AgentRAG: A Multi-Agent System Combining LLMs with Knowledge Graphs for Trustworthy Auditing

Sara Buchmann[a,*]

[a]Department of Computer Science and Mathematics, OTH Regensburg
ORCiD ID: Sara Buchmann https://orcid.org/https://orcid.org/0009-0009-5729-7902

Abstract. Integrating Large Language Models (LLMs) with graph-based retrieval-augmented generation (GraphRAG) in modular multi-agent AI systems promises to automate complex data-analytic processes. In highly regulated domains such as financial auditing, the AI-assisted automation of routine tasks like question answering must meet strict requirements for accuracy, explainability, and compliance, i.e., inaccurate or intransparent AI outputs risk undermining trust and cause severe financial and reputational consequences. We propose *AgentRAG*, a hybrid multi-agent architecture that couples LLMs with structured semantic knowledge to build a conversational analytical chatbot that lets auditors interact with their data. A three-step process involves retrieving data and generating an LLM response, verifying the LLM's answer against background knowledge, and explaining the validation and retrieval process. Leveraging GraphRAG, the system improves contextual grounding, traceability, and answer consistency, enabling auditors to validate and interpret AI-driven insights effectively. This approach enhanced trustworthiness and reliability in auditing workflows, facilitating the responsible adoption of AI in high-stakes environments.

1 Introduction and Research Questions

"Chatting with data" is an emerging paradigm in interactive data analytics, letting users pose natural-language questions to derive insights. This interaction can be realized by a modular multi-agent system, in which specialized agents collaborate to translate user queries into formal representations, to compute analytic results and to generate domain-specific, data-driven responses. A key challenge lies in the verbalization of analytical outcomes, transforming data analytics results into data stories grounded in rich domain knowledge. LLMs are attractive for this narrative role due to their context-aware generation abilities. According to the EU AI Act, domains such as financial auditing [7, 21] or healthcare [30, 22] are potentially categorized as high risk. Consequently, LLMs face limitations: their likelihood to hallucinate facts [12, 10] and the lack of transparency in their decision-making processes undermine trust in their outputs. Therefore, LLMs deployed in these domains must comply with strict reproducibility, traceability, and explainability requirements, and also broader trustworthyness criteria such as robustness and privacy [33, 14]. Verbalizing data-analytic results or knowledge-base queries [28] must be grounded in structured, semantically rich representations. *AgentRAG* addresses this by combining LLMs with factual and domain knowledge graphs (KGs) [1, 26] to ensure consistency, compliance, and thus correctness of the generated answers. The latter capture not only factual information but also the underlying semantic structures, enabling agents to interpret entity relationships with greater contextual awareness [18, 19, 17]. Within human-in-the-loop systems, explainability is just as essential as consistency: the mulit-agent system [8] must clearly show how it reached a conclusion, which data were involved, and why a particular result is consistent, valid and thus can be considered as correct. This leads to the following research question:

> *How can answers generated by a multi-agent system combining LLMs and knowledge graphs for auditing be transparently traced, validated, and explained?*

We decompose this into three sub-questions:

RQ1: How can the consistency of generated answers in a multi-agent system be systematically verified against the underlying knowledge?

RQ2: How can an agent-based validation ensure that generated answers are correct and comply with formal, domain-specific rules and constraints?

RQ3: Which methods enable agents to explain answer generation and validation results in a traceable, transparent, and auditable manner that ensures reproducibility for users?

Before outlining our approach, we briefly review related work.

2 Background

Combining LLMs and KGs has become a prominent research direction for improving the factual reliability of generated answers and the interpretability of the reasoning process [5, 22, 26]. A key motivation behind this integration is the use of structured knowledge to reduce hallucinations, enhance traceability, and provide context-aware reasoning. Recent work shows that graph-based retrieval (GraphRAG) [9] can improve both the factual correctness and the comprehensibility of LLM answers. Relevant subgraphs are integrated into the input context in order to make reasoning processes semantically sound and interpretable [16, 15]. Beyond retrieval, there is growing interest in verifying whether generated answers comply with formal constraints or domain-specific rules. Methods in this area combine LLMs with validation agents and symbolic or rule-based reasoning methods to enforce consistency with ontological structures or logical constraints [5, 13, 26]. Furthermore, the validation of KGs can be made more reliable and traceable by fusing LLMs with human feedback [27], as well as by using interactive explanation methods that provide counterexamples and support transparent reasoning [25]. Structured

* Email: sara.buchmann@oth-regensburg.de.

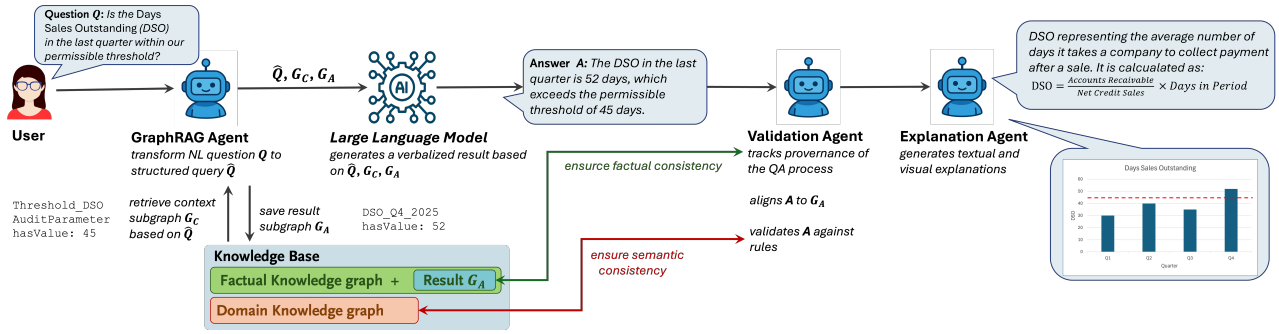

Figure 1: AgentRAG: Mulit-agent system combining LLMs with KGs for trustworthy auditing.

background knowledge improves explainability by grounding outputs in verifiable facts and yielding interpretable reasoning chains when KGs are injected into generation [3, 31]. Hybrid schemes that pair explainable clustering with generative models mitigate interpretability gaps in complex domains [2], and multi-level semantic annotation frameworks clarify what is captured and validated [20], enhancing auditability. Nevertheless, a unified architecture offering traceability, validation and explainability remains an open challenge, as most KG-augmented approaches enhance answer generation with KG-based reasoning, yet lack systematic validation and consistency checks.

3 Proposed Conceptual Approach

We introduce *AgentRAG*, a multi-agent system built on GraphRAG, combining retrieval, validation, and explanation agents to ensure verifiable and compliant outputs for question answering (QA) in auditing. Figure 1 illustrates the system's architecture and the interactions between agents. The knowledge base consists of two interrelated graphs that enable structured access to information:

- **Factual Knowledge Graph**: Stores instance-level data from source systems in a retrieval graph maintaining auditable memory, where each QA run appends its analytic fragment, preserving traceability and cumulative knowledge.
- **Domain Knowledge Graph**: Captures domain knowledge-taxonomies, semantic relations, and audit rules from audit standards such as IDW [6] and ISA [11] to enable semantic interpretation, regulatory reasoning, and compliance checks.

The multi-agent system can be conceptually divided into three phases: the QA process, the subsequent validation and the explanation process. The first phase is operationalized through the interaction of the *User*, *GraphRAG Agent*, and *LLM* components. The user poses a question Q in natural language, which is processed by the *GraphRAG* Agent transforming it into a structured form $\hat{Q}$ for querying the underlying knowledge graphs. The agent retrieves a relevant context subgraph G_C and derives an analytic subgraph G_A, which it appends to the factual KG. The LLM then verbalizes an answer A from query $\hat{Q}$, context G_C, and result G_A. Two specialised agents subsequently carry out validation and explanation.

Validation Agent First, a *backward-reasoning* step checks semantic consistency between the verbal answer A and the analytic subgraph G_A. A hypothetical subgraph $\tilde{G}_A$ is reconstructed from A and aligned with G_A using graph matching techniques, such as classical structural algorithms [24] and attributed graph matching methods [32, 29]. The resulting alignment map captures structural and semantic correspondences and helps identify potential hallucinations or inconsistencies introduced during verbalization. Second, *rule-based validation* consults the domain KG to verify that inferred relations comply with domain rules and standards [4], ensuring normative correctness and auditability. Third, to establish procedural transparency and traceability, comprehensive *provenance tracking* records data sources, agents, and all transformations throughout the QA process for each response, attaching this audit trail as metadata to the factual KG. Together these mechanisms yield evidence-based reasoning [22] that extends beyond traditional chain-of-thought methods [23].

Explanation Agent generates explanations using a KG that captures key elements of the QA and validation processes, presenting retrieval and reasoning steps transparently in a user-friendly format. To balance transparency and usability, the graph could be structured into a *three-level model* that adapts the amount of detail to different user needs [3]. The *Core Level* shows the natural language question Q, context subgraph G_C, and verbalized answer A, providing an overview of the retrieval context. The *Extended Level* adds the structured query $\hat{Q}$, analytic result subgraph G_A, and alignment mapping, enabling deeper tracing of the Validation Agent's backward reasoning. The *Audit Level* offers a comprehensive rule-based validation report and provenance tracking for expert users requiring full traceability. The explanation graph supports both visual and textual rendering, clearly conveying the system's rationale and allowing users to explore each derivation step interactively.

4 Conclusion and Next Steps

In summary, the proposed multi-agent architecture enables verifiable, auditable and explainable QA for the auditing domain. By decomposing the workflow into specialized, collaborating agents, the approach achieves modularity, end-to-end transparency, and rigorous control over answer generation. The knowledge base enables both efficient retrieval of data-driven context and strict compliance checking against domain-specific rules and auditing standards. The *Validation Agent* anchors every answer in formal reasoning, regulatory requirements, and provenance evidence, ensuring outputs that are plausible, reproducible, and fully aligned with domain requirements. Complementing this, the *Explanation Agent* provides user-centred, traceable narratives that abstract complex internal reasoning, thereby deepening trust in the system's results. The next phase involves developing a GraphRAG system as the foundational agent to establish the core of the multi-agent framework. Subsequently, research will focus on identifying and implementing methods that empower the Validation Agent to robustly enforce consistency, rule compliance, and auditability. Likewise, coordination strategies among agents will be investigated to further enhance the system's robustness and operational efficiency.

Acknowledgements

Research Project Data Tales, funded by the Bavarian StMWi (DIK0660)

References

[1] B. Abu-Salih, 'Domain-specific knowledge graphs: A survey', *Journal of Network and Computer Applications*, **185**, 103076, (2021).

[2] J. Amling, E. Slany, C. Dormagen, M. Kretschmann, and S. Scheele, 'Bridging the interpretability gap in process mining: A comprehensive approach combining explainable clustering and generative ai', in *Proceedings of the 2025 Conference on Explainable Artificial Intelligence (XAI 2025)*, to appear, (2025). Accepted for publication.

[3] V. Armant, A. Mouakher, F. Vargas-Rojas, D. Symeonidou, J. Guérin, I. Mougenot, and J.-C. Desconnets, 'Can knowledge graphs and retrieval-augmented generation be combined to explain query/answer relationships truthfully?', in *Proceedings of the 4th International Workshop on Data meets Ontologies in Explainable AI (DAO-XAI) at ECAI 2024*, volume 3833, (2024).

[4] K. Baclawski, M. Bennett, G. Berg-Cross, T. Schneider, R. Sharma, M. Underwood, and A. Westerinen, 'Ontologies, neuro-symbolic and generative ai technologies: Toward trustworthy ai systems', *Journal of the Washington Academy of Sciences*, **110**(1), 1–38, (2024). Communiqué of the Ontology Summit 2024.

[5] L. Cao, 'Graphreason: Enhancing reasoning capabilities of large language models through a graph-based verification approach', in *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*, pp. 1–12, (2024).

[6] Institut der Wirtschaftsprüfer in Deutschland e.V. (IDW). Official auditing and professional standards, 2025.

[7] T. Föhr, K.-U. Marten, and M. Schreyer, 'Deep learning meets risk-based auditing: a holistic framework for leveraging foundation and task-specific models in audit procedures', *SSRN Electronic Journal*, (01 2023).

[8] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, 'Large language model based multi-agents: A survey of progress and challenges', in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 8048–8057, (2024).

[9] H. Han, Y. Wang, H. Shomer, K. Guo, J. Ding, Y. Lei, M. Halappanavar, R. A. Rossi, S. Mukherjee, X. Tang, Q. He, Z. Hua, B. Long, T. Zhao, N. Shah, A. Javari, Y. Xia, and J. Tang. Retrieval-augmented generation with graphs (graphrag), 2025.

[10] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, and T. Liu, 'A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions', *ACM Transactions on Information Systems*, **43**(2), (2025).

[11] International Auditing and Assurance Standards Board (IAASB), *2023–2024 Handbook of International Quality Management, Auditing, Review, Other Assurance, and Related Services Pronouncements*, International Federation of Accountants (IFAC), 2024.

[12] J. Li, J. Chen, R. Ruiyang, X. Cheng, X. Zhao, J. y. Nie, and J.-R. Wen, 'The dawn after the dark: An empirical study on factuality hallucination in large language models', in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10879–10899, (2024).

[13] X. Li, Y. Zhang, and E. C. Malthouse, 'Large language model agentic approach to fact checking and fake news detection', in *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, volume 392, pp. 2572–2579, (2024).

[14] X. Liu, Y. Zhu, Y. Lan, C. Yang, and Y. Qiao, 'Safety of multimodal large language models on images and text', in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 8151–8159. International Joint Conferences on Artificial Intelligence Organization, (2024).

[15] L. Luo, Y.-F. Li, G. Haffari, and S. Pan, 'Reasoning on graphs: Faithful and interpretable large language model reasoning', in *The Twelfth International Conference on Learning Representations*, (2024).

[16] S. Ma, C. Xu, X. Jiang, M. Li, H. Qu, C. Yang, J. Mao, and J. Guo, 'Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation', in *The Thirteenth International Conference on Learning Representations*, (2025).

[17] N. Mimouni and J.-C. Moissinac, 'Towards efficient exploitation of large knowledge bases by context graphs', *Studies on the Semantic Web*, **60**, 294–310, (2024).

[18] J. Z. Pan, S. Razniewski, J.-C. Kalo, S. Singhania, J. Chen, S. Dietze, H. Jabeen, J. Omeliyanenko, W. Zhang, M. Lissandrini, R. Biswas, G. de Melo, A. Bonifati, E. Vakaj, M. Dragoni, and D. Graux, 'Large language models and knowledge graphs: Opportunities and challenges', *Transactions on Graph Data and Knowledge*, **1**(1), 2:1–2:38, (2023).

[19] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, 'Unifying large language models and knowledge graphs: A roadmap', *IEEE Transactions on Knowledge and Data Engineering*, **36**(7), 3580–3599, (2024).

[20] C. T. Pedretti, M. F. Bocchi, F. Tomasi, and F. Vitali, 'What do we annotate when we annotate? towards a multi-level approach to semantic annotations', *Studies on the Semantic Web*, **60**, 370–385, (2024).

[21] T. Seidenstein, K.-U. Marten, G. Donaldson, T. Föhr, V. Reichelt, and L. Jakoby, 'Innovation in audit and assurance: A global study of disruptive technologies', *Journal of Emerging Technologies in Accounting*, **21**, 1–18, (2024).

[22] T. Sekar, K. Kushal, S. Shankar, S. Mohammed, and J. Fiaidhi, 'Investigations on using evidence-based graphrag pipeline using llm tailored for usmle style questions', *medRxiv*, (2025).

[23] M. Shirdel, J. Rorseth, P. Godfrey, L. Golab, D. Srivastava, and J. Szlichta, 'Aprèscot: Explaining llm answers with knowledge graphs and chain of thought', in *Proceedings of the 28th International Conference on Extending Database Technology (EDBT)*, pp. 1142–1145, (2025).

[24] K. Skitsas, K. Orłowski, J. Hermanns, D. Mottin, and P. Karras, 'Comprehensive evaluation of algorithms for unrestricted graph alignment', in *Proceedings of the 26th International Conference on Extending Database Technology (EDBT)*, pp. 260–272, (2023).

[25] E. Slany, S. Scheele, and U. Schmid, 'Explanatory interactive machine learning with counterexamples from constrained large language models', in *KI 2024: Advances in Artificial Intelligence*, eds., Andreas Hotho and Sebastian Rudolph, volume 14992, 324–331, Springer, (2024).

[26] T. Sun, J. Carr, and D. Kazakov, 'A hybrid question answering model with ontological integration for environmental information', in *Proceedings of the 4th International Workshop on Data meets Ontologies in Explainable AI (DAO-XAI) at ECAI 2024*, volume 3833, (2024).

[27] S. Tsaneva, D. Dessì, F. Osborne, and M. Sabou, 'Knowledge graph validation by integrating llms and human-in-the-loop', *Information Processing & Management*, **62**(5), 104145, (2025).

[28] D. Vollmers, P. Sharma, H. M. Zahera, and A.-C. Ngonga Ngomo, 'Enhancing answers verbalization using large language models', in *SEMANTICS*, pp. 345–352, (2024).

[29] Z. Wang, N. Zhang, W. Wang, and L. Wang, 'On the feasible region of efficient algorithms for attributed graph alignment', in *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 1163–1168, (2022).

[30] J. Wu, X. Wu, and J. Yang, 'Guiding clinical reasoning with large language models via knowledge seeds', in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 7491–7499. International Joint Conferences on Artificial Intelligence Organization, (2024).

[31] Rong Wu, Pinlong Cai, Jianbiao Mei, Licheng Wen, Tao Hu, Xuemeng Yang, Daocheng Fu, and Botian Shi. KG-TRACES: Enhancing Large Language Models with Knowledge Graph-constrained Trajectory Reasoning and Attribution Supervision, 2025. Preprint.

[32] N. Zhang, W. Wang, and L. Wang, 'Attributed graph alignment', in *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 1829–1834, (2021).

[33] Y. Zhou, Y. Liu, X. Li, J. Jin, H. Qian, Z. Liu, C. Li, Z. Dou, T.-Y. Ho, and P. Yu. Trustworthiness in retrieval-augmented generation systems: A survey, 2024.