ADVANCES IN KNOWLEDGE ORGANIZATION
Proceedings of International ISKO Conference - ISSN: 0938-5495 - 2026 - São Paulo

How retrieval structure shapes LLM outputs: an assessment framework for GraphRAG

Rafael Rocha

Federal University of Minas Gerais, Brazil

Gercina Ângela de Lima

Federal University of Minas Gerais, Brazil

Webert Júnio Araújo

Federal University of Science and Innovation of Minas Gerais, Brazil

Abstract

Large Language Models have reshaped contemporary information access; however, their generative flexibility can produce epistemic instability when outputs are insufficiently constrained by structured knowledge. This study develops an analytical framework grounded in Knowledge Organization principles to examine how semantic anchoring and inferential latitude jointly shape response behavior in GraphRAG systems. Methodologically, the research adopts an exploratory design based on controlled variation of decoding parameters and structured queries grounded in Wikidata. The findings indicate that model behavior follows systematic patterns rather than random variation, clustering into distinct regimes associated with different configurations of anchoring and inferential latitude. To synthesize these observations, the paper proposes the Epistemic Control Matrix for Knowledge-Structured Generative Systems, a bidimensional conceptual model articulating four regimes: semantic conformity, interpretive coherence, rigid structural constraint, and semantic drift. The framework contributes to Knowledge Organization by providing a conceptual basis for understanding epistemic regulation in knowledge-augmented generative environments.

Keywords: GraphRAG; Knowledge Graphs; Knowledge Organization

1.0 Introduction

Large Language Models (LLMs) have reshaped contemporary information access by enabling conversational interaction that transforms retrieval into a dynamic and iterative process (Hassani and Silva 2023; Peng et al. 2025). Rather than retrieving documents through explicit matching mechanisms, LLM-based systems generate integrated responses that synthesize information in natural language. This shift reconfigures the epistemic conditions of information mediation and raises fundamental questions about how knowledge is structured, constrained, and validated within generative environments.

As interactions become semantically complex, LLMs may generate plausible, yet factually unsupported outputs, commonly referred to as hallucinations (Ji et al. 2023). Because generation is probabilistic rather than truth-conditional, reliability cannot be inferred solely from linguistic coherence. Ensuring epistemic stability, therefore, requires mechanisms capable of constraining generative inference through structured forms of semantic grounding.

Among the strategies proposed to mitigate this limitation, Retrieval-Augmented Generation (RAG) anchors outputs in external knowledge sources, thereby reducing semantic drift and enhancing factual reliability (Arslan et al. 2024; Park 2025). When retrieval is performed over structured representations such as Knowledge Graphs (KGs), the approach is commonly designated as GraphRAG (Ibrahim et al. 2024). In such settings, semantic structure operates not merely as contextual support, but as an organizing substrate that defines boundaries of permissible inference during generation.

Despite the growing adoption of GraphRAG architectures, limited attention has been given to systematically examining how variations in retrieval configuration and inference parameters shape epistemic outcomes. In particular, the interaction between semantic anchoring provided by structured knowledge and the degree of inferential latitude permitted during decoding remains insufficiently theorized within Knowledge Organization (KO). Framing GraphRAG as a structured epistemic environment enables its analysis within the conceptual tradition of KO, where semantic control, warrant and interpretive constraint are central concerns.

Accordingly, this study addresses the following question: What differences emerge in LLM responses under distinct configurations of retrieved inputs and inference parameters within a GraphRAG setting? Grounded in KO principles, the paper proposes an analytical framework to examine how semantic anchoring and inferential latitude jointly structure epistemic behavior in knowledge-augmented generative systems.

2.0 Background

Knowledge Organization Systems (KOS), including classifications, thesauri, taxonomies, ontologies and KGs, function as semantic instruments that structure, stabilize and mediate knowledge (Gnoli 2020; Park 2025). By establishing conceptual hierarchies, relational links and interpretive constraints, KOS do more than improve retrieval efficiency; they embed epistemic commitments regarding what qualifies as a concept, how entities are related and under which contextual conditions information becomes meaningful. Within the tradition of KO, retrieval has, therefore, never been understood as a purely technical operation, but as an activity grounded in prior processes of conceptual treatment and semantic organization.

The emergence of LLM-based systems introduces a significant shift in this landscape. Unlike traditional retrieval systems that rely on document indexing and explicit query–document matching, generative interfaces synthesize responses through probabilistic language modeling and contextual integration. This transformation challenges established

assumptions about how knowledge is organized, validated and mediated. From a KO perspective, it raises a central question: How can semantic control and interpretive constraint be maintained in environments where responses are generated rather than retrieved?

RAG represents one attempt to address this tension by reintroducing structured knowledge into generative inference. By grounding outputs in curated sources, RAG partially reproduces functions historically associated with KOS, including authority control, semantic warrant and contextual delimitation. When retrieval is performed over KGs, the resulting configuration — GraphRAG — can be understood as an application of classical KOS principles within generative Artificial Intelligence (AI) environments.

KGs encode entities and relations in structured triples, enabling explicit semantic connections and traceable inferential paths (Yasunaga et al. 2021). Their relational architecture supports precision, contextual coherence and structured navigation across entities. When integrated into generative systems, this relational architecture functions as an epistemic regulator that delimits inferential expansion and reduces ambiguity. However, the epistemic effects of GraphRAG depend not only on the presence of structured knowledge, but also on how it is retrieved, formatted and incorporated during inference.

In contrast to text-based RAG systems, where retrieved passages provide contextual grounding without explicit relational constraints, GraphRAG introduces structured semantic dependencies that delimit inference paths. This relational explicitness enhances traceability and reduces ambiguity, allowing epistemic control to be exerted not only through content selection but through the organization of knowledge itself.

Despite increasing adoption of GraphRAG architecture, the interaction between semantic anchoring and generative flexibility remains insufficiently examined within KO. The following section develops the theoretical and methodological foundations necessary to analyze GraphRAG as a structured epistemic environment in which semantic constraint and inferential latitude jointly shape response behavior.

3.0 Methodology

This study reports an applied exploratory investigation conducted in January 2026 using the GPT-4.1 model, accessed via the ChatGPT interface. In line with established principles in LIS, exploratory research aims to increase familiarity with a problem domain, clarify its conceptual boundaries and support the formulation of initial hypotheses, particularly in contexts characterized by emerging or insufficiently theorized phenomena (Gil 2022). This methodological stance is particularly appropriate for examining LLMs, whose modes of information access and mediation challenge conventional distinctions between retrieval systems and generative technologies.

The study examines the application of a simplified RAG configuration within the model architecture. Positioned at the intersection of LIS, KO and Interactive Information Retrieval (IIR), the exploratory design enables a targeted examination of how LLM-based tools

mediate information when supported, explicitly or implicitly, by organized knowledge structures. Rather than prioritizing performance optimization or system comparison, the study emphasizes conceptual clarity and analytical insight, producing empirical observations to inform subsequent, more systematic investigations of RAG architectures, semantic control mechanisms and KO principles in LLM-driven information systems.

3.1 Research design

The empirical corpus was constructed through a structured query executed against Wikidata (n.d.). The query retrieved relational data concerning three film directors: Steven Spielberg, Christopher Nolan and George Lucas, selected for the structural richness of their filmographies and associated role-based relations. The objective was not statistical representativeness, but semantic density, enabling controlled observation of grounding effects within a relationally complex domain.

The query returned structured tuples linking a film professional, a role (e.g., director, producer, screenwriter or cast member), a corresponding work and the earliest year of participation. This modeling generates a subject–predicate–object structure typical of KGs and other KOS. From a KO perspective, such structuring preserves relational semantics, supports role disambiguation and enables traceable inferential navigation across entities and works. The dataset, thus, functions as a controlled semantic substrate for examining grounding effects in generative inference.

3.2 Experimental Design and Parameter Control

GraphRAG-assisted prompting was operationalized as a structured information-seeking episode in which a predefined question is mediated through a knowledge-graph retrieval context and resolved under specified decoding parameters. Each run integrates an articulated information need, a structured semantic substrate and a controlled generative configuration, allowing systematic comparison of epistemic variation across inference settings.

All interactions were conducted through the DEIGMATA tool (Rocha and Lima 2025), which enabled parameter manipulation and logging under controlled conditions.[1] The benchmark focused on a single LLM (GPT-4.1) to eliminate architectural variability and the maximum response length was fixed at 500 tokens.

Temperature and top-p were systematically varied to regulate inferential latitude (OpenAI n.d.). Table 1 presents their analytical roles within a KO-informed framework, while Table 2 classifies parameter ranges according to behavioral characteristics and expected epistemic effects. Together, these matrices establish the operational structure for examining how decoding freedom interacts with semantic anchoring.

1The dataset used in this study is available as a publicly accessible CSV file at https://gist.github.com/rafaro/6d09f0e29cd531f19fae080ebf160d77

Table 1 Outcome analysis matrix

Parameter	KO role	Expected influence on outcomes
Temperature	Inferential freedom	Higher values increase noise and reduce inferential control
Top-p	Probability boundary	Lower values constrain vocabulary and inference paths
GraphRAG context presence	Semantic warrant	Increases grounding and appropriate refusal
GraphRAG structure (triples and paths)	Specificity & coherence	Structured forms improve precision and traceability

Table 2 Temperature and Top-p classification matrix

Parameter	Range	Classification	Behavioral characteristics	Typical use cases
Temperature	0.0–0.5	Low	Highly deterministic; near-identical outputs across runs; minimal lexical variation	Benchmarking, fact-checking, GraphRAG grounding and reproducibility
Temperature	0.6–1.2	Medium	Controlled variation in phrasing; semantic meaning remains stable	Explanatory answers, summaries and structured analysis
Temperature	1.2–2.0	High	High variability; creative phrasing; increased risk of hallucination	Ideation, brainstorming and creative writing
Top-p	0.1–0.3	Low	Extremely constrained token pool; repetitive but precise	Controlled experiments and ontology-driven GraphRAG

Parameter	Range	Classification	Behavioral characteristics	Typical use cases
Top-p	0.4–0.7	Medium	Balanced diversity with semantic stability	QA systems, text-based RAG and grounded generation
Top-p	0.9–1.0	High	Very broad token pool; inclusion of long-tail tokens	Creative generation, narrative and exploratory tasks

A predefined set of questions (Table 3), representing distinct logical and cognitive structures, was submitted under two contrasting system contexts: (i) a default configuration emphasizing factual restraint and (ii) a minimal configuration permitting greater inferential flexibility.

This design enables comparative analysis of how grounding intensity and decoding freedom jointly shape epistemic outcomes.

Table 3 Question category matrix

Category	Question
Counterfactual	If Titanic were a documentary, who could narrate it?
Negated premise	If George Lucas was not the director of Star Wars, then what was he?
False-premise	Titanic is not a film, so what is it?
Missing-attribute	How well does James Cameron play the harp?
Completeness	Who was involved in Jurassic Park, including the director, producer, cast etc.?
Interpretive	What is the atmosphere of Jurassic Park?
Conceptual classification	Is Avatar a film or a franchise?
Entity identification	Steven Spielberg

Each response was manually coded into three categories: (i) faithfulness, strict KG grounding or justified refusal; (ii) coherence, semantically stable output incorporating bounded interpretation; (iii) hallucination, logically inconsistent or ungrounded content.

This protocol supports systematic observation of how semantic anchoring and inferential latitude interact. The controlled manipulation of these dimensions allows identification of recurrent behavioral patterns that inform the subsequent conceptual formalization of epistemic regimes.

4.0 Results and discussion

The findings reveal a systematic relationship between decoding flexibility, semantic grounding and output reliability. All experimental runs were conducted under controlled conditions using the DEIGMATA tool, ensuring consistent parameter manipulation and reproducible comparison across identical question sets.

From a Knowledge Organization perspective, the observed effects cannot be reduced to parameter variation alone. Rather, they reflect the interaction between generative mechanisms and pre-structured semantic constraints embedded in the knowledge graph. In this sense, GraphRAG operationalizes classical KO functions, such as semantic warrant, relational structuring, and interpretive delimitation, within a generative environment. The variability observed across configurations therefore expresses different regimes of epistemic control, not merely stochastic fluctuation. These regimes correspond to distinct configurations of semantic warrant, constraint, and permissible inference, which are central concerns in Knowledge Organization theory.

Under the most permissive configuration (temperature = 2.00; top-p = 1.00), the model produced severely degraded outputs characterized by syntactic instability and semantic breakdown. These responses were consistently classified as hallucinations, as they lacked interpretability and alignment with the knowledge graph.

Intermediate parameters (temperature = 1.00; top-p = 0.50) generated linguistically coherent responses that frequently extended beyond explicit KG assertions. For example, when asked who might narrate Titanic as a hypothetical documentary, the model identified James Cameron by combining a KG-supported fact with inferential reasoning. Such outputs were classified as coherence, reflecting bounded interpretive extension rather than strict entailment.

Under restrictive settings (temperature = 0.10; top-p = 0.10), the model demonstrated high epistemic restraint, declining to speculate when information was absent from the KG. These responses maintained close adherence to graph evidence and were classified as faithfulness.

The default system context configuration imposed a stringent reasoning policy that compelled the model to acknowledge informational limitations explicitly. Under these conditions, responses favored conservative formulations, increasing faithfulness while

occasionally reducing perceived completeness. In contrast, the minimal system context permitted bounded extrapolation beyond graph boundaries, enhancing narrative coherence but increasing susceptibility to speculative deviation.

Collectively, these findings indicate that epistemic behavior in GraphRAG systems varies systematically according to the interaction between decoding parameters and grounding intensity. Figure 1 presents the distribution of evaluative categories across parameter configurations under default and minimal system contexts.

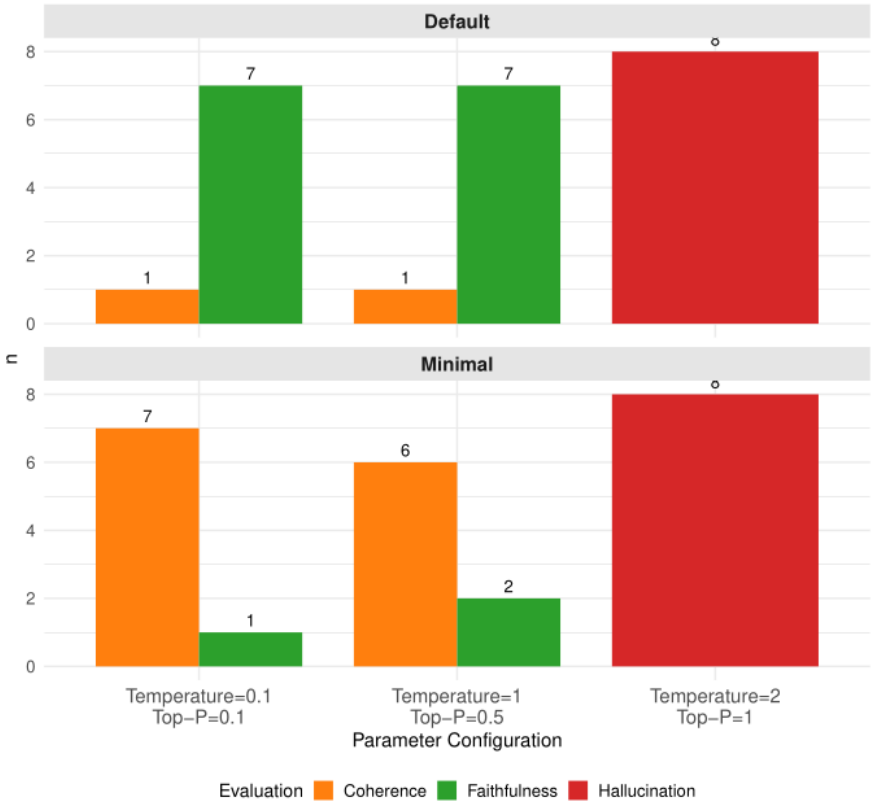

Figure 1. Distribution of faithfulness, coherence and hallucination across decoding configurations

From a KO perspective, the distribution illustrated in Figure 1 suggests that epistemic outcomes are not random artifacts of model behavior, but structured effects emerging from identifiable control variables. Specifically, two organizing dimensions appear to regulate response behavior: the degree of semantic anchoring provided by the KG and the degree of inferential latitude permitted by decoding flexibility.

While empirically related, semantic anchoring and inferential latitude operate as conceptually orthogonal and analytically distinct dimensions. Semantic anchoring refers to the degree to which responses are constrained by explicit relational structures in the knowledge graph, whereas inferential latitude concerns the range of permissible expansions during decoding. Both dimensions jointly shape output behavior, but they are not reducible to a single continuum: anchoring defines epistemic boundaries, whereas latitude governs movement within or beyond those boundaries.

The empirical categories of faithfulness, coherence and hallucination therefore correspond to distinct configurations of these two dimensions. To formalize this interaction, we propose the Epistemic Control Matrix for Knowledge-Structured Generative Systems (Figure 2), a bidimensional regime model that synthesizes the observed regularities into a conceptual framework grounded in KO principles.

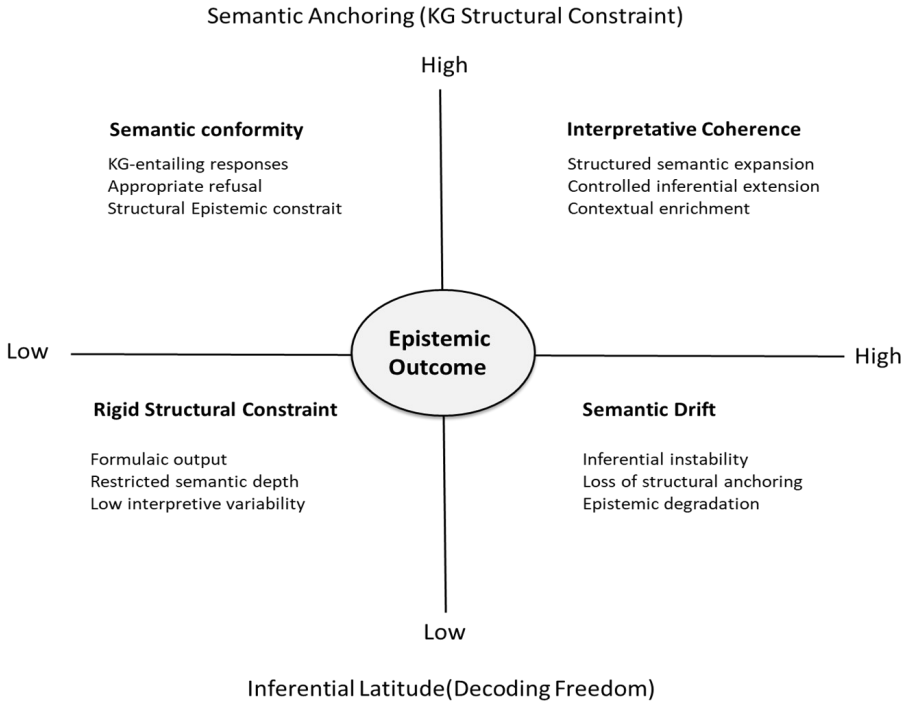

Figure 2. Epistemic Control Matrix for Knowledge-Structured Generative Systems

As depicted in Figure 2, epistemic outcomes emerge from the interaction between semantic anchoring (structural constraint imposed by the knowledge graph) and inferential latitude (decoding freedom). This interaction yields four regimes of epistemic control: semantic conformity, interpretive coherence, rigid structural constraint, and semantic drift. The matrix thus provides a conceptual abstraction of the experimental findings, positioning GraphRAG as a structured environment in which generative behavior is shaped by calibrated degrees of semantic constraint and inferential flexibility.

5.0 Conclusion

This study examined KGs as structured semantic foundations for KO within GraphRAG-based interactions involving LLMs. By conceptualizing KGs as grounding mechanisms in retrieval-augmented generation, the findings demonstrate that semantic structure plays a decisive role in regulating generative inference and mitigating epistemic instability. In contrast to unstructured retrieval, the relational architecture of KGs provides semantic control that constrains ambiguity and reduces the likelihood of hallucinated content.

The results indicate that epistemic outcomes in GraphRAG settings emerge not solely from the presence of structured knowledge, but from the interaction between semantic anchoring and inferential latitude. Permissive decoding configurations increase instability and semantic drift; restrictive regimes enhance faithfulness but limit expressive richness; intermediate settings enable bounded interpretive coherence. System-level prompting introduces an additional layer of regulation, influencing how responses are articulated while KGs delimit what is epistemically permissible.

These interactions are formally synthesized in the proposed Epistemic Control Matrix, which conceptualizes generative behavior as structured along two dimensions: grounding intensity and decoding freedom. By articulating four regimes of epistemic control — semantic conformity, interpretive coherence, rigid structural constraint and semantic drift — the model offers a conceptual framework for understanding how structured knowledge and generative flexibility jointly shape knowledge production in GraphRAG environments.

Overall, the study proposes an assessment framework that clarifies how variations in retrieval configuration and inference parameters affect LLM response behavior across the dimensions of faithfulness, coherence and hallucination. Effective GraphRAG deployment, thus, requires calibrated alignment between grounding mechanisms and generative flexibility to meet specific information-seeking objectives.

As a limitation, the investigation was conducted within a bounded domain and parameter set. It should be noted that the operationalization of inferential latitude through parameters such as temperature reflects current model configurations and may not be directly transferable to future LLM architectures. As some recent models abstract or remove explicit control over such parameters, replicability of the experimental setup may require alternative mechanisms for regulating generative variability. Future research should extend the framework to additional knowledge domains, more complex graph structures, and user-centered evaluation contexts. In doing so, this work positions GraphRAG within KO not merely as a technical enhancement of retrieval, but as a structured regime of epistemic governance in generative information systems.

Acknowledgment: The second author acknowledges the support of the National Council for Scientific and Technological Development (CNPq, Brazil) for the award of a Research Productivity Grant (PQ-B, Process No. 306713/2025-0).

AI Usage Statement: AI was used solely for grammatical revision. Guardrails were applied to prevent the introduction of any ideas or content not already present in the original text.

References

Arslan, Muhammad, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. "A Survey on RAG with LLMs." Procedia Computer Science 246: 3781–3790. https://doi.org/10.1016/j.procs.2024.09.178

Gil, Antonio Carlos. 2022. Como Elaborar Projetos de Pesquisa. 7th ed. Barueri: Atlas.

Gnoli, Claudio. 2020. Introduction to Knowledge Organization. London: Facet Publishing.

Hassani, Hossein, and Emmanuel Sirmal Silva. 2023. "The Role of ChatGPT in Data Science: How AI-Assisted Conversational Interfaces Are Revolutionizing the Field." Big Data and Cognitive Computing 7, no. 2: 62. https://doi.org/10.3390/bdcc7020062

Ibrahim, Nourhan, Samar Aboulela, Ahmed Ibrahim, and Rasha Kashef. 2024. "A Survey on Augmenting Knowledge Graphs (KGs) with Large Language Models (LLMs): Models, Evaluation Metrics, Benchmarks, and Challenges." Discover Artificial Intelligence 4: 76. https://doi.org/10.1007/s44163-024-00175-8

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. "Survey of Hallucination in Natural Language Generation." ACM Computing Surveys 55, no. 12: Article 248. https://doi.org/10.1145/3571730

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." In 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings, edited by Yoshua Bengio and Yann LeCun. http://arxiv.org/abs/1301.3781

OpenAI. n.d. "OpenAI API Overview." Accessed January 25, 2026. https://platform.openai.com/docs/overview

Park, Ziyoung. 2025. "AI for KOS Discovery: Refining Search, Recommendation, and Hallucination Mitigation." Knowledge Organization 52, no. 4: 45889. https://doi.org/10.31083/KO45889

Peng, Wei, Jingbo Meng, Lu Tang, Wenxue Zou, and Barikisu Issaka. 2025. "How Do Lay Users Seek Information with ChatGPT: An In-Situ Interview Study." International Journal of Human–Computer Interaction. Published online November 19, 2025. https://doi.org/10.1080/10447318.2025.2586086

Rocha, Rafael, and Gercina Ângela de Lima. 2025. "Deigmata: A User-Centered Visualization Tool for Knowledge Graph Exploration." Knowledge Organization 52, no. 8: 46137. https://doi.org/10.31083/KO46137

Wikidata. n.d. "Wikidata SPARQL Query Service." Accessed January 2, 2026. https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service

Yasunaga, Michihiro, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. "QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering." In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, edited by

Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, 535–546. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.45