Towards LLM-KG Symbiosis for Reducing Factual Hallucinations

Tommaso Dolci[1,*], Milos Jovanovik[1,2] and Katja Hose[1]

[1]*Institute of Logic and Computation, TU Wien, Austria*

[2]*Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia*

Abstract

The widespread adoption of Large Language Models (LLMs) has increased concerns about hallucinations, i.e., the generation of incorrect or nonsensical claims. While popular approaches to reduce hallucinations (e.g., RAG) are promising, they still suffer from intrinsic hallucinations and remain largely limited to closed-domain scenarios, where the external source of knowledge is complete and sufficient to generate a response. Recently, knowledge graphs (KGs) have emerged as trustworthy sources to detect and mitigate hallucinations either before or after generation, but their adoption remains challenging for open-domain questions and long responses containing a mixture of correct, incorrect, and opinionated claims. This paper discusses the main opportunities and limitations of current approaches for reducing LLM hallucinations by KG grounding, and presents a framework for LLM-KG symbiosis to address the following open challenges: factuality assessment of multiple-claim responses, KG-grounded retrieval under incomplete data, and uncertainty management.

Keywords

Large Language Models, Knowledge Graphs, Hallucinations, Uncertainty

1. Introduction

Recent advances of Large Language Models (LLMs) have accelerated the adoption of AI across multiple domains, including high-risk scenarios such as legal or healthcare support. While LLMs offer great potential for task automation, their widespread adoption and the increasing trust of users in LLM-based chatbots (e.g., ChatGPT[1] and Claude[2]) raise concerns about AI reliability, especially when these systems are treated as authoritative sources of knowledge [1]. In fact, LLMs often generate hallucinations, i.e., incorrect or nonsensical information [2, 3]. Hallucinations represent both a technical challenge and a risk for society. Technically, factual inaccuracies are considered an intrinsic problem of LLMs under current training and evaluation methods [4, 5]. Socially, hallucinations can misinform users and pose safety threats, as exemplified by the recently-introduced Google AI Overviews[3] suggesting adding glue on pizza or eating one small rock per day.[4]

Recently, knowledge graphs (KGs), knowledge bases including semantics and fact representations in the form of triples, have emerged as a prominent solution for reducing LLM hallucinations [6]. In fact, KGs contain verified factual knowledge curated from trustworthy sources, e.g., Wikipedia [7]. Many KG-enhanced methods for hallucination mitigation have been proposed, from KG-augmented retrieval [8] to response retrofitting [9] and knowledge injection [10]. However, many solutions rely entirely on external knowledge retrieval, which may fail due to KG incompleteness or absence of up-to-date information. These approaches are often evaluated on information-retrieval tasks (e.g., KG question-answering [11]), assuming that the external source of knowledge is complete and sufficient to answer the question (*closed-knowledge scenario*). However, general-purpose chatbots frequently operate

Published in the Proceedings of the Workshops of the EDBT/ICDT 2026 Joint Conference (March 24-27, 2026), Tampere, Finland

*Corresponding author.

tommaso.dolci@tuwien.ac.at (T. Dolci); milos.jovanovik@tuwien.ac.at (M. Jovanovik); katja.hose@tuwien.ac.at (K. Hose)
0000-0002-1403-7766 (T. Dolci); 0000-0001-7360-8015 (M. Jovanovik); 0000-0001-7025-8099 (K. Hose)

[1] openai.com/index/chatgpt

[2] anthropic.com/news/introducing-claude

[3] search.google/ways-to-search/ai-overviews

[4] bbc.com/news/articles/cd11gzejgz4o

Table 1
Comparison between current approaches for reducing LLM hallucinations by KG grounding and our LLM-KG symbiosis framework, highlighting the characteristic opportunities of each approach. The symbol ❍ indicates partial coverage.

Mitigation Approach	Notable Works	Opportunities				
		Implicit Knowledge	External Knowledge	Reasoning on KG	Multiple-Claim Evaluation	Uncertainty
No Mitigation (standard LLM)	[16, 17]	✓	–	–	–	–
KG-Based Retrieval	[18]	–	✓	–	–	–
KG-Based Retrieval with Reasoning	[10, 19]	❍	✓	✓	–	–
KG-Based Comparison	[9]	✓	✓	–	–	–
KG-Based Comparison on Long Text	[12, 20]	✓	✓	–	❍	❍
LLM-KG Symbiosis		✓	✓	✓	✓	✓

under *open-knowledge scenarios*, where the required knowledge is not restricted to a specific domain or a predefined source. Moreover, long responses frequently contain a mixture of factual, non-factual, and unverifiable claims due to their opinionated nature or the lack of external reference knowledge.

In this context, the following limitations emerge: *i)* evaluating hallucinations beyond binary classification [12] and accuracy in question-answering [13, 14], *ii)* insufficient address of KG incompleteness in open-knowledge scenarios, where multiple (potentially contradicting) sources may be needed and factual counterparts for claim verification may not be available, and *iii)* a limited consideration of uncertainty in evaluating LLM hallucinations [5]. While the combination of LLMs and KGs is considered a viable solution to address each other's limitations, e.g., LLM hallucinations and KG incompleteness [15], the implementation of synergistic LLM-KG solutions remains largely unexplored.

This paper proposes an LLM-KG framework for reducing hallucinations after text generation, suitable for open-domain questions and evaluation of long responses under uncertainty. Long responses are especially challenging because they can contain a mixture of factual claims, hallucinated claims to mitigate, and opinionated claims to be assessed for misinformation and polarization. Adopting a strategy for reducing hallucinations after text generation allows both to evaluate LLM hallucinations for testing purpose (even for closed-source models) and to mitigate factual inaccuracies at runtime without overly limiting the generative power of LLMs. Our framework is a first step towards LLM-KG symbiosis for reducing factual hallucinations, i.e., an approach to strengthen bi-directional synergy between LLMs and KGs to mitigate hallucinations and misinformation. In this context, KGs support LLMs by providing factual evidence and trustworthy knowledge, while LLMs enhance KGs with reasoning and semantic-awareness to expand incomplete knowledge (e.g., by link prediction).

The rest of this paper is organized as follows: Section 2 discusses current approaches to reduce factual hallucinations by KG grounding, Section 3 highlights open challenges in current approaches, Section 4 describes our LLM-KG symbiosis framework for reducing factual hallucinations, and Section 5 concludes the paper and outlines future work.

2. State of the Art

Surveys typically distinguish two types of hallucinations: *intrinsic*, where generated content contradicts the provided input or context (e.g., in text summarization) and *extrinsic*, where claims can only be verified from external sources [1, 3]. While intrinsic hallucinations can be detected by comparing against the LLM context, factual hallucinations require access and retrieval from external trustworthy sources, e.g., documents or knowledge bases. In this paper, we focus on *factual hallucinations*, i.e., extrinsic hallucinations that either contradict real-world information or fabricate new false knowledge [2]. Factual hallucination detection commonly relies on self-consistency via multiple invocation [21] or confidence estimation [22]. However, these methods assess internal consistency rather than comparing claims with verifiable facts, failing when LLMs are over-confident in false knowledge or under-confident in correct claims. KGs can address this issue by providing curated, verified knowledge for reducing hallucinations, typically through *KG-based retrieval* (pre-generation) or *KG-based comparison* (post-generation).

KG-Based Retrieval KG-based retrieval is typically enabled by RAG (retrieval-augmented generation) [23], a popular approach to reduce factual hallucinations by grounding generation in external documents or knowledge bases [24]. GraphRAG, first introduced by Microsoft Research [18], expanded standard RAG to include graph structures, leading to higher accuracy in question-answering tasks, fewer hallucinations, and improved LLM reasoning [25]. However, both RAG and GraphRAG can only retrieve the knowledge contained in the external sources. Therefore, they adopt a *closed-knowledge assumption* that limits applicability in real-world scenarios, where even curated encyclopedic KGs like Wikidata [7] or YAGO [26] may turn out to be incomplete, outdated, or insufficient to answer a question. Additionally, intrinsic hallucinations remain a latent issue of KG-based retrieval, which emerges when the LLM summarizes and rewrites the retrieved knowledge into the final response.

To overcome the limitations of standard KG-based retrieval approaches, recent works proposed to detect and mitigate hallucinations by synergizing LLM reasoning and KG-grounded information. Among pre-generation approaches, MindMap [10], Think-on-Graph [13], and Reasoning-on-Graphs [14] leverage LLM reasoning to extract more information from graphs. Generate-on-Graph [19] addresses KG incompleteness by leveraging both LLM parametric knowledge and reasoning to augment externally retrieved graphs, while Knowledge Injection [27] injects KG information into the LLM prompt for in-context learning. Despite adding LLM reasoning, these approaches still rely on external knowledge retrieval, constraining applications to closed-knowledge scenarios.

KG-Based Comparison KG-based comparison verifies LLM responses by detecting and mitigating hallucinations after text generation. These approaches extract claims from LLM responses to compare against external KG facts, e.g., KG-based Retrofitting [9] verifies LLM claims against external facts and retrofits incorrect claims in the original response according to factual evidence. To assess long responses containing multiple claims, FactScore [28] decomposes text and compares LLM claims against Wikipedia to estimate hallucination metrics, although without relying on graph structures. GraphEval [12] transforms responses into graphs and compares them with contextual factual triples by framing the problem as a *natural language inference* task, albeit focusing only on intrinsic hallucinations. Most notably, FLEEK [20] extracts individual claims from LLM responses, verifies each claim against a KG triple, and suggests factual corrections. FLEEK classifies LLM claims as supported by the external KG, unsupported or questionable, including a degree of uncertainty in factuality evaluation. However, the system does not address KG incompleteness or calculate sentence-level hallucination scores.

In summary, KG-based retrieval approaches are limited by KG incompleteness and intrinsic hallucinations. KG-based comparison approaches avoid intrinsic hallucinations but still require the completeness of reference knowledge (e.g., to be improved by reasoning on graphs) and evaluation strategies for multiple-claim assessment under uncertainty. Table 1 compares the current KG-based approaches for reducing hallucinations, highlighting the main opportunities for each approach and describing how LLM-KG symbiosis fits into the current state of the art.

3. Open Challenges

In this section, we describe the three main open challenges in current KG-based approaches for reducing hallucinations, exemplified in Figure 1.

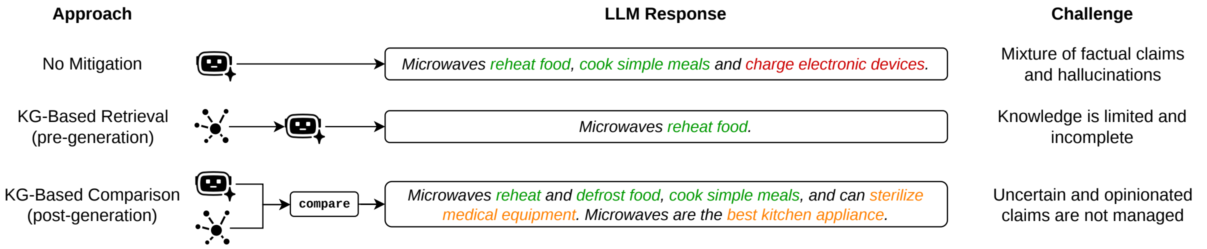
Approach
LLM Response
Challenge
No Mitigation
Microwaves reheat food, cook simple meals and charge electronic devices.
Mixture of factual claims and hallucinations
KG-Based Retrieval (pre-generation)
Microwaves reheat food.
Knowledge is limited and incomplete
KG-Based Comparison (post-generation)
compare
Microwaves reheat and defrost food, cook simple meals, and can sterilize medical equipment. Microwaves are the best kitchen appliance.
Uncertain and opinionated claims are not managed

Figure 1: Open challenges of current approaches for reducing LLM hallucinations by KG grounding, answering the question *What are common uses of a microwave oven?* No mitigation approaches produce hallucinations, KG-based retrieval approaches (e.g., GraphRAG [25]) reduce factual hallucinations but limit the contents of the response to the source of knowledge, KG-based comparison approaches fall short of addressing uncertainty and opinionated claims (e.g., KG-based Retrofitting [9]).

C1 – Multiple-Claim Evaluation Longer LLM responses contain multiple claims (some trustworthy and some hallucinated) to be extracted and evaluated both separately and in combination. After claim-level evaluation, appropriate response-level hallucination metrics should be computed. Addressing this challenge involves *(i)* improving claim decomposition for long-form text, *(ii)* developing strategies for aggregating claim-level evaluations into an overall response-level hallucination score, and *(iii)* defining the meaning and application of hallucination metrics: while fine-grained evaluations are generally more informative, binary classifiers at response-level may be more adequate for critical use cases such as medical diagnosis.

C2 – Incomplete Knowledge While KGs are trustworthy sources of knowledge, they are also limited and often incomplete. As a result, KG-based retrieval is limited outside of closed-knowledge scenarios where the reference knowledge is assumed complete, e.g., domain-specific question-answering. KG incompleteness is also a challenge for KG-based comparison approaches, which also need fact extraction from KGs. Therefore, addressing this challenge involves *(i)* considering multiple KG sources for detection and mitigation across domains, and combining evidence from heterogeneous sources (e.g., encyclopedic [7], commonsense [29], domain-specific [30] knowledge bases) by handling inconsistencies and potential contradictions, *(ii)* improving KG navigation to expand the available reference knowledge through LLM reasoning and link prediction, accounting for different degrees of factual evidence (e.g., logically-entailed facts and semantically-related facts), and *(iii)* combining different approaches for claim-fact comparison, e.g., converting triples into text, or generating triples from text.

C3 – Uncertainty Not all content generated by LLMs should be evaluated. A flexible solution for detecting and mitigating hallucinations must differentiate between scenarios that require factual correctness and scenarios where responses should include uncertainty and multiple perspectives. Addressing this challenge involves *(i)* differentiating between claims that require fact-checking and claims that should be treated as opinions or unverifiable products of creativity (e.g., in scenarios such as storytelling, use-case definition, or hypothesis formulation), and *(ii)* avoiding opinionated claims that may favor misinformation and unfairness even if they do not contain hallucinations (e.g., fostering Western-centric perspectives). While improving factual accuracy is fundamental, polarized opinions can propagate historical biases, misinformation, and lack of diversity.

Finally, these challenges intersect each other: multiple-claim evaluation requires navigating and integrating multiple sources to overcome limited and incomplete knowledge in KGs; uncertainty should be accounted for when computing hallucination metrics, because longer responses can contain a mixture of factual, non-factual, and opinionated claims; uncertainty should also be considered during KG-retrieval, when inconsistent or contrasting facts arise from incomplete KGs or from reasoning on multiple external sources.

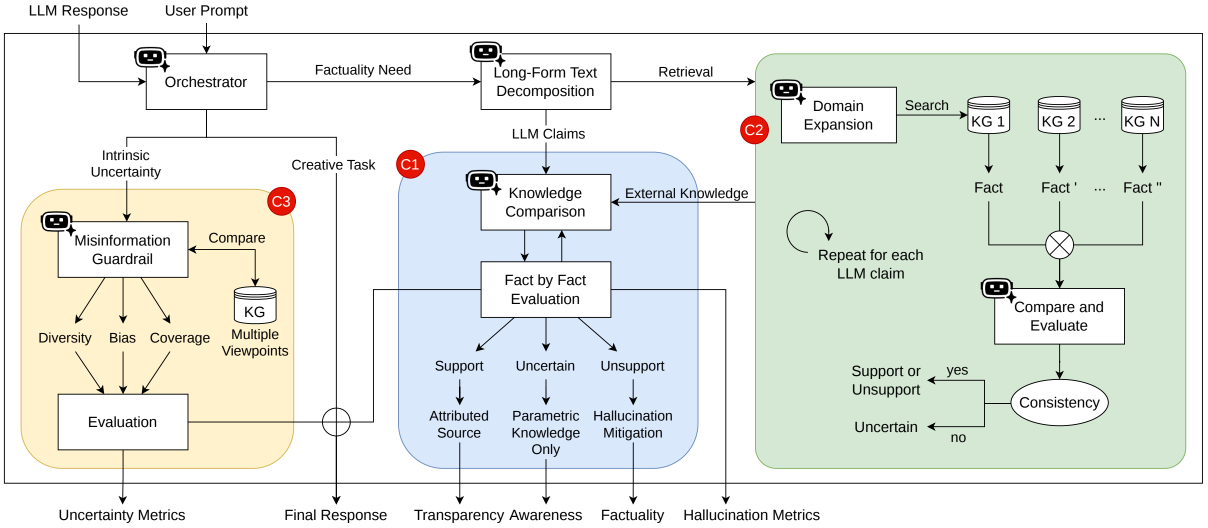

Figure 2: LLM-KG symbiosis framework for hallucination detection and mitigation. In the center, the *Evaluation Module* compares LLM claims with external factual knowledge to identify and mitigate hallucinations. On the right, the *Retrieval Module* searches for relevant knowledge across multiple KGs (encyclopedic or domain-specific) to extract factual evidence. On the left, the *Uncertainty Module* identifies uncertainty and safeguards against misinformation and bias.

4. Towards LLM-KG Symbiosis

While many approaches for reducing LLM hallucinations by KG grounding have recently emerged [6, 8], addressing the open challenges outlined in Section 3 requires a deeper LLM-KG integration [15]. To this end, we envision a modular framework for LLM-KG symbiosis (Figure 2) to achieve hallucination reduction by combining LLM parametric knowledge, reasoning, and semantic understanding with external, deterministic processes for KG retrieval and knowledge expansion. The modular nature of the framework allows combining existing solutions (e.g., for claim-to-fact comparison [12, 31]) with new approaches for evaluating hallucinations in multiple-claim responses under uncertainty. This framework distinguishes responses that require factual accuracy from creative or intrinsically uncertain responses, such as personal recommendations or political discussions. Text demanding factual accuracy is decomposed into claims to compare against external KGs. Intrinsically uncertain responses are subject to further actions to ensure trustworthiness beyond factuality (e.g., evaluating diversity and bias to avoid misinformation). Creative texts do not require factual assessment. The *Orchestrator* is tasked to identify whether an LLM response should be assessed for factual accuracy and evaluated for misinformation and bias, depending on the content of the response and the nature of the user prompt. This module can be implemented similarly to orchestration in multi-agentic systems, where the orchestrator agent receives an input task and distributes it accordingly to the worker agents [32].

Our framework addresses the open challenges discussed in Section 3 in three separate modules, outlined in the paragraphs below.

Evaluation Module (blue box in the center in Figure 2) addresses **C1** by comparing LLM claims to KG facts, supporting different comparison strategies and aggregating claim-level evaluations into overall response-level metrics. This module is motivated by the need for new evaluation metrics and benchmarks in hallucination detection [5, 33]. Claims supported by external knowledge are attributed for increased transparency, while unsupported claims are mitigated with external factual knowledge. Unverifiable claims (e.g., factual fabrication [2] or inconsistent retrieval results) are mitigated by informing the user, raising awareness and caution about LLM-generated claims. Hallucination metrics aggregate claim-level evaluation across two dimensions. A *quantitative* dimension measures the number of supported and unsupported claims in the response. A *qualitative* dimension measures the distance between unsupported claims and KG facts both in terms of semantics and graph distance. Finally, for unverifiable LLM claims, a plausibility score should be estimated by measuring the semantic and structural similarity to existing graph content, i.e., entities and relations are evaluated for meaning alignment and structural compatibility with known triples.

Retrieval Module (green box on the right in Figure 2) addresses **C2** by KG-grounded hallucination detection from multiple and incomplete sources. This module tackles open-knowledge scenarios by considering multiple KGs from different domains depending on the user question, managing uncertainty and inconsistency (e.g., lack of supporting evidence or multiple sources providing conflicting information). Here, support from LLMs introduces a deeper layer of reasoning and semantic-awareness to address the limitation of incomplete knowledge (e.g., performing link prediction). Reasoning-based approaches for expanding incomplete knowledge have obtained promising results in question-answering tasks, showing that LLM reasoning can effectively address KG incompleteness [19].

Uncertainty Module (yellow box on the left in Figure 2) addresses **C3** for uncertainty identification and evaluation. Opinionated and intrinsically uncertain statements require an assessment that goes beyond factual accuracy. To avoid misinformation and polarization, adequate evaluation of viewpoint diversity and cultural coverage must be achieved (e.g., by comparing against multi-cultural knowledge bases [34]). Moreover, introducing a set of *uncertainty metrics* complements factual accuracy to achieve reliable and trustworthy AI. For instance, uncertainty metrics can evaluate diversity in LLM recommendations [35] or assess multi-cultural coverage and biases using external KGs that map cultural knowledge and stereotypes [36].

5. Conclusion and Future Work

This paper presented a framework with a modular architecture to enable LLM-KG symbiosis for hallucination detection and mitigation. LLM-KG symbiosis addresses the main open challenges in current KG-based approaches: factuality assessment and evaluation in the context of multiple-claim responses, KG-grounded retrieval under incomplete data, and uncertainty management.

The main challenge of the proposed framework is about computational complexity, especially for real-time factuality evaluation. In fact, the Retrieval Module searches for factual evidence for each LLM claim, querying large encyclopedic or domain-specific KGs. To address this challenge, a first step is to consider a *hierarchy of knowledge*, where external KGs are selectively and orderly queried based on their probability of containing relevant triples, e.g., biomedical KGs queried first for medical and pharmaceutical questions.

Future work includes implementing and testing the three modules composing the framework. Additionally, an important step is to define aggregated hallucination metrics under uncertainty, to enable factuality assessment in long responses containing multiple claims. Modules are designed as stand-alone solutions to a challenge and can incorporate different LLM techniques for advanced reasoning on KGs (e.g., chain-of-thought [37] or plan-and-solve [38]). Moreover, this framework can support agentic AI integration, which has been recently introduced to data exploration tasks [39]. The ReAct agent in particular has been successfully tested for mitigating hallucinations [40]. To safeguard against misinformation, an important step is to consider historical and cultural biases in KGs [41], their impact on evaluating hallucinations, and how they can affect uncertainty verification (e.g., defining “famous people” primarily from Western-centric sources).

Acknowledgments

This work has been partially supported by ARMADA, funded by the European Union’s Horizon Europe Marie Skłodowska-Curie Actions (MSCA), under grant agreement No. 101168951.

Declaration on Generative AI

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

References

[1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of Hallucination in Natural Language Generation, ACM Comput. Surv. 55 (2023) 248:1–248:38.

[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) 42:1–42:55.

[3] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T. Luu, W. Bi, F. Shi, S. Shi, Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models, Computational Linguistics (2025) 1–46.

[4] S. Banerjee, A. Agarwal, S. Singla, LLMs Will Always Hallucinate, and We Need to Live with This, in: Intelligent Systems and Applications, Springer, 2025, pp. 624–648.

[5] A. T. Kalai, O. Nachum, S. S. Vempala, E. Zhang, Why Language Models Hallucinate, CoRR abs/2509.04664 (2025).

[6] E. Lavrinovics, R. Biswas, J. Bjerva, K. Hose, Knowledge Graphs, Large Language Models, and Hallucinations: An NLP Perspective, J. Web Semant. 85 (2025) 100844.

[7] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85.

[8] G. Agrawal, T. Kumarage, Z. Alghamdi, H. Liu, Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey, in: NAACL-HLT, ACL, 2024, pp. 3947–3960.

[9] X. Guan, Y. Liu, H. Lin, Y. Lu, B. He, X. Han, L. Sun, Mitigating Large Language Model Hallucinations via Autonomous Knowledge Graph-Based Retrofitting, in: AAAI, AAAI Press, 2024, pp. 18126–18134.

[10] Y. Wen, Z. Wang, J. Sun, MindMap: Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models, in: ACL (1), ACL, 2024, pp. 10370–10388.

[11] C. Ma, Y. Chen, T. Wu, A. Khan, H. Wang, Large Language Models Meet Knowledge Graphs for Question Answering: Synthesis and Opportunities, in: EMNLP, ACL, 2025, pp. 24589–24608.

[12] H. Sansford, N. Richardson, H. P. Maretic, J. N. Saada, GraphEval: A Knowledge-Graph Based LLM Hallucination Evaluation Framework, in: KiL@KDD, volume 3894, CEUR-WS.org, 2024, pp. 20–31.

[13] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, L. Ni, H.-Y. Shum, J. Guo, Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph, in: ICLR, 2023.

[14] L. Luo, Y.-F. Li, G. Haffari, S. Pan, Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning, in: ICLR, 2023.

[15] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying Large Language Models and Knowledge Graphs: A Roadmap, IEEE Trans. Knowl. Data Eng. 36 (2024) 3580–3599.

[16] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. R. et al., Training language models to follow instructions with human feedback, in: NeurIPS, 2022.

[17] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. B. et al., Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023.

[18] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, J. Larson, From local to global: A graph RAG approach to query-focused summarization, CoRR abs/2404.16130 (2024).

[19] Y. Xu, S. He, J. Chen, Z. Wang, Y. Song, H. Tong, G. Liu, J. Zhao, K. Liu, Generate-on-Graph: Treat LLM as both Agent and KG for Incomplete Knowledge Graph Question Answering, in: EMNLP, ACL, 2024, pp. 18410–18430.

[20] F. Fatahi Bayat, K. Qian, B. Han, Y. Sang, A. Belyy, S. Khorshidi, F. Wu, I. Ilyas, Y. Li, FLEEK:

Factual Error Detection and Correction with Evidence Retrieved from External Knowledge, in: EMNLP (Demos), ACL, 2023, pp. 124–130.

[21] P. Manakul, A. Liusie, M. Gales, SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models, in: EMNLP, ACL, 2023, pp. 9004–9017.

[22] N. Varshney, W. Yao, H. Zhang, J. Chen, D. Yu, A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation, CoRR abs/2307.03987 (2023).

[23] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. e. a. Rocktäschel, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: NeurIPS, 2020.

[24] C. Wang, X. Liu, Y. Yue, Q. Guo, X. Hu, X. Tang, T. Zhang, C. Jiayang, Y. Yao, X. e. a. Hu, Survey on Factuality in Large Language Models, ACM Comput. Surv. 58 (2025) 13:1–13:37.

[25] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, S. Tang, Graph Retrieval-Augmented Generation: A Survey, ACM Trans. Inf. Syst. 44 (2025) 35:1–35:52.

[26] F. M. Suchanek, G. Kasneci, G. Weikum, Yago: A core of semantic knowledge, in: WWW, ACM, 2007, pp. 697–706.

[27] A. Martino, M. Iannelli, C. Truong, Knowledge Injection to Counter Large Language Model (LLM) Hallucination, in: The Semantic Web: ESWC 2023 Satellite Events, Springer, 2023, pp. 182–185.

[28] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, H. Hajishirzi, FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation, in: EMNLP, ACL, 2023, pp. 12076–12100.

[29] R. Speer, J. Chin, C. Havasi, ConceptNet 5.5: An Open Multilingual Graph of General Knowledge (2017) 4444–4451.

[30] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (2004) 267–270.

[31] M. Rashad, A. Zahran, A. Amin, A. Abdelaal, M. Altantawy, FactAlign: Fact-Level Hallucination Detection and Classification Through Knowledge Graph Alignment, in: Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024), ACL, 2024, pp. 79–84.

[32] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, X. Zhang, Large Language Model Based Multi-agents: A Survey of Progress and Challenges, in: IJCAI, volume 9, 2024, pp. 8048–8057.

[33] A. Kulkarni, Y. Zhang, J. R. A. Moniz, X. Ge, B.-H. Tseng, D. Piraviperumal, S. Swayamdipta, H. Yu, Evaluating Evaluation Metrics – The Mirage of Hallucination Detection, in: EMNLP (Findings), ACL, 2025, pp. 19013–19032.

[34] W. Shi, R. Li, Y. Zhang, C. Ziems, S. Yu, R. Horesh, R. A. D. Paula, D. Yang, CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies, in: EMNLP (Findings), ACL, 2024, pp. 4996–5025.

[35] S. K. Sakib, A. Bijoy Das, Challenging Fairness: A Comprehensive Exploration of Bias in LLM-Based Recommendations, in: IEEE Big Data, 2024, pp. 1585–1592.

[36] A. Deshpande, D. Ruiter, M. Mosbach, D. Klakow, StereoKG: Data-Driven Knowledge Graph Construction For Cultural Knowledge and Stereotypes, in: Sixth Workshop on Online Abuse and Harms, ACL, 2022, pp. 67–78.

[37] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, in: NeurIPS, 2022.

[38] L. Wang, W. Xu, Y. Lan, Z. Hu, Y. Lan, R. K.-W. Lee, E.-P. Lim, Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models, in: ACL (1), ACL, 2023, pp. 2609–2634.

[39] S. Amer-Yahia, Intelligent Agents for Data Exploration, Proc. VLDB Endow. 17 (2024) 4521–4530.

[40] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, Y. Cao, ReAct: Synergizing Reasoning and Acting in Language Models, in: ICLR, 2023.

[41] A. Kraft, R. Usbeck, The Lifecycle of “Facts”: A Survey of Social Bias in Knowledge Graphs, in: AACL/IJCNLP (1), ACL, 2022, pp. 639–652.