Evaluating Knowledge Graph-Grounded Retrieval-Augmented Generation for Biomedical Question Answering: A Preliminary Analysis

Roberta Angioni[1], Sergio Consoli[2,*], Silvia M. Massa[1], Daniele Riboni[1] and Diego Reforgiato Recupero[1]

[1]Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy
[2]European Commission, Joint Research Centre (JRC), Ispra, Italy

Abstract

Large Language Models (LLMs) often produce plausible but factually incorrect outputs in specialized biomedical domains, and conventional Retrieval-Augmented Generation (RAG) methods relying on unstructured text retrieval may miss essential structured relationships. Here we present a Knowledge Graph-grounded RAG (KG-RAG) pipeline that extracts structured knowledge from biomedical ontologies via Linked Open Data, combining an ensemble of Named Entity Recognition models, ontology linking, RDF triple retrieval, and cross-encoder reranking to provide context for LLM-based question answering. Evaluated on a subset of PubMedQA and an expert-curated dataset concerning Virtual Human Twins (VHTs), our approach demonstrates the feasibility of ontology-grounded retrieval, with smaller LLMs showing notable improvements when augmented with KG-derived context. These findings suggest that KG-RAG can enhance reliable and resource-efficient biomedical question answering, particularly in privacy-sensitive or computationally constrained environments.

Keywords

Retrieval-Augmented Generation, Knowledge Graphs, Biomedical Question Answering, Named Entity Recognition, Biomedical Ontologies

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation, achieving impressive performance across a wide range of tasks [1]. However, their application to specialised domains such as biomedicine remains challenged by well-documented limitations, including the tendency to generate plausible yet factually incorrect content — a phenomenon known as hallucination [2, 3] — and the inability to access up-to-date or domain-specific knowledge not seen during pre-training.

Retrieval-Augmented Generation (RAG) [4] has emerged as a prominent paradigm to mitigate these shortcomings by conditioning LLM responses on external information retrieved at inference time. In the biomedical domain, RAG systems have shown promise in improving the factual grounding of generated answers [5, 6]. The standard RAG approach retrieves relevant text passages from unstructured document corpora based on vector similarity, and injects them as context into the generation prompt. While effective in many scenarios, this document-centric retrieval has inherent limitations when applied to professional domains: vector similarity does not always correlate with knowledge relevance, and the retrieved passages may lack the structured relationships — hierarchical, causal, or compositional — that are essential for sound biomedical reasoning [7].

Knowledge Graphs (KGs) [8] offer a complementary source of structured, curated knowledge. In the biomedical domain, large-scale ontologies such as SNOMED-CT, MeSH, ICD-10, Gene Ontology, and NCIT encode millions of concepts and their interrelationships in machine-readable form, and are widely available as Linked Open Data through SPARQL endpoints. Recent work has begun exploring

RAGE-KG 2026: Retrieval-Augmented Generation Enabled by Knowledge Graphs, October 25–29, 2026, Bari, Italy
*Corresponding author.
roberta.angioni@unica.it (R. Angioni); sergio.consoli@ec.europa.eu (S. Consoli); silviam.massa@unica.it (S. M. Massa); riboni@unica.it (D. Riboni); diego.reforgiato@unica.it (D. Reforgiato Recupero)
© 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 2026-11-04

CEUR-WS.org/Vol-4312/paper25.pdf

the integration of KGs into RAG pipelines to enhance the quality and reliability of generated responses. The Knowledge Augmented Generation (KAG) framework [9] demonstrated significant improvements over standard RAG in multi-hop question answering by jointly leveraging KGs and vector retrieval with a logical-form-guided reasoning engine. Similarly, PLATO-KAG [10] showed the benefits of grounding generation in external structured knowledge for improved factual accuracy.

However, existing KG-enhanced RAG approaches have been developed primarily for general-domain or conversational settings, and their application to the biomedical domain — where the richness of available ontologies and the Linked Open Data ecosystem provides unique opportunities — remains largely unexplored. The biomedical domain is particularly suited for KG-grounded RAG due to the availability of authoritative, standardised terminologies that can serve as both a disambiguation tool (through Named Entity Linking) and a source of structured factual context (through triple extraction from ontology graphs).

In this paper, we present a preliminary study of a KG-grounded RAG architecture for biomedical question answering. Our pipeline leverages an ensemble of Named Entity Recognition (NER) models to identify biomedical concepts in the input questions, links them to standard ontologies via SPARQL endpoints, retrieves and linearises the associated RDF triples, and uses a cross-encoder reranker to select the most relevant knowledge before injecting it into the LLM prompt. We evaluate this approach on two datasets from distinct health domains — a subset of PubMedQA covering classical biomedical research questions, and a set of questions on Virtual Human Twins (VHT) representing an emerging interdisciplinary field — providing initial evidence on the feasibility and potential of ontology-grounded RAG for biomedicine.

The remainder of this paper is organised as follows. Section 2 describes the system architecture in detail. Section 3 presents the preliminary experimental evaluation. Section 4 discusses conclusions and future work directions.

2. System Architecture

The proposed system implements a KG-grounded Retrieval-Augmented Generation (KG-RAG) pipeline for biomedical question answering. Rather than retrieving passages from unstructured document corpora, it extracts structured knowledge from biomedical ontologies exposed as Linked Open Data (LOD), linearises the retrieved triples into natural-language context, and injects this context into the LLM prompt. The pipeline[1] comprises five sequential stages (Figure 1): NER, Entity Fusion and Filtering, Named Entity Linking and Context Extraction, Context Aggregation, and RAG-based Question Answering.

2.1. Named Entity Recognition

The first stage extracts biomedical entities from the input questions. Each question is assigned a unique sentence identifier (a consecutive integer, shared by identical questions so that duplicates are processed once) so that downstream stages can track provenance: it makes the recorded character offsets interpretable, confines duplicate and overlap resolution to annotations of the same question, and acts as the grouping key for aggregating per-entity contexts and for associating the generated answer with its input question. Entity extraction is performed by an ensemble of heterogeneous NER models spanning three architectural families: (i) transformer-based token-classification models via Hugging Face [11] pipelines, including *WikiNEuRaL*[2] [12] for multilingual entity types and *Medical-NER*[3] for fine-grained biomedical categories; (ii) span-based zero-shot models from the GLiNER [13] and GLiNER2 [14] families, which recognise entities conditioned on user-defined labels (e.g., "disease", "location"); and (iii) the NCBO BioPortal Annotator+ [15], which performs knowledge-driven matching

[1] A web-based demo of the system is available at: https://jrc-ai-biokg-rag.hf.space (weekdays, 08:00–20:00 CET), while the software code is available upon reasonable request from the authors.

[2] https://huggingface.co/Babelscape/wikineural-multilingual-ner

[3] https://huggingface.co/blaze999/Medical-NER

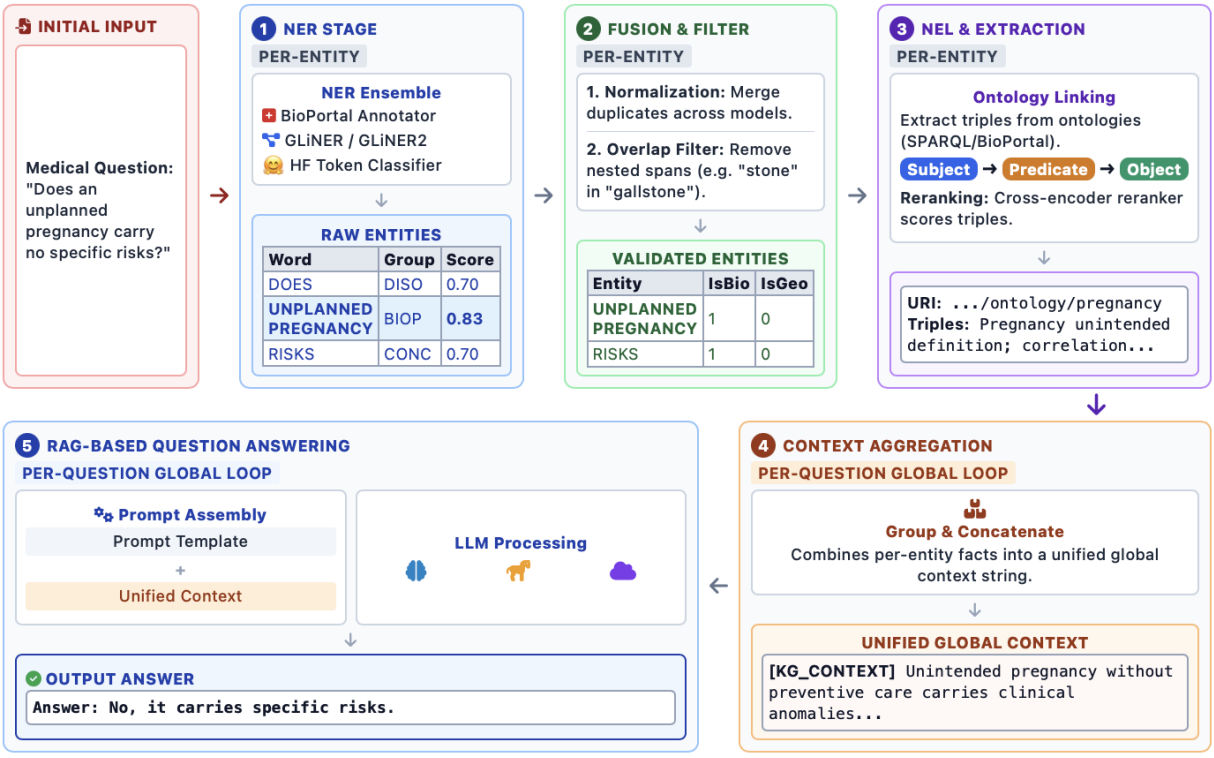


Figure 1: Architecture of the KG-RAG pipeline. The five-stage pipeline processes biomedical questions from entity extraction to knowledge-grounded answer generation.

against biomedical ontology concepts. For each entity, the system records the surface form, the entity type label assigned by the annotating model (the `entity_group` field in our implementation), confidence score (a model-specific value, whose differing scales are harmonised in the fusion stage described next), and character offsets (the start and end positions of the entity mention within the question string). Because the ensemble is heterogeneous, these labels are drawn from different inventories: *WikiNEuRaL* produces classic CoNLL-style types (PER, ORG, LOC, MISC); *Medical-NER* produces fine-grained clinical types (e.g., DISEASE_DISORDER, DOSAGE, LAB_VALUE, DURATION); the GLiNER models return the user-defined labels they are prompted with ("disease", "location", "date", "numerical value", "number"); and the BioPortal Annotator+ returns UMLS Semantic Groups (e.g., ANAT, CHEM, DISO, GENE, LIVB, PHEN, PHYS, PROC), with a generic biomedical label assigned when no semantic group is returned. These model-specific labels are subsequently normalised into the coarse biomedical/geographical distinction used by the downstream stages. Outputs from all models are stacked into a unified set of raw entities (block 1 in Figure 1).

2.2. Entity Fusion & Filtering

Since multiple models annotate the same text independently, their outputs contain redundant and conflicting annotations. The fusion stage resolves these through a multi-step strategy. First, model-specific offset inconsistencies are corrected: the BioPortal API returns 1-indexed positions, *Medical-NER* may prepend whitespace to entity starts, and GLiNER models may include leading spaces — all normalised to zero-indexed character spans. Second, confidence scores from different models are placed on a comparable scale. The transformer-based and GLiNER models already emit probabilistic confidences in $[0, 1]$, whereas the BioPortal Annotator+ returns unbounded and strongly skewed ontology-matching scores that are not probabilities; the latter are therefore monotonically remapped, through a logarithmic transformation followed by min–max normalisation, onto the $[0.7, 1.0]$ interval. The lower bound is deliberately aligned with the confidence threshold applied in the next step, so that dictionary-based matches against curated ontology labels and synonyms are retained as candidates and compete with the

neural annotations on a comparable scale, while the threshold prunes low-confidence predictions from the neural models. Third, a word-validation step removes entities whose surface forms do not correspond to actual words in the source text, and entities below a configurable confidence threshold (default 0.7) are discarded. Fourth, duplicates are resolved by retaining only the highest-scoring annotation for each unique span. Finally, a cross-inside resolution algorithm handles overlapping or nested entities (e.g., "carcinoma" inside "basal cell carcinoma"): a sliding window iterates over sorted annotations within the same sentence and marks lower-confidence overlapping spans for removal, prioritising entities that are linkable to KGs. Entities surviving this process are categorised as biomedical and form the validated entities dataset (block 2 in Figure 1).

2.3. Named Entity Linking & Context Extraction

The third stage (block 3 in Figure 1) resolves each validated entity to its canonical URI in the KG and extracts structured context. For biomedical entities, the system queries a private Virtuoso [16] SPARQL endpoint or the NCBO BioPortal API, searching across `skos:prefLabel`, `rdfs:label`, `skos:altLabel`, `skosxl:literalForm`, and `obo:hasRelatedSynonym` predicates. The search spans over 50 biomedical ontologies including SNOMED-CT, LOINC, ICD-10, MeSH, NCIT, CHEBI, GO, and MONDO [15]. Matched concepts are ranked by edge count, computed as the number of outgoing triples of each candidate concept in the union of the queried ontology graphs — that is, the degree of the candidate in the global KG, rather than in a subgraph induced by the other candidates of the same question— as a proxy for information richness, and the top-ranked concept is selected as the primary link; all alternative URIs across ontologies are also collected. This score is therefore a per-candidate prior that favours richly described concepts but is insensitive to the remaining entities mentioned in the question. Scoring candidates also by their relatedness within the subgraph induced by all candidates of a question would provide a more context-sensitive disambiguation signal, and we regard it as a natural extension of the present pipeline.

For each linked concept and its alternatives, the system retrieves all associated RDF triples — standard subject-predicate-object statements, literal values, and OWL restrictions — via SPARQL queries. The two query templates used, respectively, for concept lookup and for triple retrieval are reported in Appendix A. Triple collection is capped at a configurable threshold (default 10,000) to bound memory consumption. The retrieved triples are then linearised into textual passages by concatenating the human-readable labels of each subject, predicate, and object.

To distil relevant knowledge from this potentially large set of linearised triples, the system employs a cross-encoder reranker (mxbai-rerank-large-v1[4]) that scores each passage against the original question. The reranker first returns a candidate pool whose size grows with the number of linearised triples: all of them when they are at most ten, ten plus a tenth of them for larger sets, and up to 400 for the largest ones. A dynamic thresholding mechanism starts at a score of 0.80 and decreases in steps of 0.05 until a minimum number of relevant passages is retained, that is, until more than ten passages exceed the current threshold, or more than half of the candidates when the pool contains ten or fewer passages, ensuring that context quality is maintained while avoiding empty results; the reduction stops once the threshold falls below 0.20, so that passages judged clearly irrelevant are never admitted. When the target count is not reached even at that point, only the passages above the last threshold are retained, and the entity may contribute no context at all — the condition that determines the context-yielding subsets used in Section 3. The top-scoring passages constitute the extracted context for each entity.

Geographical entities are separately linked via the GeoNames[5] API but without triple extraction, since their role is limited to spatial grounding: this branch resolves place names to canonical GeoNames URIs and, more importantly, diverts them from biomedical linking, so that toponyms colliding with ontology labels do not inject spurious triples into the retrieved context. Geographical mentions are infrequent in the two question sets considered here, so this branch plays a marginal role in the reported

[4]https://huggingface.co/mixedbread-ai/mxbai-rerank-large-v1
[5]GeoNames: https://www.geonames.org

results. All SPARQL queries and API responses are persistently cached to avoid redundant computations across runs.

2.4. Context Aggregation

The fourth stage assembles the per-entity contexts into a unified knowledge context for each question. Entity contexts are grouped by question identifier and concatenated with newline separators, producing a single global context string per question. Empty entity contexts are ignored during aggregation. This global context represents the aggregated factual knowledge extracted from all ontology concepts identified within the question (block 4 in Figure 1).

2.5. RAG Question Answering

In the final stage (block 5 in Figure 1), the aggregated context is injected into a prompt template that instructs the LLM to reason step-by-step using the provided facts. The prompt follows a context-then-question structure, asking the model to produce a coherent explanation grounded in the supplied knowledge. When the global context exceeds the model's token limit, it is truncated to the first chunk via a token-based splitter; if no context is available for a question, the system falls back to a simpler prompt without external knowledge. The prompt is submitted to an LLM served on-premises (via a vLLM-based inference endpoint) and the generated answer is recorded. This design decouples the knowledge retrieval from the generation model, allowing different LLMs to be evaluated on the same retrieved contexts.

3. Preliminary Experiments

This section presents a preliminary experimental evaluation aimed at assessing the feasibility and potential of the proposed KG-RAG pipeline. The goal at this stage is not a comprehensive evaluation, but rather an initial investigation into whether grounding LLM responses in biomedical KGs can improve answer quality compared to a baseline without external knowledge. To this end, we selected the following two small datasets from distinct health domains, enabling us to probe the system's behaviour across different types of biomedical questions and terminology. The two datasets, namely *PubMedQA* and *Virtual Human Twin*, are publicly available through the Joint Research Centre Data Catalogue[6], respectively at https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/DIGLIFE/OPEN/RagEv-Bench/ds-PubMedQA_sample100.csv and https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/DIGLIFE/OPEN/RagEv-Bench/ds-VHT.csv.

PubMedQA. The first dataset is a subset of PubMedQA [17], a publicly available benchmark for biomedical research question answering built upon PubMed articles. PubMedQA contains questions constructed from paper titles, each answerable in a yes/no fashion, with ground-truth answers derived from the corresponding abstracts. This structure makes it well-suited for automated evaluation using standard classification metrics. For our preliminary experiments, we selected a context-yielding subset of 24 questions covering diverse biomedical topics (e.g., disease mechanisms, treatment efficacy, genetic factors), providing a controlled setting with well-established scientific terminology that can be reliably linked to biomedical ontologies. The subset was obtained automatically rather than by manual curation: we retained the questions for which the pipeline of Section 2 produced a non-empty context, that is, those containing at least one entity that was successfully linked to an ontology concept and yielded triples above the reranking threshold. This filter depends solely on the outcome of the retrieval stage and is therefore independent of the language models and of the correctness of their answers. In terms of label distribution, this 24-question subset comprises 9 questions with a "yes" ground-truth answer and 15 with a "no" ground-truth answer.

[6] https://data.jrc.ec.europa.eu/

Virtual Human Twin (VHT). The second dataset focuses on the VHT, a relatively new scientific topic concerning digital representations of different levels of human anatomy (e.g., cells, tissues, organs, or organ systems) for modelling states of human health or disease[7]. The VHT initiative represents a flagship European digital health strategy. This dataset comprises open-ended questions prepared by domain experts, ranging from conceptual questions requiring synthesis (e.g., *"What differentiates a Human digital twin from a digital twin?"*) to factual queries targeting specific information. Unlike the well-established terminology of PubMedQA, the VHT domain features non-homogeneous information coverage and interdisciplinary vocabulary spanning engineering, medicine, and computer science — posing a distinct challenge for entity recognition and ontology linking. For this evaluation, the dataset was reformulated as a binary decision task: the open-ended questions were converted into a forced-choice format with "yes" or "no" answers, producing an initially balanced dataset across the two outcomes. Subsequently, instances were selected based on their ability to leverage the RAG pipeline, retaining only the 36 questions for which informative context could be retrieved from the KG. As for PubMedQA, this filtering was automatic and applied the same criterion: a question was retained only if at least one of its entities was linked to an ontology concept and yielded triples above the reranking threshold. Questions whose vocabulary falls outside biomedical terminology are typically discarded by this criterion; for instance, *"What standard is ZigBee developed atop?"* contains no biomedical concept that can be linked to the ontologies, so no KG-derived context is produced and the augmented condition would coincide with the unaugmented one. As a consequence of this filtering step, the final evaluation subset is no longer class-balanced, containing 32 questions with a "yes" ground-truth answer and 4 questions with a "no" ground-truth answer.

We evaluate the pipeline on these two datasets, each representing complementary challenges in the biomedical domain. The choice of these two datasets was deliberate: PubMedQA provides a mature biomedical setting with standardised terminology well-represented in ontologies like SNOMED-CT and MeSH, while VHT represents an emerging interdisciplinary field where ontology coverage may be sparser and entity boundaries less clear-cut. Together, they allow us to compare ontology-grounded augmentation in a mature biomedical domain and in a more challenging emerging interdisciplinary domain. It is important to note that both evaluation sets were restricted to questions for which the retrieval pipeline produced informative KG-derived context. Therefore, the experiments assess the effect of knowledge-graph augmentation under successful retrieval conditions rather than the end-to-end coverage of the complete pipeline on unfiltered questions. Evaluating retrieval success, ontology coverage, and downstream answer quality jointly on the full datasets is left for future work. A further caveat concerns the statistical reliability of the reported figures. Each model–dataset configuration was executed once, with near-deterministic decoding (sampling at temperature 0.01) and without repetitions across random seeds, so no variability estimates or confidence intervals are reported; the values in Table 1 should therefore be read as indicative of qualitative trends rather than as precise performance estimates, and repeated runs with multiple seeds and confidence intervals on larger question sets are left for future work.

To evaluate the performance of the models on this decision task, we extract the prediction from the text following an explicit answer marker. The possible predictions can be represented as $m \in \{\text{'Yes'}, \text{'No'}, \text{'I don't know'}\}$, where if no explicit binary decision is found or if the model expresses uncertainty, the prediction defaults to 'I don't know'; to penalize these non-committal or uncertain responses without completely discarding them from evaluation, we introduce weighted metrics parameterized by a penalty factor $\alpha \in \{1/2, 1/3\}$. In this way, uncertainty always brings a penalty, although with a smaller impact than incorrect answers.

To distinguish abstentions from fully incorrect binary predictions, we define a weighted accuracy measure in which each abstention ($m =$ 'I don't know') contributes a fractional penalty (α) to the denominator:

[7] https://digital-strategy.ec.europa.eu/en/policies/virtual-human-twins

$$\text{Accuracy}_{\text{weighted},\alpha} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN}) + \alpha \cdot N_{\text{IDK}}} \tag{1}$$

where TP and TN are True Positives and True Negatives on valid binary predictions, and N_{IDK} is the total number of samples where the model predicted 'I don't know' or something other than the expected outcomes. In our test setup, as anticipated, we evaluate two different penalty values. To compute the F_1 score across both classes, we incorporate weighted confusion matrix counts separately for the 'Yes' and 'No' classes based on the ground-truth label g:

- **For the 'Yes' class:** An abstention (m = 'I don't know') on a target g = 'Yes' adds α to FN_{Yes}, whereas if g = 'No', it adds α to FP_{Yes}.
- **For the 'No' class:** An abstention on a target g = 'No' adds α to FN_{No}, whereas if g = 'Yes', it adds α to FP_{No}.

Accordingly, an abstention is treated as a partial misclassification in the one-vs-rest evaluation of both classes, with the severity controlled by the penalty factor α. The increment is α rather than a full unit by design: an incorrect binary prediction contributes one whole count to the corresponding FN or FP term, so choosing $\alpha < 1$ guarantees that abstaining is penalised less severely than answering incorrectly, symmetrically on the false-negative and false-positive sides, and consistently with the treatment of N_{IDK} in Eq. (1).

Using these adjusted confusion terms ($\text{TP}'_c, \text{FP}'_c, \text{FN}'_c$), we calculate precision (P'_c) and recall (R'_c) for each class $c \in \{\text{'Yes'}, \text{'No'}\}$ (TN'_c is not included in the computation for precision and recall), and compute the overall Weighted Macro F_1 as:

$$\text{Macro } F_{1,\text{weighted},\alpha} = \frac{F'_{1,\text{Yes}} + F'_{1,\text{No}}}{2}, \quad \text{where} \quad F'_{1,c} = \frac{2 \cdot P'_c \cdot R'_c}{P'_c + R'_c}. \tag{2}$$

To assess the impact of KG-derived context, we evaluate four autoregressive generative models and one encoder-only biomedical classifier, with parameter counts ranging from 0.6B to 1.7B. In the direct setting, the models answer using only the information encoded in their parameters. In the KG-augmented setting, the retrieved ontology context is supplied as additional input.

It is worth noting that for the autoregressive models, the KG-derived context is included directly in the generation prompt, while for PubMedBERT, the direct condition receives `{question} [SEP]` and the KG-augmented condition uses `{question} [SEP] Context: {kg_context}`. Since PubMedBERT always predicts one of the two target classification labels, it does not produce abstentions ($N_{IDK} = 0$).

Table 1
Performance comparison between direct generation (Dir) and KG-augmented generation (+KG) across models and datasets. Bold values indicate the better result between the two settings for each model–dataset configuration; lower values are preferred for N_{IDK}.

		Acc$_{w(1/2)}$ (%)		**Macro F$_{1,w(1/2)}$ (%)**		**Acc$_{w(1/3)}$ (%)**		**Macro F$_{1,w(1/3)}$ (%)**		N_{IDK}	
Model	**Dataset**	**Dir**	**+KG**	**Dir**	**+KG**	**Dir**	**+KG**	**Dir**	**+KG**	**Dir**	**+KG**
Qwen3-0.6B	PubMedQA	20.68	**37.50**	19.35	**27.27**	26.47	**37.50**	24.32	**27.27**	19	**0**
	VHT	4.44	**88.88**	6.45	**47.05**	5.55	**88.88**	9.00	**47.05**	27	**0**
PubMedBERT	PubMedQA	54.16	54.16	49.42	49.42	54.16	54.16	49.42	49.42	0	0
	VHT	50.00	**72.22**	43.75	**45.00**	50.00	**72.22**	43.75	**45.00**	0	0
Llama-3.2-1B	PubMedQA	37.50	**41.66**	27.27	**32.12**	37.50	**41.66**	27.27	**32.12**	0	0
	VHT	80.59	**84.50**	44.62	**48.56**	82.65	**84.90**	45.25	**48.70**	5	**1**
DeepSeek-R1-Distill-Qwen-1.5B	PubMedQA	38.88	**53.33**	30.43	**50.54**	43.75	**54.54**	33.33	**51.20**	12	**3**
	VHT	23.25	**65.67**	23.25	**48.91**	30.00	**67.34**	29.64	**49.88**	29	**4**
SmolLM2-1.7B	PubMedQA	**51.06**	36.37	**42.35**	28.07	**51.42**	37.50	**42.42**	28.23	**1**	4
	VHT	71.42	**72.22**	44.71	**45.00**	72.11	**72.22**	44.98	**45.00**	2	**0**

*N_{IDK} indicates the number of non-committal or uncertain responses ("I don't know").

Table 1 presents the experimental results. The performance of the LLMs in the direct generation setting, reported in the "Dir" columns, reveals significant fragility in predictions when the models have to rely solely on parametric memory. Very compact models are often uncertain about which answer to provide, leading to high abstention rates. For example, `Qwen3-0.6B`[8] abstains on 27 out of 36 questions in the VHT dataset and on 19 PubMedQA questions. Similarly, `DeepSeek-R1-Distill-Qwen-1.5B`[9] abstains on 29 VHT queries. These non-committal outputs can be interpreted as a kind of model indecision, and since they are penalized under the weighted accuracy formulation (Acc_w), the models with high abstention counts also report performance drops. Furthermore, a clear distinction emerges between domain-specialized and general-purpose architectures, as well as their underlying modeling paradigms: the domain-tuned encoder-only model `PubMedBERT`[10] achieves a baseline accuracy of 54.16% on PubMedQA with zero abstentions, reflecting its specialized pre-training on biomedical text, but its rigid classification-head nature may limit its flexibility across diverse tasks, hence its poorer performance on the VHT dataset. In contrast, some generalist autoregressive models such as `Llama-3.2-1B`[11] and `SmolLM2-1.7B`[12] perform remarkably well on this dataset in the baseline setting but they struggle more on PubMedQA: this discrepancy may stem from the nature of the two benchmarks, since PubMedQA queries target highly specific clinical research findings derived from specialized biomedical literature. Such specific findings are unlikely to be represented faithfully in the parameters of small models, potentially causing them to guess or abstain.

In contrast, when we inject structured context retrieved from biomedical ontologies, we obtain overall improvement in weighted accuracy and abstention rates for most model–dataset configurations, as shown in column "+KG" of Table 1. In particular, the retrieved context acts as an anchor during generation, reducing model uncertainty and mitigating abstentions: `Qwen3-0.6B` completely eliminates non-committal responses across both benchmarks ($N_{\text{IDK}} = 0$), raising its weighted accuracy on VHT from 4.44% to 88.88% ($\text{Acc}_{w(1/2)}$) and on PubMedQA from 20.68% to 37.50%. This particular figure must, however, be read together with the corresponding Macro F_1 of 47.05%. Since the VHT subset contains 32 "yes" and 4 "no" ground-truth answers, an accuracy of 88.88% coincides with what a degenerate classifier answering "yes" to every question would obtain, and the low Macro F_1 indicates that the "no" class is essentially not recovered. For this configuration the gain therefore reflects the suppression of abstentions in favour of affirmative answers rather than an improved ability to discriminate between the two classes; this is precisely why we report Macro F_1 alongside weighted accuracy, and why the class-imbalanced VHT subset requires the two metrics to be read jointly. Similarly, `DeepSeek-R1-Distill-Qwen-1.5B` reduces its VHT abstentions from 29 to 4, and its performance improves significantly in terms of weighted accuracy and F_1 scores.

In general, the integration of ontology-derived contexts brings improvements in most model–dataset configurations, suggesting that the approach can help lightweight LLMs compensate for limitations in their parametric knowledge. This gain is particularly evident on the VHT dataset, where the retrieved context appears to compensate for knowledge that may be less extensively represented in the models' parameters.

However, performance gains are not uniform across all model-dataset configurations, revealing key insights into how small language models interact with external context. In certain instances, such as `SmolLM2-1.7B` on PubMedQA, the introduction of RAG leads to a substantial degradation in performance compared to the direct baseline: one possible explanation is that some models rely heavily on the supplied context. When structured context is provided within the prompt, the model prioritizes the retrieved facts over its own internal parametric memory so, if the retrieved ontology information does not contain the precise fact required to answer the binary query even though it is relevant, the model becomes overly conservative. Rather than falling back on its internal pre-trained knowledge,

[8] Qwen3-0.6B, https://huggingface.co/Qwen/Qwen3-0.6B

[9] DeepSeek-R1-Distill-Qwen-1.5B, https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

[10] Specific PubMedBERT model version: *BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext-finetuned-pubmedqa*, available at: https://huggingface.co/blizrys/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext-finetuned-pubmedqa

[11] Llama-3.2-1B, https://huggingface.co/meta-llama/Llama-3.2-1B

[12] SmolLM2-1.7B, https://huggingface.co/HuggingFaceTB/SmolLM2-1.7B

which might have yielded a correct prediction in the unaugmented setting, the model either abstains or makes an incorrect inference, suggesting a trade-off between strict context grounding and parametric knowledge fallback.

4. Conclusions and Future Work

In this paper, we have introduced a KG-grounded RAG architecture for biomedical question answering. We conducted a preliminary evaluation on two biomedical datasets representing complementary domains: PubMedQA and VHT. Our experiments focused exclusively on lightweight models, and direct comparison with larger LLMs remains future work. The preliminary results indicate that KG-derived context can be beneficial for smaller models, whose limited parameter capacity may reduce their coverage of specialised biomedical knowledge. By supplying these models with authoritative, structured context retrieved from curated biomedical ontologies at inference time, the proposed KG-RAG architecture has the potential to substantially bridge the knowledge gap without requiring expensive fine-tuning or larger hardware. This finding has important implications for democratising generative AI in healthcare: in deployment scenarios where computational resources are limited — such as on-premises inference in hospitals, edge devices, or privacy-sensitive environments where data cannot leave the institution — small models augmented with KG-grounded RAG may offer a promising path towards resource-efficient biomedical question answering that would otherwise require access to large proprietary models.

At the same time, our empirical observations highlight important trade-offs inherent to ontology grounding in small LLMs. Although external factual context substantially reduced abstentions, some models appeared to rely excessively on the retrieved knowledge. When the retrieved context was incomplete or only indirectly relevant to the question, it could introduce noise and lead to incorrect inferences. These results should also be interpreted in light of the context-yielding selection procedure, since the present evaluation does not measure system performance on questions for which KG retrieval fails or produces insufficient context.

Looking ahead, the limited number of datasets considered in this preliminary study is a clear constraint on the generality of our findings, and broadening the evaluation is our primary next step: we will expand our evaluation of the system across other biomedical datasets of diverse nature and complexity, spanning established benchmarks as well as emerging interdisciplinary domains such as antimicrobial resistance, and neonatal and perinatal health, so as to verify whether the trends observed here hold beyond the two question sets used in this work. Future development directions will also include expanding the NER ensemble in the system to incorporate additional models, integrating graph-based reasoning mechanisms that exploit multi-hop relationships in the ontologies, including context-sensitive entity disambiguation that ranks candidate concepts within the subgraph induced by all entities detected in a question, supporting multilingual biomedical queries, extending the evaluation from the binary decision setting adopted here to open-ended answer generation, together with evaluation protocols suited to free-form outputs, and conducting further tests to analyze in greater detail how retrieved context interacts with the models' internal knowledge.

Acknowledgments

We would like to thank the colleagues of the Digital Health Unit of the Joint Research Centre of the European Commission for guidance and support. The views expressed are purely those of the authors and may not in any circumstance be regarded as stating an official position of the European Commission. This article is based on work from COST Action CA24121 KGELL supported by COST (European Cooperation in Science and Technology).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.1, available under the GPT@EC service[13], in order to improve the readability and language of the manuscript. After using this service, the authors reviewed and edited the content as needed and assume full responsibility for the content of the published article.

References

[1] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, A. Mian, A comprehensive overview of large language models, ACM Trans. Intell. Syst. Technol. 16 (2025). doi:10.1145/3744746.

[2] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Trans. Inf. Syst. 43 (2025). doi:10.1145/3703155.

[3] R. Azamfirei, S. R. Kudchadkar, J. Fackler, Large language models and the perils of their hallucinations, Critical Care 27 (2023). doi:https://doi.org/10.1186/s13054-023-04393-x.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., US, 2020, pp. 9459–9474.

[5] L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, J. Seidel, Retrieval augmented generation for large language models in healthcare: A systematic review, PLOS Digital Health 4 (2025). doi:10.1371/journal.pdig.0000877.

[6] R. Yang, Y. Ning, E. Keppo, M. Liu, C. Hong, D. S. Bitterman, J. C. L. Ong, D. S. W. Ting, N. Liu, Retrieval-augmented generation for generative artificial intelligence in health care, npj Health Systems 2 (2025) 2. doi:10.1038/s44401-024-00004-1.

[7] Y. Gao et al., Retrieval-Augmented Generation for Large Language Models: A Survey, 2024. doi:10.48550/arXiv.2312.10997. arXiv:2312.10997.

[8] A. Hogan, E. Blomqvist, M. Cochez, C. D'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Computing Surveys 54 (2021). doi:10.1145/3447772.

[9] L. Liang, Z. Bo, Z. Gui, Z. Zhu, L. Zhong, P. Zhao, M. Sun, Z. Zhang, J. Zhou, W. Chen, W. Zhang, H. Chen, KAG: Boosting LLMs in Professional Domains via Knowledge Augmented Generation, in: WWW Companion 2025 - Companion Proceedings of the ACM Web Conference 2025, 2025, p. 334 – 343. doi:10.1145/3701716.3715240.

[10] X. Huang, H. He, S. Bao, F. Wang, H. Wu, H. Wang, PLATO-KAG: Unsupervised Knowledge-Grounded Conversation via Joint Modeling, in: NLP for Conversational AI, NLP4ConvAI 2021 - Proceedings of the 3rd Workshop, 2021, p. 143 – 154.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, et al., Transformers: State-of-the-Art Natural Language Processing, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.

[12] S. Tedeschi, V. Maiorca, N. Campolungo, F. Cecconi, R. Navigli, WikiNEuRal: Combined Neural and Knowledge-based Silver Data Creation for Multilingual NER, in: Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021, 2021, p. 2521 – 2533. doi:10.18653/v1/2021.findings-emnlp.215.

[13] GPT@EC (©European Union 2026, https://gpt.tech.ec.europa.eu/home) is an internal tool of the European Commission that aims to provide scientific and non-scientific staff with the means to explore securely the application of Generative AI technology to their daily work.

[13] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, Gliner: Generalist model for named entity recognition using bidirectional transformer, 2023. URL: https://arxiv.org/abs/2311.08526. arXiv:2311.08526.
[14] U. Zaratiana, G. Pasternak, O. Boyd, G. Hurn-Maloney, A. Lewis, Gliner2: An efficient multi-task information extraction system with schema-driven interface, 2025. URL: https://arxiv.org/abs/2507.18546. arXiv:2507.18546.
[15] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, M. A. Musen, Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications, Nucleic Acids Research 39 (2011) W541–W545.
[16] O. Erling, Virtuoso, a hybrid rdbms/graph column store, in: Proceedings of the Linked Data Benchmark Council, 2012.
[17] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, X. Lu, PubMedQA: A dataset for biomedical research question answering, EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference (2019) 2567 – 2577. doi:10.18653/v1/D19-1259.

A. SPARQL Queries

For readability, in the listings below `LBL` abbreviates the property path `skos:prefLabel|rdfs:label|skos:altLabel|skosxl:literalForm|obo:hasRelatedSynonym`, used throughout to obtain human-readable labels.

Listing 1: Concept lookup and ranking. One `FROM` clause is generated per selected ontology graph; the clauses are omitted when the search is not restricted.

```
PREFIX skosxl: <http://www.w3.org/2008/05/skos-xl#>
SELECT ?concept ?label (COUNT(?edge) AS ?score)
FROM <ONTOLOGY_1> ... FROM <ONTOLOGY_n>
WHERE {
  ?concept LBL ?label .
  FILTER (LCASE(STR(?label)) = "<surface form>")
  ?concept ?edge ?o .
}
GROUP BY ?concept ?label
ORDER BY DESC(?score)
```

Listing 2: Triple retrieval for a linked concept. The query is issued for the primary concept URI and, separately, for each alternative URI.

```
PREFIX skosxl: <http://www.w3.org/2008/05/skos-xl#>
SELECT DISTINCT ?labelS ?labelP ?labelO
WHERE {
  {                                 # (a) object triples
    <CONCEPT_URI> ?p ?o .
    <CONCEPT_URI> LBL ?labelS .
    ?p LBL ?labelP .
    ?o LBL ?labelO .
  } UNION {                         # (b) literal values
    SELECT ?labelS ?labelP ?labelO WHERE {
      <CONCEPT_URI> ?p ?labelO .
      <CONCEPT_URI> LBL ?labelS .
      ?p LBL ?labelP .
      FILTER (isLiteral(?labelO))
    }
  } UNION {                         # (c) OWL existential restrictions
    SELECT DISTINCT ?labelS ?labelP ?labelO WHERE {
      <CONCEPT_URI> ?ppp ?ooo .
      ?ooo rdf:type owl:Restriction .
```

```
      ?ooo owl:onProperty ?p .
      ?ooo owl:someValuesFrom ?o .
      <CONCEPT_URI> LBL ?labelS .
      ?p LBL ?labelP .
      ?o LBL ?labelO .
    }
  }
}
```