

From Fragmentation to Fidelity: Can Partially Harmonized Knowledge Graphs Support Reliable Clinical Question Answering?

Astha Anand¹, Martin Hofmann-Apitius¹, Maria Vehreschild², Jörg Janne Vehreschild³ and Marc Jacobs¹

¹Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Sankt Augustin, Germany

²University Hospital Frankfurt, Department of Internal Medicine, Infectious Diseases, Goethe University Frankfurt, Germany

³Goethe University Frankfurt, Faculty of Medicine, Institute for Digital Medicine and Clinical Data Sciences, Germany

Abstract

Biomedical and clinical knowledge remains fragmented across sources. Fully harmonized knowledge graphs (KGs) are rarely available, raising the question of whether partially integrated KGs can still support reliable clinical reasoning. We present a constrained KG-RAG prototype for clinical QA over a partially integrated comorbidity KG. The system performs deterministic, entity- and context-aware evidence selection over KG-linked clinical guidelines, trials, and mechanistic relations. Across diagnostic evaluation, clinical assessment, and neurodegenerative risk reasoning tasks, the agent produces provenance-linked guidance.

Keywords

Knowledge Graphs, Retrieval-Augmented Generation (RAG), Generative AI and Knowledge Graphs, Reasoning on RAG, Provenance, Explainable AI, Biomedicine and Health

1. Introduction

Clinical question answering (QA) often requires evidence from disconnected clinical guidelines, clinical trials, and biological knowledge. This is a semantic interoperability problem, not only a text-generation problem, because relevant evidence may exist across partially aligned knowledge layers [1]. Large Language Models (LLMs) are increasingly applied in medicine due to their high linguistic fluency; however, they present significant risks, including unsupported claims or hallucinated references [2]. Existing Retrieval-Augmented Generation (RAG) systems often focus on guidelines, trials, or graph-based reasoning separately, whereas clinical QA may require these evidence types to be considered together. We address these gaps through a multi-stage KG-RAG pipeline over a previously developed semi-harmonized comorbidity KG [1]. Here, we integrate clinical guidelines, clinical trials, and mechanistic biomedical pathways while imposing determinism at the retrieval layer rather than open-ended graph exploration.

2. Methods

2.1. Knowledge graph and bridge entities

The KG contains two semantically distinct but linked layers: (i) a clinical evidence subgraph containing guideline chunks, clinical trial metadata, and biomedical entity mentions, and (ii) a biomedical hypothesis subgraph containing disease, phenotype, gene, and biological-process relations. Interoperability between these layers is mediated through a limited set of shared "bridging" entities. At runtime, the system derives these bridge nodes by selecting entities that participate in both clinical evidence relations and non-clinical mechanistic relations. These bridge entities are the only admissible anchors for downstream retrieval. This prevents open-ended graph exploration and ensures that selected concepts are grounded in both clinical and mechanistic evidence spaces.

Posters, Demos, Blue Sky, and Tutorials at SEMANTiCS 2026, Sep 2026, Ghent, Belgium



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

2.2. KG-RAG pipeline

Figure 1 summarizes the complete KG-RAG workflow, from user question input to provenance-linked answer generation. In the first LLM stages (steps 1–5 of Figure 1), the question is decomposed into 4–8 clinical anchor terms representing disease context and explicitly mentioned phenotypes or symptoms. These anchors are then mapped exclusively to the shared bridge entities. If no node directly captures a queried concept, an additional proxy-selection step chooses the closest semantically useful entity. To ensure precision, up to three final entities are then selected by prioritizing explicit phenotypes over broader surrogates.

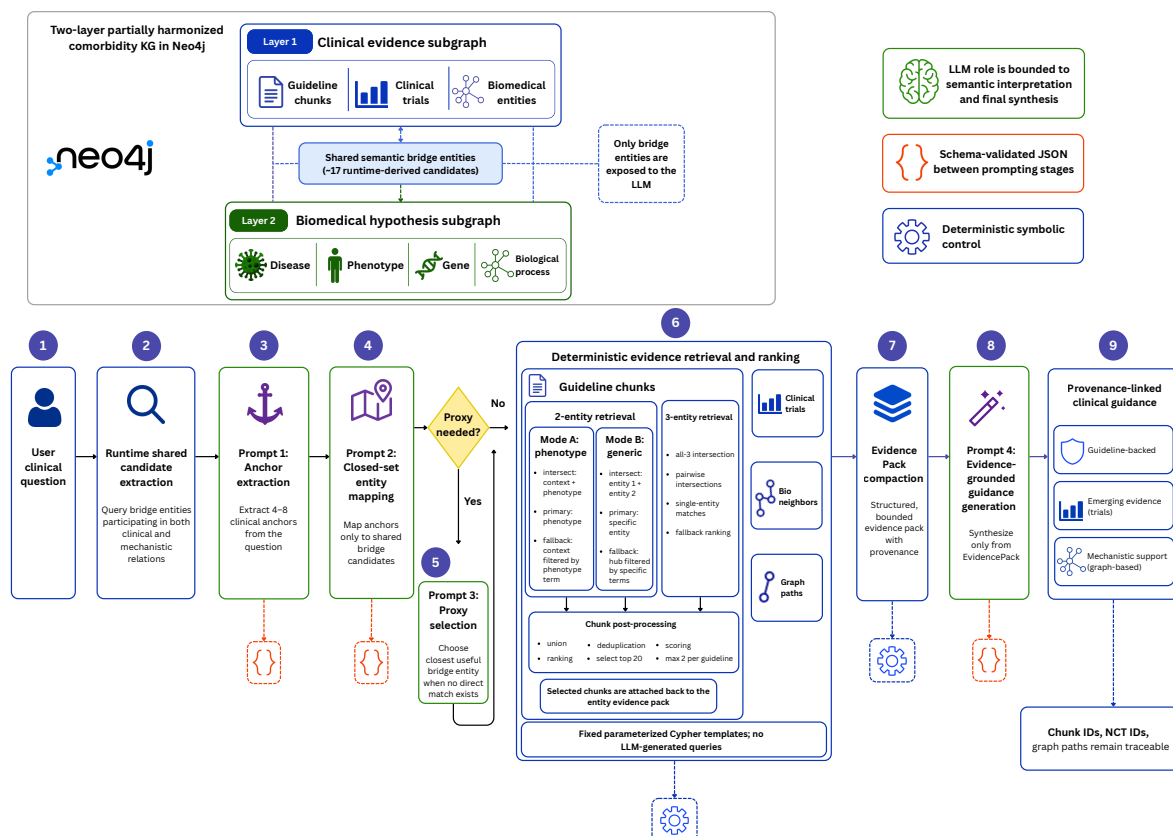


Figure 1: Overview of the clinical reasoning pipeline: User questions are first mapped to runtime-derived shared bridge candidates (or proxies when no direct match is available) from the KG (steps 1-5). Evidence retrieval (step 6) is then performed deterministically using Cypher templates to extract and rank relevant guideline chunks. Additionally trials, biomedical nodes, and graph paths are returned. The compacted EvidencePack (steps 7-9) is finally used for evidence-grounded guidance generation, producing provenance-linked outputs that distinguish guideline-backed statements, trial-supported evidence, and graph-based mechanistic support.

Steps 6–9 of Figure 1 cover deterministic evidence retrieval, EvidencePack compaction, and final evidence-grounded answer generation. For each selected entity, the system retrieves three evidence classes: linked clinical trial nodes, linked guideline chunks, and non-clinical biomedical neighbors. Guideline retrieval is ranked symbolically prioritizing chunks where multiple selected entities intersect. This reduces the inclusion of noisy chunks from infiltrating the evidence pool. Retrieved chunks are deduplicated, and ranked using weighted intersection scores. The top 20 chunks are selected with a maximum of 2 chunks per guideline to maintain diversity across guideline sources. In parallel, the system retrieves mechanistic support by tracing short non-clinical graph paths from the selected disease context through the chosen bridge entity toward downstream neurodegenerative endpoints. These paths are not used as clinical recommendations themselves, but serve as supplementary biological plausibility signals in the final reasoning stage.

All retrieved evidence is compacted into an EvidencePack containing selected entities, guideline chunks, NCT-linked trial metadata, biomedical neighbors, and graph-path summaries. Fixed upper bounds are imposed on the number of entities (max 3), chunks (max 10 per entity), trials (max 5 per entity), neighbors (max 10 per entity), and graph paths (max 5) to avoid context bloat. The final LLM prompt receives only this compact EvidencePack and is instructed to distinguish guideline-backed recommendations, trial-supported emerging evidence, unsupported areas, and graph-based mechanistic support. Final answers retain provenance through chunk IDs, trial identifiers, textual snippets verbatim, and graph traversals with ontology identifiers.

3. Results

We evaluated the proposed agent on three clinical question answering scenarios centered on persistent olfactory dysfunction after COVID-19. For each question, two responses were prepared for the domain experts: one generated by the agentic system and the other by a standalone LLM.

The use cases were:

1. **Q1:** How should persistent olfactory dysfunction be clinically evaluated in post-COVID patients?
2. **Q2:** What clinical assessments are appropriate for patients with prolonged olfactory dysfunction after COVID-19?
3. **Q3:** How should clinicians assess neurodegenerative risk in patients with persistent smell loss after COVID-19?

3.1. Evidence Decomposition for Use Case 1

For the first question “**How should persistent olfactory dysfunction be clinically evaluated in post-COVID patients?**”, the agent selected COVID-19 (DOID:0080600) as the disease-context entity and sense of smell impaired (MESH:D000857) as the phenotype link. The final response was organized into guideline-backed clinical evaluation, emerging trial evidence, and graph-supported extrapolation.

Guideline-backed evidence: The main clinical evaluation steps were supported by chunks which stated that assessment of an olfactory deficit should include patient history, duration of deficit, sinonasal symptoms, fluctuations, distortions, preceding events such as trauma or upper respiratory tract infection, objective measurement of olfactory function, and nasal endoscopy. The agent therefore generated patient history, olfactory function measurement, and nasal endoscopy as the core evaluation components. Additional chunks backed the use of structured smell testing, including odor stimuli presented to the subject with multiple descriptors. The final answer named examples including Sniffin Sticks, Connecticut Chemosensory Clinical Research Center Test, Barcelona Smell Test-24, and TT Olfactometer. Certain chunks stated that olfactory testing is useful for differential diagnosis, prediction of disease progression, and early detection of cognitive decline in neurodegenerative diseases.

Emerging trial evidence: The final response used NCT04696601, a trial investigating rapid screening for olfactory disorders in COVID-19 infection using an olfactory test/Olfascan compared with RT-PCR analysis.

Graph-based support: The graph signal linked COVID-19 to inflammatory response, SNCA, MAPT, and Alzheimer disease. This was used only as evidence-grounded extrapolation, suggesting that persistent post-COVID olfactory symptoms may warrant consideration of broader neurological assessment, rather than as a direct clinical recommendation.

3.2. Evidence Decomposition for Use Case 2

For the question “**What clinical assessments are appropriate for patients with prolonged olfactory dysfunction after COVID-19?**”, the agent extracted anchors such as *after COVID-19* and *prolonged olfactory dysfunction*, which were mapped to the bridge entities COVID-19 (DOID:0080600) and sense of smell impaired (MESH:D000857).

Guideline-backed evidence: The final answer used guideline chunks to support patient history, duration of olfactory deficit, concomitant sinonasal symptoms, fluctuation, distortions, preceding events, objective olfactory function measurement, and nasal endoscopy. These chunks supported culturally adjusted smell testing and named olfactory test modalities, including Sniffin Sticks, Connecticut Chemosensory Clinical Research Center Test, Barcelona Smell Test-24, and TT Olfactometer. Additionally, the relevant chunks mentioned the role of olfactory testing in differential diagnosis, prediction of disease progression, and early detection of cognitive decline.

Emerging trial evidence: The final answer incorporated trial-supported assessment signals from NCT06589375, NCT04361474, and NCT05364125. These trials supported olfactory pathway assessment using Sniffin Sticks and EEG-based measures, ODORATEST-based improvement in COVID-19-related hyposmia, and olfactory training outcomes including threshold, identification, quality of life, and taste/smell changes.

Graph-based support: The retrieved graph paths linked COVID-19 to inflammatory response or IL6, followed by SNCA, sense of smell impaired, MAPT, and Alzheimer disease. The final answer used these paths as mechanistic context to support cautious neurodegenerative-risk interpretation, while keeping graph evidence separate from formal clinical guidance.

3.3. Evidence Decomposition for Use Case 3

For “**How should clinicians assess neurodegenerative risk in patients with persistent smell loss after COVID-19?**”, the agent selected three entities: COVID-19 (DOID:0080600), sense of smell impaired (MESH:D000857), and Neurodegenerative Diseases (MESH:D019636).

Guideline-backed evidence: The final answer used relevant chunks to support neurodegenerative-risk assessment, stating that olfactory testing is useful for differential diagnosis, prediction of disease progression, and early detection of cognitive decline in neurodegenerative diseases. The answer also supported baseline olfactory evaluation, including patient history, objective olfactory function measurement, and nasal endoscopy.

Trial evidence: No clinical trial evidence was used in the final Q3 response.

Graph-based support: The retrieved graph paths connected COVID-19 to inflammatory response or IL6, then to SNCA, sense of smell impaired, MAPT, and Alzheimer disease. The final response explicitly labeled these as graph-based mechanistic support. The graph signal was used to support biological plausibility for considering further neurological evaluation in post-COVID patients with persistent smell loss, but it was not presented as a standalone clinical recommendation.

3.4. Expert Scores

Table 1

Question 1 expert scoring table. Scores are reported for Expert 1 (E1) and Expert 2 (E2).

Dimension	Response A		Response B	
	E1	E2	E1	E2
Guideline Adherence	4	4	5	4
Evidence Traceability	5	5	1	2
Evidence Reliability	5	3	1	1
Clinical Plausibility and Safety	3	3	5	4
Question-Intent Alignment	3	3	5	4

Table 2

Question 2 expert scoring table. Scores are reported for Expert 1 (E1) and Expert 2 (E2).

Dimension	Response A		Response B	
	E1	E2	E1	E2
Guideline Adherence	5	4	5	4
Evidence Traceability	5	5	1	2
Evidence Reliability	4	3	1	1
Clinical Plausibility and Safety	3	3	4	4
Question-Intent Alignment	3	3	4	4

Table 3

Question 3 expert scoring table. Scores are reported for Expert 1 (E1) and Expert 2 (E2).

Dimension	Response A		Response B	
	E1	E2	E1	E2
Guideline Adherence	4	4	5	4
Evidence Traceability	5	5	1	2
Evidence Reliability	5	3	1	1
Clinical Plausibility and Safety	3	3	5	4
Question-Intent Alignment	3	3	5	4

4. Conclusion

The expert scores show a consistent trade-off. The agent received substantially higher scores for evidence traceability and evidence reliability. The standalone LLM produced more clinically complete or directly framed answers; however, it did not expose the sources used to support its claims. The agent improves evidence fidelity, and serves as a reusable design for partially harmonized cross-domain infrastructures. This design also reduces the black-box character of end-to-end LLM systems because evidence discovery is not delegated to the model.

5. Limitations

The evaluation was restricted to olfaction-centred use cases because olfactory dysfunction is the most consistently represented bridge phenotype in the current KG. This enables controlled testing but does not yet establish generalizability across other post-COVID neurological or psychiatric manifestations. Other symptom domains, such as depression, anxiety, sleep disturbances or speech-related symptoms, currently rely on weaker proxy mappings. The system is therefore inherently limited by the completeness and connectivity of the utilized KG. Broader evaluation will therefore require additional symptom domains with stronger cross-layer connectivity.

6. Outlook

The lower scores for clinical completeness and question-specific alignment reflect current KG coverage limits. Future work will expand neurological concepts, guideline coverage, entity connectivity, and cross-layer evidence density. This framework is intended to be integrated into the EU-funded COMMUTE [3] project's EvidenceBase. Understanding how COVID-19 interacts with long-term neurological conditions is not a single-data problem and we cannot make knowledge talk to each other easily. The EvidenceBase is a multi-agent reasoning system connecting this guidance agent with other agents over multi-layered networks, patient cohorts, and experimental data.

Funding

This work was supported by the Fraunhofer Institute for Algorithms and Scientific Computing (SCAI). Funding came from the COMMUTE project, which is funded by the European Union under grant agreement number 101136957.

Code and Data Availability

The source code, along with documentation, is available at https://github.com/SCAI-BIO/Guidelines_Store_Agent.

Ethics, Regulatory Scope, and Environmental Footprint

This work is a research prototype, not intended for patient care, diagnosis, treatment, regulatory approval, or clinical deployment. All outputs require expert review. The environmental footprint was limited to prototype-level API-based LLM inference and local knowledge graph queries; no large-scale training was performed, and carbon emissions were not measured.

Declaration on Generative AI

The authors used generative AI and language support tools, including ChatGPT, Gemini, and Grammarly, for grammar, spelling, paraphrasing, rewording, and readability improvements. All content was reviewed, edited, and verified by the authors, who take full responsibility for the final publication.

References

- [1] E. Gupta, A. Anand, M. Hofmann-Apitius, M. Jacobs, Integrating clinical evidence for multi-condition care, *Artificial Intelligence in the Life Sciences* 9 (2026) 100164. URL: <https://www.sciencedirect.com/science/article/pii/S2667318526000127>. doi:<https://doi.org/10.1016/j.aailsci.2026.100164>.
- [2] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, D. S. W. Ting, Large language models in medicine, *Nature Medicine* 29 (2023) 1930–1940. URL: <https://doi.org/10.1038/s41591-023-02448-8>. doi:10.1038/s41591-023-02448-8.
- [3] Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Commute project, <https://www.commute-project.eu/en/about.html>, 2022. Accessed: 2026-07-01.