-F1*, a category-level faithfulness metric over four clinically salient categories: clinical problems, medications, procedures, and follow-up intent. We further propose a hybrid Evidence and Knowledge-Grounded Retrieval-Augmented Generation (*EKG-RAG*), an evidence and knowledge-grounded framework combining hybrid retrieval over PubMed and MedlinePlus resources with UMLS (Unified Medical Language System)-aligned ontology grounding. Evaluating LLMs LLaMA-3 and Qwen2.5 across zero-shot, few-shot, and QLoRA settings on MeQSum and medical question-pair (MQP) datasets revealed that base models exhibit category-level faithfulness failure rates (CHR) exceeding 40%, invisible to standard metrics, while EKG-RAG with QLoRA reduces CHR to 26.75%, achieving MedFaith-F1 of 0.73. Our findings call for faithfulness-aware evaluation in clinical query rewriting, and MedFaith-F1 provides a reproducible step in that direction. https://github.com/digihealth1230-cmd/RAG-KG

1 Introduction

Consumer health platforms and clinical information retrieval systems increasingly rely on automated medical query rewriting to transform verbose, informal consumer health questions into concise professional-style queries that improve downstream retrieval effectiveness (Lee et al., 2024). As large language models (LLMs) have become the dominant tool for this task, they are routinely evaluated using lexical and semantic similarity metrics (e.g., ROUGE, BLEU, METEOR, and BERTScore) inherited from general-purpose text generation benchmarks (Zhou et al., 2025).

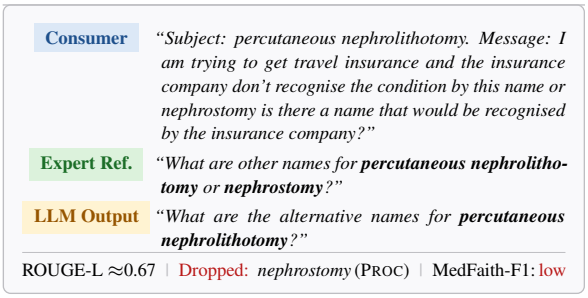

Figure 1: Example of faithfulness gap in medical query rewriting (MeQSum).

However, *similarity is not faithfulness.* The example in Figure 1 from MeQSum (Abacha and Demner-Fushman, 2019) illustrates this concretely. The LLM output scores high on ROUGE-L, yet silently drops *nephrostomy*, a clinically related procedure equally central to the patient's query, while standard metrics register this as strong performance. This failure mode of high surface similarity masking category-level entity omission reflects a structural mismatch between what similarity metrics measure and what clinical faithfulness requires. Prior work on faithfulness evaluation has largely addressed long clinical documents, such as discharge summaries and radiology reports (Zhang et al., 2023), where omissions affect multiple sentences and remain detectable through redundancy. In a short-form query rewriting, a single dropped entity, such as a medication name, a procedure alternative, or a follow-up instruction, can redirect the retrieval toward clinically irrelevant content while leaving similarity scores intact (Abbasian and Khatibi, 2024; Yim et al., 2025). Despite the growing awareness of hallucination in LLM-generated clinical text (Rahman et al., 2026), category-level omission in short-form medical query rewriting has not been systematically studied across learning

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 93–102

paradigms and grounding configurations.

This paper addresses that gap through the following research questions:

RQ1: To what extent do standard similarity metrics capture category-level clinical faithfulness in short-form medical query rewriting?

*We address this by introducing **MedFaith-F1**, a category-level faithfulness metric that evaluates retention of clinical problems (Dx), medications (Rx), procedures (Proc), and follow-up intent (Fup) via UMLS-normalized binary labels, making category-level omissions directly measurable alongside surface similarity scores.*

RQ2: Can structured knowledge grounding systematically reduce the divergence between surface fluency and clinical information retention?

*We address this by proposing **EKG-RAG**, an evidence and knowledge-grounded reformulation framework that couples hybrid sparse-dense retrieval over biomedical corpora PubMed and MedlinePlus with ontology-guided knowledge graph construction via SapBERT, SNOMED, RxNorm, and UMLS-based semantic linkage whose clinical categories directly mirror those of MedFaith-F1, providing structured grounding that unstructured retrieval alone cannot supply.*

RQ3: Which clinical category is most prone to omission under short-form compression, and can grounding mitigate it?

We address this by conducting per-category ablation across 36 configurations, finding that clinical problem retention is the most vulnerable category and that EKG-RAG's ontology coverage determines which categories grounding can and cannot rescue.

Our results show that base models in zero-shot settings exhibit category-level hallucination rates exceeding 40%, invisible to ROUGE and BERTScore. EKG-RAG with QLoRA reduces this to 26.75% on MeQSum, achieving MedFaith-F1 of 0.73, while ROUGE-L improvement remains modest.

In summary, our contributions are the following:

- We present a systematic evaluation of LLM-based medical query rewriting across zero-shot, few-shot, and QLoRA settings on MeQSum and MQP datasets using two model families (LLaMA-3 and Qwen2.5), comparing base generation, RAG, and EKG-RAG configurations.
- We propose **MedFaith-F1**, a reproducible category-level faithfulness metric measuring retention of clinical problems, medications, procedures, and follow-up intent, revealing that high lexical similarity scores do not reflect equivalent gains in clinical information retention across all evaluated settings.
- We develop **EKG-RAG**, a reformulation framework that grounds generation in an ontology-aligned clinical knowledge graph co-designed with MedFaith-F1, outperforming text-only RAG on MedFaith-F1 across all 36 configurations on MeQSum and MQP, with the largest gains observed under parameter-efficient fine-tuning (PEFT).
- We conduct a per-category faithfulness analysis demonstrating that clinical problem retention is the most vulnerable clinical information type, and that category-level hallucination rates exceeding 40% in zero-shot settings remain invisible to standard lexical and semantic metrics.

2 Related Work

2.1 Medical query rewriting and summarization

Consumer health question summarization was formalized by Abacha and Demner-Fushman (2019), who introduced MeQSum and showed that neural abstractive models can compress verbose patient queries into professional-style summaries. Subsequent work extended this through entailment-guided summarization (Mrini et al., 2021), chain-of-thought reformulation (Lee et al., 2024), and cross-lingual frameworks using LLaMA-2 (Abrar et al., 2025). Van Veen et al. (2024) demonstrated that adapted LLMs can match or exceed clinical expert performance on medical text summarization Ayalew et al. (2026). However, all of these approaches rely exclusively on lexical and semantic similarity metrics and none examined whether clinically salient categories are faithfully preserved under short-form compression.

2.2 Faithfulness evaluation in natural language generation

The limitations of surface-level metrics have been studied extensively in general summarization. Fabbri et al. (2021) showed through SummEval that automatic metrics correlate poorly with human faithfulness judgments, particularly for abstractive models. Zhang et al. (2023) extended this to

medical summarization, demonstrating systematic failure to detect errors through standard metrics. More recently, Wan et al. (2025) showed that LLM-generated summaries exhibit a U-shaped faithfulness trend across long documents. However, these works address longer documents where omissions remain detectable through lexical redundancy; a single dropped entity in short-form query rewriting produces no redundant signal, making category-level failures systematically invisible to existing metrics.

2.3 Hallucination in clinical language models

Hallucination in LLM-generated clinical text has received growing attention (Abbasian and Khatibi, 2024; Pal et al., 2023). Zuo and Jiang (2024) introduced MedHallBench to assess factual errors across diagnostic and treatment categories, while Rahman et al. (2026) surveyed fact-checking approaches for clinical LLMs. These efforts focused on unsupported claims against a verifiable knowledge base. The specific problem of entity omission, where clinically relevant information is dropped rather than fabricated remains largely unaddressed by existing hallucination benchmarks.

2.4 Knowledge-grounded retrieval-augmented generation

Hybrid sparse-dense retrieval (Formal et al., 2021) and ontology-aligned medical concept encoding (Liu et al., 2021) have substantially improved biomedical information retrieval. Xiong et al. (2024) benchmarked RAG systems for medical question answering, showing that corpus choice and retrieval strategy significantly affect factual accuracy. Cao et al. (2026) extended RAG to long-horizon EHRs using structured document representations. However, existing medical RAG frameworks were evaluated primarily for answer accuracy rather than preservation of specific clinical categories, a gap our EKG-RAG and MedFaith-F1 directly address.

3 Methodology

3.1 Problem Formulation

Let q denote a consumer-authored medical question and $\hat{q}$ its reformulated professional-style version. The goal of medical query rewriting is to generate a concise and fluent $\hat{q}$ that preserves the clinical information need expressed in q. We decompose this information need into four clinically salient categories:

$$\mathcal{C} = \{Dx,\ Rx,\ Proc,\ Fup\} \tag{1}$$

corresponding to clinical problems (Dx), medications (Rx), procedures ($Proc$), and follow-up intent (Fup). Dx subsumes the full UMLS *Disorders* semantic group, encompassing confirmed diagnoses, reported symptoms, and suspected conditions as expressed in consumer health questions. These four categories correspond to the primary UMLS semantic groups normalized by our knowledge graph (SNOMED CT for Dx and $Proc$; RxNorm for Rx) and the dominant pragmatic intent class in consumer health questions (Fup), providing principled co-design between the metric and the grounding framework. Given a parameterized LLM-based reformulation model f_θ, the task is to estimate $\hat{q}$ such that:

$$\hat{q} = f_\theta(q \mid \mathcal{E},\ \mathcal{K}) \tag{2}$$

where $\mathcal{E}$ denotes retrieved textual evidence and $\mathcal{K}$ denotes structured medical knowledge. Standard LLM-based reformulation conditions only on q; we explicitly study the effect of grounding through $\mathcal{E}$ and $\mathcal{K}$.

3.2 EKG-RAG: Evidence and Knowledge-Grounded Reformulation

Figure 2 presents the EKG-RAG pipeline, which extends standard RAG through two parallel grounding pathways: hybrid evidence retrieval and ontology-guided knowledge construction.

Evidence retrieval ($\mathcal{E}$): Given the input q, we retrieve a candidate set $\mathcal{R}$ of top-N passages using SPLADE-v2 (Formal et al., 2021) over a combined corpus of PubMed abstracts and MedlinePlus articles. Candidates are then encoded with a dense retriever and re-ranked via cosine similarity measure. To balance relevance and coverage, we apply *budgeted Maximal Marginal Relevance* (MMR) (Carbonell and Goldstein, 1998):

$$\text{MMR}(p_i, q) = \lambda \cdot \text{sim}(p_i, q) - (1-\lambda) \cdot \max_{p_j \in \mathcal{S}} \text{sim}(p_i, p_j) \tag{3}$$

where $\mathcal{S}$ is the set of already-selected passages, $p_i \in \mathcal{R} \setminus \mathcal{S}$ restricts this selection to unselected candidates, and λ balances relevance against diversity. Section-based scoring up-weights passages from clinically informative regions (e.g., indications, contraindications), producing the final diversified evidence block $\mathcal{E}$.

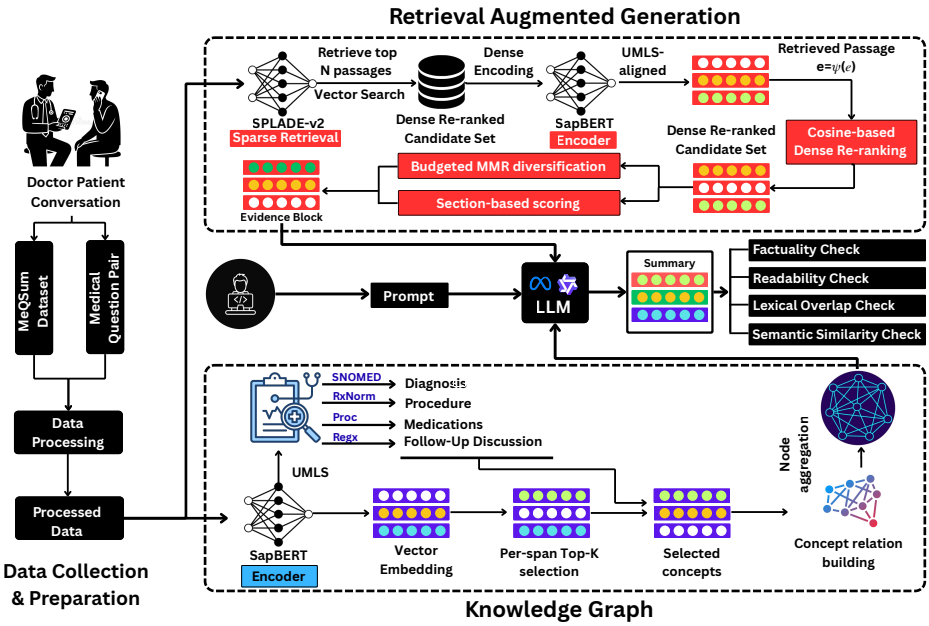

Figure 2: EKG-RAG pipeline: the input question is processed through hybrid retrieval and medical knowledge extraction, producing evidence and knowledge contexts that ground the LLM reformulation.

Knowledge construction ($\mathcal{K}$): In parallel, clinically salient spans are identified in q using a hybrid extraction approach: rule-based pattern matching identifies candidate spans, which are encoded with SapBERT (Liu et al., 2021), a biomedical encoder pre-trained via self-alignment on UMLS concept embeddings. For each extracted span, the K most similar UMLS concepts are retrieved by cosine similarity over SapBERT-encoded concept embeddings. Concepts are normalized using SNOMED (Abdulnazar et al., 2023) for diagnoses and procedures and RxNorm (Cai et al., 2023) for medications. Concept relations are established through UMLS-based semantic linkage (Afshar et al., 2024), and node aggregation yields the structured knowledge representation $\mathcal{K}$. We manually inspected extraction outputs on a representative sample during development and found the pipeline reliably identifies primary clinical entities; however, we note that extraction quality on rare or highly ambiguous clinical terms may vary, which we acknowledge as a limitation in Section 5.

The reformulation model conditions jointly on $\mathcal{E}$ and $\mathcal{K}$ via the prompt in Figure 3:

Prompt Template

```
Role: Medical question reformulation specialist.
Objective: Convert a verbose consumer query into a concise professional medical question.
Clinical Requirements: Preserve clinical problems (Dx), medications (Rx), procedures (Proc), and follow-up intent (Fup); maintain negation, uncertainty, and temporal context; avoid unsupported claims.
Evidence Context: «BEGIN» {e_1}, ..., {e_k} «END».
Knowledge Context: {K}
Input Question: {q}
Reformulated Question: ...
```

Figure 3: Prompt template used for reformulation

3.3 MedFaith-F1: Category-Level Faithfulness Metric

Standard similarity metrics measure surface overlap between $\hat{q}$ and a reference but cannot detect category-level omissions. MedFaith-F1 addresses this by treating retention of each category $c \in \mathcal{C}$ as a binary classification problem at the instance level.

Span extraction and binary labelling: For each instance $(q, \hat{q})$, the SapBERT-based extraction pipeline is applied to both texts. A category c is labelled *present* if at least one UMLS-normalized concept belonging to c is identified; otherwise it is labelled *absent*. This produces binary labels $y_c^{(q)} \in \{0,1\}$ and $y_c^{(\hat{q})} \in \{0,1\}$ for source and reformulation respectively.

Category-level F1 and metric definition: For category c, a *true positive* (TP) occurs when c is present in both q and $\hat{q}$; a *false negative* (FN) occurs when it is present in q but absent in $\hat{q}$; a *false positive* occurs (FP) when it is present in $\hat{q}$ but absent in

q. Per-category F1 scores $\mathrm{F1}_c$ are computed across the evaluation set, and MedFaith-F1 is defined as the macro-average:

$$\text{MedFaith-F1} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{F1}_c \tag{4}$$

The *Category-level Faithfulness Failure Rate* (CHR) is derived automatically as:

$$\text{CHR} = (1 - \text{MedFaith-F1}) \times 100 \tag{5}$$

CHR quantifies the expected rate of category-level clinical faithfulness failures, encompassing both omissions (false negatives, where a clinically present category is dropped) and fabrications (false positives, where an absent category is introduced). We use the term *hallucination rate* as shorthand, while acknowledging that CHR captures the broader class of category-level unfaithfulness rather than fabrication alone. We further acknowledge that the binary formulation does not capture intra-category semantic shifts such as negation or severity modification.

4 Experimentation

4.1 Learning Paradigms and Experimental Configuration

We select LLaMA-3-8B and Qwen2.5-7B as backbone models, with configurations summarized in Table 1. Both are open-source, instruction-tuned at 7-8B scale, and support QLoRA fine-tuning; their pretraining corpora include biomedical and clinical text, making them well-suited for constrained medical query generation. Evaluating two architecturally distinct model families ensures that findings reflect general reformulation behaviour rather than model-specific artifacts. We use an 800/200 train/test split on MeQSum and an 80/20 split on the medical question-pair dataset.

4.2 Main Results

Table 2 reports the overall reformulation performance across all paradigms, retrieval configurations, models, and datasets. Four patterns emerge consistently across both datasets:

Grounding and fine-tuning improve lexical and semantic evaluation scores across all settings. RAG yields ROUGE-L gains of +0.01-0.04 over base generation across paradigms and models on both datasets. EKG-RAG achieves the highest lexical overlap and semantic similarity scores in every configuration, with best ROUGE-L of 0.54 on MeQSum and 0.59 on MQP, and best BERTScore of 0.87 and 0.89 respectively, both under PEFT with Qwen. Readability remains stable or improves slightly, with Flesch-Kincaid Grade Level (FKGL) decreasing from 10.5 to 9.3 on MeQSum and from 10.3 to 9.2 on MQP.

Table 1: Ablation of retrieval components and QLoRA training configuration. ✓ = enabled; ✗ = disabled.

Config	Sparse-Dense Retrieval	Evidence Block ($\mathcal{E}$)	Knowledge Graph ($\mathcal{K}$)
Base	✗	✗	✗
+RAG	✓	✓	✗
+EKG-RAG	✓	✓	✓
QLoRA Hyperparameters (PEFT only)			
Rank r / α / Dropout	8 / 16 / 0.05		
Target modules	Query & value projections		
Epochs / Batch size	3 / 16		
Learning rate	2×10^{-4}, cosine + 6% warmup		

Category-level hallucination is high and invisible to standard metrics. Zero-shot base models exhibit CHR of 45.00% (LLaMA) and 42.50% (Qwen) on MeQSum, and 42.75% and 40.50% on MQP, indicating high category-level faithfulness failure rates across reformulations, a failure rate not reflected in ROUGE-L scores of 0.43-0.46 and BERTScore of 0.72-0.81 under the same conditions. PEFT with EKG-RAG reduces CHR to 26.75% on MeQSum and 29.00% on MQP, an absolute reduction of 18.25 and 13.75 percentage points over the weakest zero-shot base, with no degradation in readability.

EKG-RAG consistently outperforms RAG-only. EKG-RAG yields average MedFaith-F1 gains of +0.037-0.065 over base generation across all paradigms on MeQSum, compared to +0.023-0.035 for RAG-only. The gap between EKG-RAG and RAG-only is positive across all 36 configurations (18 per dataset), confirming that structured knowledge in $\mathcal{K}$ provides complementary grounding beyond unstructured retrieval in $\mathcal{E}$ alone. The largest incremental gains over RAG-only occur under PEFT, indicating that fine-tuning amplifies the model's ability to exploit the knowledge graph.

Qwen outperforms LLaMA under EKG-RAG and PEFT. Under PEFT with EKG-RAG, Qwen achieves MedFaith-F1 of 0.7325 (CHR 26.75%) on MeQSum versus 0.6650 (CHR 33.50%) for LLaMA, with the gap narrowing in zero-shot setting.

Table 2: Category-wise evaluation on MeQSum and MQP datasets across all learning paradigms and retrieval configurations. **RL**: ROUGE-L; **BS**: BERTScore; **MF-F1**: MedFaith-F1; **CHR%**: failure rate (omissions and fabrications combined), lower is better; **FKGL**: Flesch-Kincaid Grade Level (Read: readability; lower is better).

		MeQSum Dataset										**Medical-Question-Pair Dataset**									
		Similarity			***Factuality Check***						***Read.***	***Similarity***			***Factuality Check***						***Read.***
Setting	**Model**	**RL**	**BLEU**	**BS**	**Dx**	**Rx**	**Proc**	**Fup**	**MF-F1**	**CHR%**	**FKGL**	**RL**	**BLEU**	**BS**	**Dx**	**Rx**	**Proc**	**Fup**	**MF-F1**	**CHR%**	**FKGL**
Zero-shot	**LLaMA**	0.43	0.16	0.72	0.51	0.63	0.54	0.52	0.5500	**45.00**	10.5	0.47	0.18	0.77	0.54	0.65	0.56	0.54	0.5725	**42.75**	10.3
	+RAG	0.46	0.18	0.77	0.54	0.67	0.56	0.55	0.5800	**42.00**	10.3	0.51	0.21	0.80	0.55	0.68	0.60	0.58	0.6025	**39.75**	10.1
	+EKG-RAG	**0.48**	**0.19**	**0.78**	**0.55**	**0.69**	**0.58**	**0.56**	**0.5950**	40.50	**10.1**	**0.52**	**0.21**	**0.82**	**0.59**	**0.72**	**0.62**	**0.60**	**0.6325**	36.75	**9.9**
	Qwen	0.46	0.17	0.81	0.53	0.66	0.56	0.55	0.5750	**42.50**	10.3	0.49	0.20	0.81	0.56	0.68	0.58	0.56	0.5950	**40.50**	10.1
	+RAG	0.49	0.19	0.83	0.57	0.70	0.59	0.58	0.6100	**39.00**	10.1	0.51	0.22	0.83	0.60	0.71	0.62	0.58	0.6275	**37.25**	9.9
	+EKG-RAG	**0.51**	**0.20**	**0.84**	**0.58**	**0.72**	**0.61**	**0.60**	**0.6275**	37.25	**9.9**	**0.54**	**0.23**	**0.86**	**0.61**	**0.74**	**0.64**	**0.62**	**0.6525**	34.75	**9.7**
Few-shot	**LLaMA**	0.45	0.18	0.76	0.54	0.67	0.57	0.56	0.5850	**41.50**	10.3	0.49	0.20	0.79	0.56	0.68	0.58	0.56	0.5950	**40.50**	10.1
	+RAG	0.49	0.20	0.78	0.57	0.71	0.60	0.59	0.6175	**38.25**	10.1	0.52	0.23	0.81	0.59	0.72	0.61	0.59	0.6275	**37.25**	9.8
	+EKG-RAG	**0.51**	**0.21**	**0.82**	**0.59**	**0.73**	**0.62**	**0.61**	**0.6375**	36.25	**9.9**	**0.55**	**0.23**	**0.84**	**0.61**	**0.74**	**0.64**	**0.62**	**0.6525**	34.75	**9.7**
	Qwen	0.49	0.19	0.80	0.56	0.69	0.59	0.58	0.6050	**39.50**	10.1	0.51	0.21	0.81	0.58	0.70	0.60	0.58	0.6150	**38.50**	9.9
	+RAG	0.51	0.21	0.83	0.60	0.73	0.62	0.61	0.6400	**36.00**	9.9	0.54	0.23	0.84	0.61	0.73	0.63	0.61	0.6450	**35.50**	9.7
	+EKG-RAG	**0.53**	**0.22**	**0.85**	**0.61**	**0.75**	**0.64**	**0.62**	**0.6550**	34.50	**9.7**	**0.56**	**0.24**	**0.86**	**0.63**	**0.76**	**0.66**	**0.64**	**0.6725**	32.75	9.5
PEFT	**LLAMA**	0.48	0.20	0.81	0.58	0.72	0.61	0.60	0.6275	**37.25**	9.9	0.51	0.21	0.82	0.59	0.71	0.61	0.59	0.6250	**37.50**	9.9
	+RAG	0.51	0.21	0.83	0.60	0.75	0.63	0.62	0.6500	**35.00**	9.7	0.54	0.23	0.84	0.62	0.74	0.64	0.60	0.6500	**35.00**	9.7
	+EKG-RAG	**0.52**	**0.22**	**0.84**	**0.61**	**0.77**	**0.65**	**0.63**	**0.6650**	33.50	**9.6**	**0.56**	**0.24**	**0.86**	**0.63**	**0.77**	**0.66**	**0.64**	**0.6750**	32.50	9.5
	Qwen	0.52	0.20	0.84	0.62	0.77	0.65	0.63	0.6675	**33.25**	9.6	0.53	0.22	0.84	0.61	0.73	0.63	0.61	0.6450	**35.50**	9.7
	+RAG	0.53	0.22	0.86	0.63	0.80	0.68	0.65	0.6900	**31.00**	9.4	0.56	0.24	0.86	0.63	0.76	0.66	0.64	0.6725	**32.75**	9.5
	+EKG-RAG	**0.54**	**0.24**	**0.87**	**0.66**	**0.83**	**0.75**	**0.69**	**0.7325**	26.75	**9.3**	**0.59**	**0.26**	**0.89**	**0.68**	**0.80**	**0.69**	**0.67**	**0.7100**	29.00	**9.2**

Across all configurations, structured knowledge grounding systematically reduces the divergence between surface fluency and clinical information retention, confirming that the answer to our second research question is affirmative: EKG-RAG consistently and measurably improves category-level faithfulness beyond what retrieval augmentation alone achieves.

Table 3: Cohen's d effect sizes on MedFaith-F1, computed over 12 aggregate configuration scores per dataset (3 paradigms $\times$ 2 models). Thresholds follow: $d \geq 0.8$ Large, $0.5 \leq d < 0.8$ Medium, $0.2 \leq d < 0.5$ Small.

Comparison	**MeQSum**		**MQP**	
	d	*Interp.*	*d*	*Interp.*
EKG-RAG vs. Base	1.15	Large	2.21	Large
RAG-only vs. Base	0.74	Medium	1.19	Large
EKG-RAG vs. RAG-only	0.49	Small	1.12	Large

4.3 Faithfulness Gap Analysis

Pearson and Spearman correlations confirm directional agreement between standard metrics and MedFaith-F1 across all 18 configurations on MeQSum (ROUGE-L: $r = 0.93$, $\rho = 0.96$; BERTScore: $r = 0.87$, $\rho = 0.89$; all $p < 0.001$), yet ROUGE-L improves by only 14.5% in relative terms while CHR falls by 40.6%, a 2.8-fold magnitude divergence invisible to lexical evaluation alone.

Table 3 quantifies practical significance via Cohen's d effect sizes over 12 aggregate configuration scores (3 paradigms $\times$ 2 models). EKG-RAG versus base generation yields large effects on both datasets ($d = 1.15$ on MeQSum; $d = 2.21$ on MQP). RAG-only versus base yields medium to large effects ($d = 0.74$; $d = 1.19$), while EKG-RAG versus RAG-only ranges from small to large ($d = 0.49$; $d = 1.12$). A Friedman test confirms that learning paradigm significantly affects MedFaith-F1 ($\chi^2 = 12.00$, $p = 0.003$), with mean scores increasing monotonically from zero-shot (0.590) to PEFT (0.672).

This divergence speaks directly to RQ1: high lexical similarity does not imply faithful preservation of clinically salient categories under short-form compression.

4.4 Per-Category Analysis

Clinical problem retention (Dx) proves to be the hardest category across datasets (average F1: 0.581 on MeQSum), as consumer questions express problems as informal symptoms or suspected conditions rather than confirmed diagnoses, resisting normalization to structured ontology concepts. Procedure retention (Proc) benefits most from EKG-RAG, showing the largest absolute gain from zero-shot base to PEFT with Qwen (+0.21 on MeQSum), followed by medication retention (+0.20), consis-

Table 4: Category-wise ablation study of retrieval and knowledge grounding on MeQSum and MQP datasets. Baseline rows report absolute performance, while subsequent rows show incremental changes relative to the corresponding baseline. HRisk provides a qualitative interpretation of faithfulness failure severity: $\geq$40% High, 30-40% Moderate, <30% Low. Higher MF-F1 (MedFaith-F1) indicates better faithfulness

			MeQSum								Medical-Question-Pair							
Setting	**Model**	**Variant**	**Dx**	**Rx**	**Proc**	**Fup**	**MF-F1**	**CHR%**	**HRisk**	**FKGL**	**Dx**	**Rx**	**Proc**	**Fup**	**MF-F1**	**CHR%**	**HRisk**	**FKGL**
Zero-shot	LLaMA	Base	**0.51**	**0.63**	**0.54**	**0.52**	**0.5500**	**45.00**	**High**	**10.5**	**0.54**	**0.65**	**0.56**	**0.54**	**0.5725**	**42.75**	**High**	**10.3**
		+RAG	+.03	+.04	+.02	+.03	+.0300	−3.00	**High**	−0.2	+.01	+.03	+.04	+.04	+.0300	−3.00	**Mod.**	−0.2
		+EKG-RAG	**+.04**	**+.06**	**+.04**	**+.04**	**+.0450**	**−4.50**	**High**	**−0.4**	**+.05**	**+.07**	**+.06**	**+.06**	**+.0600**	**−6.00**	**Mod.**	**−0.4**
	Qwen	Base	0.53	0.66	0.56	0.55	0.5750	42.50	High	10.3	0.56	0.68	0.58	0.56	0.5950	40.50	High	10.1
		+RAG	+.04	+.04	+.03	+.03	+.0350	−3.50	Mod.	−0.2	+.04	+.03	+.04	+.02	+.0325	−3.25	Mod.	−0.2
		+EKG-RAG	**+.05**	**+.06**	**+.05**	**+.05**	**+.0525**	**−5.25**	**Mod.**	**−0.4**	**+.05**	**+.06**	**+.06**	**+.06**	**+.0575**	**−5.75**	**Mod.**	**−0.4**
Few-shot	LLaMA	Base	**0.54**	**0.67**	**0.57**	**0.56**	**0.5850**	**41.50**	**High**	**10.3**	**0.56**	**0.68**	**0.58**	**0.56**	**0.5950**	**40.50**	High	**10.1**
		+RAG	+.03	+.04	+.03	+.03	+.0325	−3.25	**Mod.**	−0.2	+.03	+.04	+.03	+.03	+.0325	−3.25	**Mod.**	−0.3
		+EKG-RAG	**+.05**	**+.06**	**+.05**	**+.05**	**+.0525**	**−5.25**	**Mod.**	**−0.4**	**+.05**	**+.06**	**+.06**	**+.06**	**+.0575**	**−5.75**	**Mod.**	**−0.4**
	Qwen	Base	**0.56**	**0.69**	**0.59**	**0.58**	**0.6050**	**39.50**	**Mod.**	**10.1**	**0.58**	**0.70**	**0.60**	**0.58**	**0.6150**	**38.50**	**Mod.**	**9.9**
		+RAG	+.04	+.04	+.03	+.03	+.0350	−3.50	Mod.	−0.2	+.03	+.03	+.03	+.03	+.0300	−3.00	Mod.	−0.2
		+EKG-RAG	**+.05**	**+.06**	**+.05**	**+.04**	**+.0500**	**−5.00**	**Mod.**	**−0.4**	**+.05**	**+.06**	**+.06**	**+.06**	**+.0575**	**−5.75**	**Mod.**	**−0.4**
PEFT	LLaMA	Base	**0.58**	**0.72**	**0.61**	**0.60**	**0.6275**	**37.25**	**Mod.**	**9.9**	**0.59**	**0.71**	**0.61**	**0.59**	**0.6250**	**37.50**	**Mod.**	**9.9**
		+RAG	+.02	+.03	+.02	+.02	+.0225	−2.25	Mod.	−0.2	+.03	+.03	+.03	+.01	+.0250	−2.50	Mod.	−0.2
		+EKG-RAG	**+.03**	**+.05**	**+.04**	**+.03**	**+.0375**	**−3.75**	**Mod.**	**−0.3**	**+.04**	**+.06**	**+.05**	**+.05**	**+.0500**	**−5.00**	**Mod.**	**−0.4**
	Qwen	Base	**0.62**	**0.77**	**0.65**	**0.63**	**0.6675**	**33.25**	**Mod.**	**9.6**	**0.61**	**0.73**	**0.63**	**0.61**	**0.6450**	**35.50**	**Mod.**	**9.7**
		+RAG	+.01	+.03	+.03	+.02	+.0225	−2.25	Mod.	−0.2	+.02	+.03	+.03	+.03	+.0275	−2.75	Mod.	−0.2
		+EKG-RAG	**+.04**	**+.06**	**+.10**	**+.06**	**+.0650**	**−6.50**	**Low**	**−0.3**	**+.07**	**+.07**	**+.06**	**+.06**	**+.0650**	**−6.50**	**Low**	**−0.5**

Table 5: Qualitative examples of medical question reformulation. Color indicates faithfulness alignment: green (correct; present in input, clinician, and ours), purple (correct; present in input and clinician only), blue (correct; present in input and ours), orange (underspecified), and red (incorrect or hallucinated).

Patient Query (Input)	**Summary (Medical Professional)**	**Summary EKG-RAG (Ours)**
Input 1: What are some ways doctors would suggest to stop preterm labor?	How do doctors stop preterm labor?	What can be done to stop premature labor?
Input 2: Please help me with my brother with locked-in syndrome.	What are the treatments and support for locked-in syndrome?	Where can I find information on locked-in syndrome, and how can I help my brother?
Input 3: Do I need to prepare for a thyroid panel test? Do I need to fast?	What preparations are required for a thyroid panel test?	What are the preparation instructions for a thyroid panel blood test?

tent with the structured coverage of SNOMED and RxNorm in $\mathcal{K}$. Across all 36 configurations on both datasets, EKG-RAG improvements in MedFaith-F1 are positive with no configuration where EKG-RAG underperforms RAG-only on any individual category.

These findings answer our third research question: clinical problem retention is the clinical category most prone to omission under short-form compression, and while grounding partially mitigates this, the asymmetry across categories – strongest gains for Proc and Rx, smallest for Dx – reflects the fundamental challenge of normalizing colloquial diagnostic terms to structured ontology concepts.

4.5 Ablation Study

Table 4 isolates the contribution of each grounding component. Retrieval augmentation alone yields average MedFaith-F1 gains of +0.022-0.035 over base generation across all paradigms on MeQSum, with corresponding CHR reductions of 2-6 percentage points. Adding the knowledge graph (+EKG-RAG) provides a consistent additional gain of +0.015-0.043 over RAG-only, positive across all 36 configurations on both datasets. The largest incremental gains from EKG-RAG over RAG-only occur under PEFT with Qwen (+0.065 MedFaith-F1; CHR reduced from 31.00% to 26.75% on MeQSum), where fine-tuning amplifies the model's ca-

Table 6: Comparison with prior work on MeQSum, including MEDIQA 2021 shared-task entries and subsequent state-of-the-art systems.

Author	Method	RL	BLEU	BS	MF-F1	CHR%	FKGL	Key Limitation
Abacha et al. 2019	Ptr-Gen + data aug. (MeQSum baseline)	0.40	0.13	–	–	–	–	Pre-LLM Seq2Seq; no faithfulness eval
Yadav et al. 2021	ProphetNet + RL with QA-task rewards	0.31	–	0.68	–	–	–	RL rewards task-specific; no entity retention
Mrini et al. 2021	BART multi-task (summ + RQE)	0.44	–	–	–	–	–	No grounding; no category decomp.
Zhang et al. 2023	Contrastive CL + med. terms (FaMeSumm)	0.30	–	0.75	–	–	–	Synth. negatives; no RAG or ontology
Aloraini et al. 2025	Candidate re-ranker (LexiSem)	0.37	–	0.85	–	–	–	General-purpose; no medical grounding
Abrar et al. 2025	LLaMA-2 LoRA + NER-guided (cross-lingual)	0.47	0.18	0.74	–	–	7.0	SummaC/AlignScore; no cat.-level eval
Ours (LLaMA+EKG-RAG, PEFT)	LLM + EKG-RAG (knowledge-grounded)	0.52	0.22	0.84	0.6650	33.50	9.6	Instance-level predictions not retained
Ours (Qwen+EKG-RAG, PEFT)	**LLM + EKG-RAG (knowledge-grounded)**	**0.54**	**0.24**	**0.87**	**0.7325**	**26.75**	**9.3**	**Instance-level predictions not retained**

pacity to leverage structured clinical knowledge. Qualitative examples in Table 5 confirm these trends at the instance level, showing that EKG-RAG more reliably preserves key clinical entities in cases where similarity metrics register no failure.

4.6 Comparison with Prior Work

Table 6 compares our best configurations against prior systems on MeQSum, covering the original dataset baseline (Abacha and Demner-Fushman, 2019), two MEDIQA 2021 shared-task entries (Yadav et al., 2021; Mrini et al., 2021), and three subsequent systems (Zhang et al., 2023; Aloraini et al., 2025; Abrar et al., 2025). Our proposed Qwen with EKG-RAG (PEFT) achieves ROUGE-L of 0.54 and BERTScore of 0.87, exceeding prior methods on ROUGE-L and matching the strongest reported BERTScore (Aloraini et al., 2025). Critically, our proposed system in this comparison reports category-level faithfulness evaluation; prior systems cannot be retroactively assessed on MedFaith-F1 or CHR without access to their predictions. FKGL differences with Abrar et al. (2025) (7.0 vs. 9.3) reflect compression style variation rather than a quality deficiency.

5 Discussion and Limitations

High correlation between ROUGE-L and MedFaith-F1 (r=0.93) confirms that both metrics respond to the same quality improvements, yet they diverge substantially in magnitude: similarity gains understate faithfulness improvements by a factor of 2.8. This divergence is not noise; it reflects a structural limitation of lexical metrics in short-form clinical text, where dropping a single entity can alter clinical intent without reducing surface overlap.

The gain of EKG-RAG over RAG-only is consistent across all 36 configurations, indicating that retrieved passages alone are insufficient and that ontology-grounded knowledge provides constraints that retrieval cannot. The asymmetry across categories-strongest gains for Proc and Rx, smallest for Dx-reflects a fundamental challenge: diagnostic terms in consumer questions are often colloquial or implicit (e.g., "bad heart" instead of "coronary artery disease"), resisting normalization to UMLS concepts. Targeted grounding strategies, such as symptom-to-ICD linking, represent a natural extension of this work.

Limitations. MedFaith-F1 treats category retention as binary and does not detect partial preservation or intra-category shifts such as negation or severity modification. SapBERT-based extraction may underperform on rare or ambiguous clinical terms; as MedFaith-F1 shares infrastructure with EKG-RAG, faithfulness gains may carry optimistic bias. All experiments use English-language datasets; multilingual generalization requires further study with frontier models such as GPT-4o and Claude Opus. Extending MedFaith-F1 to additional clinical dimensions, such as symptoms, severity modifiers, and temporal context for multilingual settings, and deployed retrieval systems represent an important direction for future work.

6 Conclusion

Standard similarity metrics systematically underestimate faithfulness in LLM-based medical query rewriting, with ROUGE-L gains understating CHR reductions by a factor of 2.8 across all evaluated configurations. We introduced MedFaith-F1, a category-level metric that evaluates retention of clinical problems, medications, procedures, and follow-up intent, and EKG-RAG, a knowledge-grounded reformulation framework that reduces CHR to 26.75% under PEFT. Taken together, these contributions argue for clinically informed faithfulness evaluation as a necessary complement to surface metrics in medical IR systems.

Ethics Statement

This work uses publicly available datasets (MeQSum and the medical question-pair dataset) and does not involve the collection of new human subject data. All experiments use de-identified consumer health questions. No patient identifiers were accessed or processed.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234.

Mahyar Abbasian and Azimi Khatibi, Elahe. 2024. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai. *NPJ Digital Medicine*, 7(1):82.

Akhila Abdulnazar, Markus Kreuzthaler, Roland Roller, and Stefan Schulz. 2023. Sapbert-based medical concept normalization using snomed ct. In *Caring Is Sharing–Exploiting the Value in Data for Health and Innovation*, pages 825–826. IOS Press.

Ajwad Abrar, Nafisa Tabassum Oeshy, Prianka Maheru, Farzana Tabassum, and Tareque Mohmud Chowdhury. 2025. Faithful summarization of consumer health queries: A cross-lingual framework with llms. *arXiv preprint arXiv:2511.10768*.

Majid Afshar, Yanjun Gao, Deepak Gupta, Emma Croxford, and Dina Demner-Fushman. 2024. On the role of the umls in supporting diagnosis generation proposed by large language models. *Journal of Biomedical Informatics*, 157:104707.

Eman Aloraini, Hozaifa Kassab, Ali Hamdi, and Khaled Shaban. 2025. Lexisem: A re-ranker balancing lexical and semantic quality for enhanced abstractive summarization. *Neurocomputing*, page 130816.

Aleka Melese Ayalew, Md Rabiul Hasan, Tapio Seppänen, and Mourad Oussalah. 2026. Large Language Models for Explainable Medical Text Summarization: A Systematic Literature Review. *WIREs Data Mining and Knowledge Discovery*, 16(2):e70089.

Bryan Cai, Sihang Zeng, Yucong Lin, Zheng Yuan, Doudou Zhou, and Lu Tian. 2023. Hierarchical pretraining for biomedical term embeddings. *arXiv preprint arXiv:2307.00266*.

Lang Cao, Qingyu Chen, and Yue Guo. 2026. Ehr-rag: Bridging long-horizon structured electronic health records and large language models via enhanced retrieval-augmented generation. *arXiv preprint arXiv:2601.21340*.

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*, pages 335–336.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.

Jooyeon Lee, Luan Huy Pham, and Ozlem Uzuner. 2024. Enhancing consumer health question reformulation: Chain-of-thought prompting integrating focus, type, and user knowledge level. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pages 220–228.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4228–4238.

Khalil Mrini, Franck Dernoncourt, Walter Chang, Emilia Farcas, and Ndapandula Nakashole. 2021. Joint summarization-entailment optimization for consumer health question understanding. In *Proceedings of the second workshop on natural language processing for medical conversations*, pages 58–65.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Subhey Sadi Rahman, Md Adnanul Islam, Md Mahbub Alam, Musarrat Zeba, Md Abdur Rahman, Sadia Sultana Chowa, Mohaimenul Azam Khan Raiaan, and Sami Azam. 2026. Hallucination to truth: a review of fact-checking and factuality evaluation in large language models. *Artificial Intelligence Review*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Delbrouck, and 1 others. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2025. On positional bias of faithfulness for long-form summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8791–8810.

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

Shweta Yadav, Mourad Sarrouti, and Deepak Gupta. 2021. Nlm at mediqa 2021: Transfer learning-based approaches for consumer question and multi-answer summarization. In *proceedings of the 20th workshop on biomedical language processing*, pages 291–301.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2025. Morqa: Benchmarking evaluation metrics for medical open-ended question answering. *arXiv preprint arXiv:2509.12405*.

Nan Zhang, Yusen Zhang, Wu Guo, Prasenjit Mitra, and Rui Zhang. 2023. Famesumm: Investigating and improving faithfulness of medical summarization. *arXiv preprint arXiv:2311.02271*.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, and 1 others. 2025. Automating expert-level medical reasoning evaluation of large language models. *npj Digital Medicine*.

Kaiwen Zuo and Yirui Jiang. 2024. Medhallbench: A new benchmark for assessing hallucination in medical large language models. *arXiv preprint arXiv:2412.18947*.