U-GraphRAG: Evidence-Support Estimation in a Medical GraphRAG System

Yongfan Wang[1,*], Katarzyna Kaczmarek-Majer[2], Michal Chojnicki[3], Pawel Burchardt[4], Yanwu Ren[1] and Marek Z. Reformat[1,5,*]

[1]University of Alberta, 116 St and 85 Ave Edmonton, Alberta T6G 2R3, Canada
[2]Systems Research Institute Polish Academy of Sciences, Newelska 6, 01-147 Warsaw, Poland
[3]The Department of Medical Biology, Poznań University of Medical Sciences, 60-806 Poznań, Poland
[4]Department of Hypertension, Angiology and Internal Medicine, Poznań University of Medical Sciences, 61-848 Poznań, Poland
[5]Information Technology Institute, University of Social Sciences, 90-113 Łódź, Poland

Abstract

GraphRAG systems improve retrieval-augmented generation by organizing external knowledge as entities and relations. However, they often provide limited indication of whether retrieved evidence sufficiently supports a generated answer. This limitation is especially important in medicine, where unsupported or weakly supported responses may have serious consequences. This paper introduces U-GraphRAG, an uncertainty-aware GraphRAG framework for estimating response-level evidence support. The method constructs a hierarchical uncertainty-aware knowledge graph from medical documents, extracts relational triples using multiple LLMs, assigns logit-derived interval confidence indicators to triples, and combines these values with source-level trust, recurrence-based evidence support, and query-level relevance. These signals are propagated using fuzzy aggregation operators to obtain an evidence-support score. Preliminary experiments on synthetic answerable and unanswerable questions derived from chronic coronary syndrome guidelines show that Product–Max and Min–Max aggregation provide the strongest discrimination. The resulting score should be interpreted as evidence support, not as a calibrated probability of answer correctness.

Keywords

medical GraphRAG, evidence support, uncertainty propagation, knowledge graphs, fuzzy aggregation

CEUR-WS.org/Vol-4339/paper_4.pdf

1. Introduction

Large language models (LLMs) are increasingly used as natural-language interfaces to external knowledge sources, often through Retrieval-Augmented Generation (RAG) [1]. Although RAG reduces free-form hallucination, it does not fully address a key usability problem: users often cannot determine whether the retrieved evidence genuinely supports a generated answer, or whether the model has filled in missing information. This issue is important in GraphRAG [2], where graph-based retrieval improves structure and traceability but may also propagate errors caused by missing edges, noisy triples, or weakly supported paths. In medical question answering, indicating the degree of evidence support is critical, since unsupported or weakly supported responses may lead to misleading clinical interpretations. Related work has begun to study uncertainty in LLMs [3] and uncertainty-aware knowledge graph reasoning with LLMs [4].

In this paper, we propose *U-GraphRAG*, an uncertainty-aware GraphRAG methodology that augments each generated response with an explicit evidence support score. The method constructs an uncertainty-aware knowledge graph, denoted *uKG*, from source documents in a structure-preserving way. Documents are segmented into sections and represented hierarchically, while relational triples

EXPLIMED 2026 - Third Workshop on Explainable Artificial Intelligence for the medical domain - 15-17 July 2026, Bremen, Germany

*Corresponding author.
yongfan1@ualberta.ca (Y. Wang); kaczmar@ibspan.waw.pl (K. Kaczmarek-Majer); michalc@ump.edu.pl (M. Chojnicki); pburchardt@ump.edu.pl (P. Burchardt); yanwu@ualberta.ca (Y. Ren); reformat@ualberta.ca (M. Z. Reformat)
0009-0007-2782-4885 (Y. Wang); 0000-0003-0422-9366 (K. Kaczmarek-Majer); 0000-0003-3301-0662 (M. Chojnicki); 0000-0003-0353-3470 (P. Burchardt); 0009-0002-4558-4102 (Y. Ren); 0000-0003-4783-0717 (M. Z. Reformat)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-26

extracted from each section are attached to the corresponding structural node. Section summaries are also generated and used to support document- and section-level retrieval.

The proposed framework integrates three sources of uncertainty. First, source-level trust captures the reliability or provenance of the originating document or section. Second, triple-level confidence reflects both the quality of extracted triples, estimated from LLM token-level logits, and the amount of repeated evidence observed in the constructed *uKG*. Triples that reoccur across models, sections, or documents receive higher evidence support, whereas rare or isolated triples receive lower support. Third, query-level relevance is estimated from the semantic similarity between the user query and retrieved summaries or graph elements. These values are combined using fuzzy aggregation operators and propagated through a sequence of retrieval, aggregation, and response-scoring operations.

The proposed score should therefore be interpreted as an evidence-support score, not as a calibrated probability that the generated answer is correct.

To evaluate whether the resulting scores separate queries that are answerable from the graph from those that are not, we use synthetic question generation aligned with the documents used to build the *uKG*. This choice is motivated by the lack of public QA benchmarks with ground-truth uncertainty annotations for confidence-aware GraphRAG. We treat confidence scoring as a discriminative signal and compare combinations of t-norms and t-conorms using ROC-based analysis and classification metrics. Preliminary results show that Product-Max and Min-Max aggregation achieve the strongest discrimination, suggesting that aggregation design directly affects the usefulness of uncertainty scores.

In summary, we introduce an uncertainty-aware GraphRAG methodology that: (i) constructs a hierarchical knowledge graph of document-derived triples annotated with interval-valued confidence indicators computed from multiple LLMs' output-layer logits; (ii) incorporates source-level trust, triple-level evidence support, and query-level relevance; (iii) preserves document structure for source-aware retrieval and traceability; (iv) aggregates and propagates uncertainty through the response-scoring pipeline; and (v) evaluates fuzzy aggregation operators on a synthetic medical QA benchmark.

2. Background and Related Work

Knowledge Graphs. Knowledge graphs (KGs) provide structured and interpretable representations of knowledge through explicit entities and relations [5]. The relational structure supports good traceability, symbolic reasoning, and interpretable inference, making them well suited for explainable retrieval and reasoning systems [6].

RAG Systems Retrieval-augmented generation (RAG) frameworks combine external knowledge retrieval with LLM-based response generation to reduce hallucinations and improve factual grounding [1, 7]. In conventional RAG systems, retrieval is performed over unstructured or weakly structured text chunks, most commonly using an embedding-based similarity search strategy [8], providing limited awareness of document structure and weak support for multi-hop reasoning[9].

Existing RAG and GraphRAG systems primarily focus on improving retrieval and generation quality, while uncertainty is often implicit in retrieval similarity scores or omitted from the response-generation pipeline. Conversely, uncertainty-aware KG reasoning typically focuses on graph inference or link prediction rather than response-level support estimation in a GraphRAG setting. *U-GraphRAG* addresses this gap by representing uncertainty at multiple levels — source provenance, extracted triples, recurrence-based graph evidence, and query relevance — and propagating these signals to produce an explicit evidence-support score for generated responses.

3. Construction of Uncertain Knowledge Graph (*uKG*)

The proposed *U-GraphRAG* system, Fig. 1, is built around an uncertainty-aware knowledge graph, denoted *uKG*. Source documents are segmented according to their section and subsection structure, and relational triples are extracted from the smallest textual units containing substantive content.

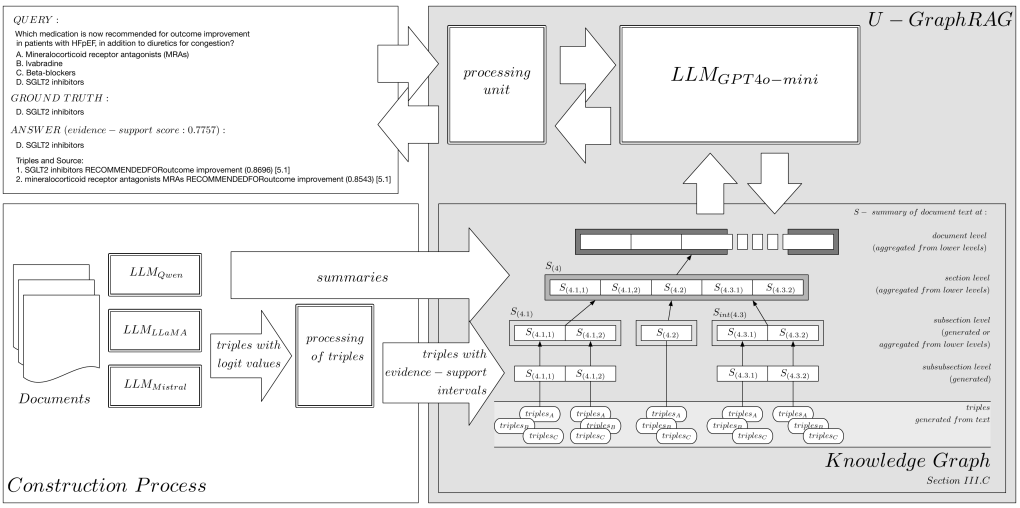


Figure 1: *U-GraphRAG* framework: construction process, knowledge graph *uKG*, exemplary usage.

This structure-aware construction preserves document hierarchy, improves retrieval granularity, and supports traceability from generated answers back to source sections.

3.1. From Text to Triples

LLM-based Text Analysis The basic building block of a KG is a triple denoted $\langle h, r, t \rangle$, where h and t correspond to the head and tail entities (nodes) and r denotes the semantic relation that connects them (edge). We deploy an ensemble of language models *LLMs* configured for information extraction to identify all entities in the text and explicit and implicit relations between them, i.e., to construct triples.

The triple extraction is performed using three *LLM*s: `Llama-3.1-8B-Instruct` [10], `Qwen3-8B` [11] and `Mistral-7B` [12]. The use of multiple models is intended to reduce the risk of missing relevant entities or relations due to model-specific biases or limitations. Three models process each textual unit. All output triples are combined, Section 3.2.

Each generated triple $\langle h, r, t \rangle$ reflects an interpretable and meaningful relationship grounded in the document. That enables the systematic transformation of unstructured text into a structured KG.

'Confidence' in Triples – Process Overview To quantify confidence in extracted knowledge, each relational triple $\langle h, r, t \rangle$ is associated with an evidence-support score derived from the language model's internal token-level logits during the extraction process.

For each candidate triple generated by an individual *LLM*, we capture the token-level logits associated with the sequences of tokens that constitute the head entity h, relation r, and tail t. We build vectors $\boldsymbol{\ell}^h = \{l_i^h\}$ for $i = 1, \ldots, i_h$ where i_h is a number of tokens composing the head. Similarly, we build $\boldsymbol{\ell}^r$, and $\boldsymbol{\ell}^t$. To derive an interval-valued confidence indicator for each triple as an indicator of the triple's quality, we define an approach based on relative logit magnitudes, with lower and upper bounds computed as follows:

$$lower = \frac{min_i(l_i^r)}{max_{i,j,k}(l_i^h, l_j^r, l_k^t)} \qquad upper = \frac{min_i(l_i^r)}{max_i(l_i^r)}$$

The upper bound reflects the internal consistency of the relation tokens, while the lower bound captures the strength of the relation relative to the most confident token across the entire triple. Together, the interval provides a comparative confidence signal indicating how strongly the model supports the generated triple as a coherent semantic unit, Fig. 2.

Both bounds use the minimum relation logit as the numerator, reflecting the observation that relation generation is often more uncertain than entity identification, especially for implicit relations. The denominators normalize the scores using the maximum logit within the relation or the entire triple,

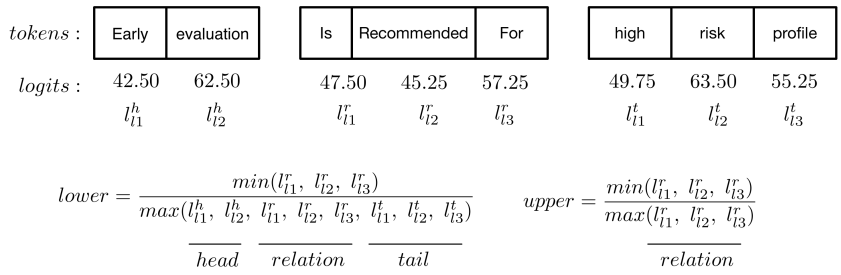


Figure 2: Calculation of lower and upper bounds for an interval-valued triple-confidence indicator; for the shown values: $lower = 0.713$ and $upper = 0.790$.

producing values in the range $[0, 1]$. The final evidence-support score is computed as the average of the lower and upper bounds and stored as a property of the relation in the *uKG*.

Importantly, this confidence should not be interpreted as a calibrated probability of correctness. Instead, it serves as a comparative indicator of the model's preference strength under the imposed prompt and contextual constraints, emphasizing internal consistency and relative preference structure rather than probabilistic certainty. When multiple extraction models are applied, evidence-support scores are computed independently for each model and subsequently combined. The logit-derived interval is used as a relative, model-internal preference indicator; it is not assumed to be calibrated across models or interpretable as a probability.

3.2. Final Construction Step: Integration

The extracted triples form the basic graph elements of our *uKG*. To preserve the **structure of source documents**, we first apply hierarchical chunking and extract triples from fine-grained structural sections. For each section, an *LLM* also generates a summary using only the raw text of that section. These summaries are stored in dedicated summary nodes, which act as semantic anchors connecting the triples extracted from the same section. The resulting *uKG* has a tree-like hierarchy, as shown in the lower-right corner of Fig. 1. The document is represented by the root node, while content sections correspond to leaf summary nodes. Each leaf is connected both to its extracted triples and to its parent node in the document hierarchy. Internal nodes connect to child summaries, other internal nodes, and ultimately to the root.

Leaf summaries are recursively propagated upward: summaries are concatenated at intermediate nodes and then further aggregated until reaching the root. The root contains a document-level summary derived from all leaf summaries. This summary supports the first retrieval stage by helping identify the most relevant document/section for a user query. In this way, the *uKG* integrates triples, summaries, and hierarchy while preserving document structure and improving interpretability and traceability.

Because several language models extract triples from the same text, redundant entity representations may arise. To consolidate semantically equivalent entities, entity nodes are embedded and merged when their cosine similarity exceeds a predefined threshold. During merging, all relations are preserved, including multiple relations between the same pair of nodes. These repeated relations are then used as evidence for trust estimation: triples that reoccur across models, sections, or source documents receive higher support, while rare or isolated triples receive lower support.

In summary, the *uKG* captures **trust** both at the **source level**, through the reliability of the originating document or section, and at the **triple level**, through repeated evidence observed during graph construction. The resulting *uKG* is a weighted, hierarchical, and interconnected knowledge base for the proposed *U-GraphRAG* system.

Note: All triples and summaries generated by *LLM*s employ deterministic decoding, with temperature set to 0 and $top\text{-}k = 1$, ensuring stable output given fixed model parameters and inputs.

4. GraphRAG System with Uncertainty

A brief description of the proposed *U-GraphRAG* system is presented here. It starts with details on the system's architecture, then clarifies how evidence-support scores of the obtained response are calculated.

4.1. Architecture Overview

The developed *U-GraphRAG* system includes an uncertainty-aware knowledge graph *uKG* constructed from documents using multiple LLMs (Section 3), and `GPT-4o-mini` [13] as the system's LLM, whose primary task is to accept queries and generate responses, Fig. 1. Special prompts are created, so the LLM performs several tasks: 1) determining the similarity between the user's query and document summaries linked to *uKG* nodes; 2) retrieving from *uKG* summaries whose semantic similarity is larger than a set threshold, together with the associated triples; and, at the time of generating a response, 3) identifying the triples that contribute mostly to the response.

The key task of U-GraphRAG is to estimate the degree to which retrieved graph evidence supports a generated response. These values estimate whether the retrieved *uKG* contains sufficient evidence to support the generated response. They are calculated using the proposed uncertainty-propagation pathway based on source/triple-level trust, triple quality, and the relevance of a query to the summarization nodes.

4.2. Uncertainty of Responses

In general, a single response of the *U-GraphRAG* system can be constructed based on an M number of subsections from different documents. It means that the information associated with several nodes of *uKG* composes the knowledge retrieved.

For each subsection $S_{(j)}$, where $j = 1, \dots, M$, the system's LLM pulls the summary of the subsection and the associated triples generated based on the text of that subsection. As a result, we have the following values representing different pieces of information:

- semantic similarity between the user's question and summary: $Sem_Sim_{S_{(j)}}$,
- triple-level evidence-support values for the N triples selected for generating a response, i.e., $Triple_Evidence_Support_{j,t_i}$ for $i = 1, \dots, N$,
- source/triple trust levels $Triple_Trust_{j,t_i}$ associated with each triple t_i for $i = 1, \dots, N$.

The value $Triple_Trust_{j,t_i}$ combines source-level trust and graph-level evidence support. Source-level trust *Source_Trust* reflects the reliability of the original document or section, while graph-level evidence support *Evidence_Support* reflects the recurrence of semantically equivalent triples across models, sections, or documents. In the present implementation, *Evidence_Support* is derived from the recurrence of semantically equivalent triples after entity merging and normalized to the interval $[0, 1]$.

$$Triple_Trust_{j,t_i} = \mathbb{T}(Source_Trust_j, Evidence_Support_{t_i}). \tag{1}$$

These values are used to determine an evidence-support score that reflects the presence of suitable information in the subsection. The calculations are as follows:

$$Triple_Evidence_Support_{S_{(j)}} = \mathbb{OWA}(\ \mathbb{T}(Triple_Evidence_Support_{j,t_i}, Triple_Trust_{j,t_i})\ ) \tag{2}$$

for $i = 1, \dots, N$. And then

$$SubSection_Support_{S(j)} = \mathbb{A}(Sem_Sim_{S(j)}, Triple_Evidence_Support_{S(j)}), \tag{3}$$

where $\mathbb{A}$ is instantiated as either a t-norm $\mathbb{T}$ or a t-conorm $\mathbb{S}$. Please note that the operator $\mathbb{T}$ or $\mathbb{S}$ can be applied to determine support associated with an individual subsection j. Further, the final score that represents the evidence support for the response is:

$$Final_Evidence_Support = \mathbb{S}_{j=1,\dots,M}(SubSection_Support_{S_{(j)}}) \tag{4}$$

In the calculations, we use three operators: $\mathbb{OWA}$, $\mathbb{T}$, and $\mathbb{S}$. The interesting question is which operators to apply. We address this question by investigating different combinations of $\mathbb{T}$ and $\mathbb{S}$ while, for simplicity, keeping $\mathbb{OWA}$ as the mean at all times; see Section 5 for details.

5. Experiments

This section presents experiments that we conducted using the *U-GraphRAG* to demonstrate and investigate its performance.

5.1. Synthetic Data Generation

We use synthetic data for evaluation. The existing QA benchmarks lack suitable uncertainty annotations and do not provide sufficient coverage of medical documents. The questions are generated directly from the same recommendations used to construct the *uKG*. It enables controlled evaluation of whether the proposed evidence-support scores distinguish supported from unsupported responses. Inspired by synthetic QA generation approaches used in RAG evaluation frameworks such as RAGAS [14], we employ NotebookLM to generate multiple-choice questions from the source guidelines. Prompt constraints are applied to produce questions together with their corresponding correct answers.

5.2. Experiment Overview

The experiments examine whether the evidence-support scores produced by *U-GraphRAG* meaningfully reflect the degree to which the generated responses are supported by the information contained in the *uKG*.

The *U-GraphRAG* includes the *uKG* constructed by processing *the 2024 ESC Guidelines for the Management of Chronic Coronary Syndromes*. It was developed by the task force for the management of chronic coronary syndromes of the European Society of Cardiology (ESC). It has been endorsed by the European Association for Cardio-Thoracic Surgery (EACTS) [15].

A set of questions has been generated from the ESC guidelines using the data generation process (Section 5.1). A total of 100 multiple-choice questions were created: 50 answerable based on the content of the document and 50 intentionally unanswerable. In the following, these questions are referred to as 50 positive and 50 negative ones.

For answerable questions, high evidence-support scores are expected, indicating that the retrieved *uKG* contains information relevant to answering the question. For unanswerable questions, low scores are expected.

To investigate the influence of different fuzzy operators, the following combinations of $\mathbb{T}$ and $\mathbb{S}$ are applied and evaluated: $\langle max - max \rangle$, $\langle min - max \rangle$, $\langle product - max \rangle$, and $\langle probsum - max \rangle$, for Eqs. (3) and (4). We use the *max* operator in Eq. (4) because it aligns better with human intuition: the score of the subsection with the strongest support should dominate the final response score.

5.3. Analysis of Results

The distribution of the final evidence-support scores of the response obtained using different combinations of fuzzy operators is illustrated in Fig. 3. The choice of the operator substantially affects both the range and separation of the response evidence-support scores, indicating that uncertainty propagation is strongly dependent on the selected aggregation strategy.

As can be seen, $\langle product - max \rangle$ and $\langle min - max \rangle$ exhibit clearer separation between answerable and unanswerable score distributions. In contrast, $\langle max - max \rangle$ and $\langle probsum - max \rangle$ produce compressed score distributions concentrated at high values, suggesting that the aggregation processes are too aggressive. This distinction aligns well with an intuitive human understanding of score semantics, where higher scores indicate stronger evidence support for answerability of the question.

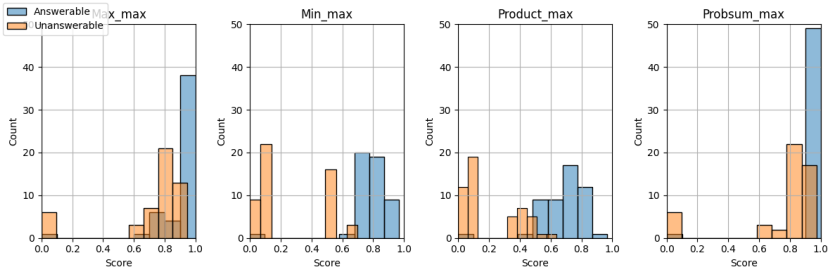


Figure 3: Distribution of final evidence-support scores obtained for different aggregation-operator combinations.

Another interpretation of the results involves categorizing responses by final score. It means that by establishing a threshold value, we can identify which answers with high scores match positive questions and which with low scores match negative questions.

Due to the variety of fuzzy (co)norm operators, which yield different score distributions with varying ranges and shapes, a fixed threshold (e.g., 0.5) is not appropriate for all settings. Therefore, separate thresholds were determined for each configuration using ROC-based analysis (Section 6). In particular, the selected threshold maximizes the trade-off between correctly identifying answerable and unanswerable cases. The resulting classification performance is summarized in Table 1.

Table 1
Classification of U-GraphRAG responses

	Threshold optimal	**Performance Metrics**				
		TP	***FP***	***FN***	***TN***	***F1***
Max-Max	0.90	38	5	12	45	0.82
Min-Max	0.61	49	3	1	47	0.96
Product-Max	0.52	48	2	2	48	0.96
Probsum-Max	0.96	40	3	10	47	0.86

These results should be interpreted as exploratory answerability-discrimination results. The questions are synthetically generated from the same guideline used to construct the uKG, and thresholds are selected on the same dataset used for reporting. Larger corpora, harder negative questions, external QA benchmarks, and similarity-only or GraphRAG baselines are left for future work.

6. Discussion

The results show that aggregation design has a substantial effect on evidence-support scores. Max–Max and Probsum–Max tend to compress scores toward high values, while Min–Max and Product–Max provide clearer separation between answerable and unanswerable questions. This distinction is important in medicine, where false support is especially concerning: unsupported responses assigned high evidence-support scores may mislead users. ROC analysis further supports this interpretation. Min–Max is preferable when false positives must be minimized, while Product–Max provides a more balanced trade-off by achieving high recall at a small false-positive rate.

Threshold Selection. To identify suitable thresholds for separating supported from unsupported responses, analysis of the Receiver Operating Characteristic (ROC) curve [16] was applied. We use ROC analysis to select thresholds for each operator configuration, since different fuzzy operators produce scores with different ranges and distributions.

The ROC curves for each combination of (co)norms are presented in Fig. 4. In the medical domain, false positives are particularly concerning because unsupported or incorrect responses assigned high evidence-support scores may be misleading and potentially dangerous. From this perspective, the $\langle min - max \rangle$ and $\langle product - max \rangle$ combinations provide the most favorable trade-off between identifying answerable questions and minimizing false confidence in unsupported responses. The following observations can be made from Fig. 4:

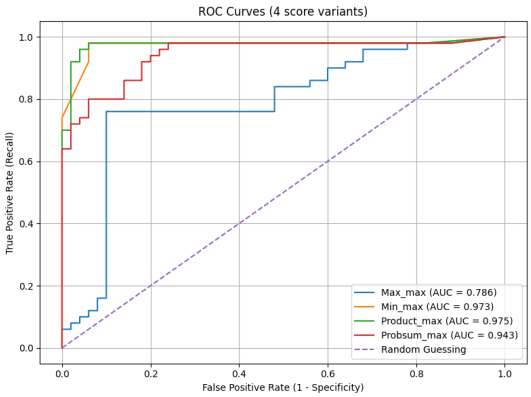


Figure 4: ROC curves for different aggregation-operator combinations.

- $\langle min - max \rangle$ (orange curve) is relatively high even when FPR is zero, it starts with $TPR = 0.77$ at $FPR = 0$: this configuration is preferable when false positives must be avoided;
- $\langle product - max \rangle$ (green curve) jumps to $TPR = 0.98$ at a small value of $FPR = 0.06$: if one can tolerate $\sim 6\%$ False Positives, then it has a high recall.

Borderline Examples. Let us examine some responses of *U-GraphRAG* when the pair $\langle product{-}max \rangle$ is used. While most answerable questions receive high evidence-support scores and most unanswerable questions receive low scores, the following examples illustrate issues near the decision threshold.

The first example is an answerable question: *When managing late graft failure after CABG, what is the preferred approach for revascularization?* The ground truth is *PCI of the bypassed native artery (D)*, whereas the model answers *PCI of the failing bypass graft (A)* with the score of 0.53, slightly above the threshold. Ideally, such responses should receive lower scores to indicate uncertainty better and prompt user caution. Although the response is incorrect, the *semantic similarity* of 0.70 and a *triple evidence-support score* of 0.7544 indicate that partially relevant evidence was retrieved from the *uKG*. This example highlights that the obtained evidence-support scores reflect the strength of evidence rather than the correctness of the direct answer.

The second borderline example illustrates the opposite situation. For the question *For patients with Heart Failure and preserved Ejection Fraction (HFpEF), which...?*, the model correctly answers *Perfusion PET (D)*, but the obtained score is 0.48, slightly below the decision threshold. In this case, a correct answer is assigned a score that is just too low to be considered trustworthy. Despite the correct prediction, the *semantic similarity* of 0.7 and a *triple evidence-support score* of 0.6902 suggest only partial evidence support from the retrieved graph information. This case demonstrates that correct responses may still receive relatively low evidence-support scores when supporting evidence is insufficient or weakly represented in the *uKG*.

These examples also illustrate that the proposed score should be interpreted as an evidence-support score rather than as a calibrated probability that the generated answer is correct.
Further, they motivate presenting the score together with retrieved triples and source sections, rather than using it as an automatic correctness guarantee.

Trust in Source of Information. To illustrate the role of source-level trust, we conducted a small sensitivity test in which the trust value of Section 6.4 was reduced from 1.0 to 0.5. The response-support scores for affected questions decreased accordingly, showing that source reliability can influence response-level support through the proposed propagation pathway.

In future deployments, source trust could encode document type, clinical evidence level, guideline status, recency, or institutional provenance.

In general, the trust values could represent multiple things; it depends on how *uKG* is constructed. The confidence in different information sources – source level – is the most straightforward indicator of trust. Yet, extracting triples from other sources, such as clinical trials, may be associated with varying

levels of confidence in the presented outcomes. A significant 'origin' of trust could be associated with the processing of patients' information and with linking it to the strength of evidence or the commonality/uniqueness of different bits of information – triple level, such as symptoms, and treatments.

The present evaluation is preliminary. Since the questions are synthetically generated from the same guideline used to construct the uKG, the results should be interpreted as an initial test of answerability discrimination rather than as evidence of general clinical QA performance.

7. Conclusion

This paper presented *U-GraphRAG*, an uncertainty-aware GraphRAG framework for estimating evidence support in medical question answering. The method constructs a hierarchical *uKG*, annotates extracted triples with logit-derived confidence indicators and recurrence-based evidence support, and propagates source trust, triple-level evidence, and query relevance using fuzzy aggregation. Preliminary experiments on synthetic guideline-based questions show that Product–Max and Min–Max aggregation best separate answerable from unanswerable cases. The resulting score estimates whether sufficient supporting evidence is present in the retrieved graph; it is not a calibrated probability of answer correctness.

Future work will evaluate larger medical corpora, stronger baselines, calibrated uncertainty estimates, and human interpretation of evidence-support scores.

Acknowledgments

Katarzyna Kaczmarek-Majer acknowledges support from the project 'ExplainMe: Explainable Artificial Intelligence for Monitoring Acoustic Features extracted from Speech' (FENG.02.02-IP.05-0302/23) carried out within the First Team programme of the Foundation for Polish Science co-financed by the European Union under the European Funds for Smart Economy 2021-2027 (FENG).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[2] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From local to global: A graph rag approach to query-focused summarization, arXiv preprint arXiv:2404.16130 (2024).

[3] F. Ye, M. Yang, J. Pang, L. Wang, D. F. Wong, E. Yilmaz, S. Shi, Z. Tu, Benchmarking llms via uncertainty quantification, arXiv preprint arXiv:2401.12794 (2024).

[4] B. Ni, Y. Wang, L. Cheng, E. Blasch, T. Derr, Towards trustworthy knowledge graph reasoning: An uncertainty aware perspective, in: Proceedings of the AAAI, volume 39, 2025, pp. 12417–12425.

[5] L. Ehrlinger, W. Wöß, Towards a definition of knowledge graphs., SEMANTiCS 48 (2016) 2.

[6] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al., Knowledge graphs, CSUR 54 (2021) 1–37.

[7] K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, Retrieval augmentation reduces hallucination in conversation, arXiv preprint arXiv:2104.07567 (2021).

[8] K. Guu, K. Lee, Z. Tung, P. Pasupat, M. Chang, Retrieval augmented lm pre-training, in: ICML, 2020, pp. 3929–3938.
[9] M. Yasunaga, J. Leskovec, P. Liang, Linkbert: Pretraining language models with document links, arXiv preprint arXiv:2203.15827 (2022).
[10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
[11] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).
[12] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).
[13] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al., Gpt-4o system card, arXiv preprint arXiv:2410.21276 (2024).
[14] S. Es, J. James, L. E. Anke, S. Schockaert, Ragas: Automated evaluation of retrieval augmented generation, in: 18th Conf. of the European Chapter of the ACL: System Demons., 2024, pp. 150–158.
[15] C. Vrints, F. Andreotti, K. C. Koskinas, X. Rossello, M. Adamo, J. Ainslie, A. P. Banning, A. Budaj, R. R. Buechel, G. A. Chiariello, et al., 2024 esc guidelines for the management of chronic coronary syndromes: developed by the task force for the management of chronic coronary syndromes of the european society of cardiology (esc) endorsed by the european association for cardio-thoracic surgery (eacts), European heart journal 45 (2024) 3415–3537.
[16] T. Fawcett, An introduction to roc analysis, Pattern recognition letters 27 (2006) 861–874.