Mitigating Hallucination by Integrating Knowledge Graphs into LLM Inference – a Systematic Literature Review

Robin Wagner **Emanuel Kitzelmann** **Ingo Boersch**
Brandenburg University of Applied Sciences
Brandenburg an der Havel, Germany
{robin.wagner, emanuel.kitzelmann, ingo.boersch}@th-brandenburg.de

Abstract

Large Language Models (LLMs) demonstrate strong performance on different language tasks, but tend to hallucinate – generate plausible but factually incorrect outputs. Recently, several approaches to integrate Knowledge Graphs (KGs) into LLM inference were published to reduce hallucinations. This paper presents a systematic literature review (SLR) of such approaches. Following established SLR methodology, we identified relevant work by systematically search in different academic online libraries and applying a selection process. Nine publications were chosen for in-depth analysis. Our synthesis reveals differences and similarities of how the KG is accessed, traversed, and how the context is finally assembled. KG integration can significantly improve LLM performance on benchmark datasets and additionally to mitigate hallucination enhance reasoning capabilities, explainability, and access to domain-specific knowledge. We also point out current limitations and outline directions for future work.

1 Introduction

The performance of large language models (LLMs) has made significant progress in recent years (Zhao et al., 2024; Wang et al., 2024). Their ability to understand and answer questions in natural language makes them popular tools in many industries (Hadi et al., 2023). However, due to their architecture, LLMs tend to "hallucinate" plausible but factually incorrect answers (Huang et al., 2024). This reduces the applicability of LLMs, especially in sensitive domains such as, e.g., medicine. The aim of this review is to investigate how the integration of knowledge graphs (KGs) into the inference processes of LLMs can help mitigate hallucinations. We analyze how KGs can be used as a structured source of knowledge to improve the reliability and factual accuracy of model answers, what other advantages this integration offers and what challenges

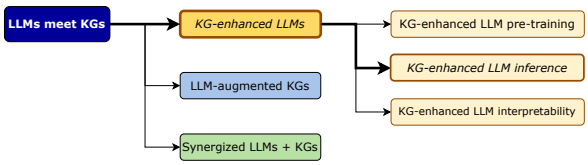

Figure 1: Categorization of current approaches to integrate LLMs and KGs according to (Pan et al., 2024).

are associated with it. For this purpose, a systematic literature review (Keele et al., 2007) of publications that propose approaches for integrating KGs into the LLM inference phase is conducted.

The combination of LLMs and KGs has already been investigated in other systematic literature reviews. Ibrahim et al. (Ibrahim et al., 2024) provide a comprehensive survey on integrating KGs with LLMs, highlighting key paradigms, methodologies, and challenges in this rapidly evolving field. (Pan et al., 2024) provide a comprehensive overview of how LLMs and KGs can be combined for different purposes. To this end, they categorize previous research into three groups and each group into subgroups (Fig. 1). The literature examined in this review could be categorized as "KG-enhanced LLMs" and therein as "KG-enhanced LLM inference", according to (Pan et al., 2024). Furthermore, the focus in this review is on the mitigation of hallucinations. (Agrawal et al., 2024) investigate the integration of KGs for the mitigation of hallucinations in LLMs. In addition to inference, they also consider other LLM-related processes such as pre-training, fine-tuning and validation for the integration of KGs (Fig. 2). Our review is limited to the area of "knowledge-aware inference" in the context of KGs.

The rest of the paper is structured as follows: In Section 2 we provide necessary background on LLMs and KGs. In Section 3 we describe the methodology that we used to conduct the literature review, including research questions, databases and

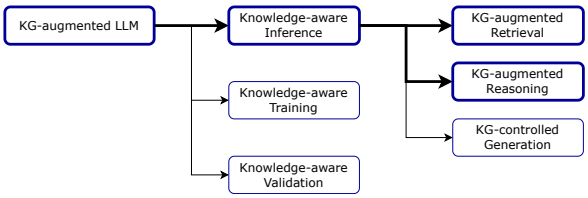

Figure 2: Categorization of current approaches to KG-supported mitigation of hallucinations according to (Agrawal et al., 2024).

criteria for selecting and evaluating relevant literature. In Section 4 we briefly overview all reviewed papers that present different approaches to integrate KGs into LLMs. Section 5 contains the synthesis of the results of the literature review to identify patterns, benefits and challenges. Finally, we conclude with Section 6 where we summarize the key findings.

2 Background

LLMs (Zhao et al., 2024; Wang et al., 2024) are language models that can understand and answer queries in natural language. In a complex training phase, they learn language patterns from huge text corpora. In the inference phase, the learned knowledge (in the form of model weights) is used to generate answers to queries. LLMs use learned language patterns to calculate probabilities for possible next tokens based on the query and the tokens generated so far. Due to their statistical and probabilistic nature, LLMs are prone to hallucinations (Huang et al., 2024). Hallucinations are coherent, plausible, but factually wrong answers. In order to increase the reliability of LLMs, various methods for mitigating hallucinations have been proposed in recent years.

Retrieval Augmented Generation (RAG) (Lewis et al., 2020) combines LLMs with external knowledge sources. Traditional RAG systems compare semantic vector representations ("embeddings") of the query and of chunks of the external knowledge, i.e., semantic similarity of query and knowledge chunks, in order to retrieve suitable chunks that contain the necessary knowledge to answer the question. This knowledge is then inserted as context to answer the query into the prompt for the LLM. Thereby, the probability of hallucinations can significantly be reduced.

In addition to documents, knowledge graphs (Hogan et al., 2021) can serve as an external source of knowledge. Knowledge graphs consist of a set of entities (nodes) and relations (directed edges) between them. A graph therefore basically consists of *triples* with subject entity, relation and object entity (e.g. Berlin $-$capital_of$\rightarrow$ Germany). A *reasoning path* is a concatenation of such triples and can serve the LLM as a context for answering complex questions (e.g. Berlin $-$capital_of$\rightarrow$ Germany $-$in_continent$\rightarrow$ Europe). To find such paths, patterns in the form of *relation paths* can be used to find entities based on a start entity: (Berlin $-$capital_of$\rightarrow$? $-$in_continent$\rightarrow$?).

3 Methodology

The present paper aims at answering the following research questions: i) How can KGs be integrated into LLM inference to mitigate hallucinations? ii) What is the structure of the integrated KGs and where do they come from? iii) To what extent does the integration of KGs improve the quality of LLM answers? iv) What other advantages does the integration of KGs have? v) What challenges arise when integrating KGs?

The following academic databases were used: IEEE Xplore, ACM Digital Library and Google Scholar. IEEE Xplore and ACM Digital Library are internationally important libraries for scientific and technical literature. Google Scholar is a freely accessible search engine for scientific literature. According to the research questions, the search focused on LLMs, KGs and hallucinations. Since the search at the ACM Digital Library led to many irrelevant results, the search string here was restricted by excluding irrelevant tasks. Search strings and results are shown in Tab. 1.

Only publications fulfilling the following conditions were kept: i) The publication is in English. ii) It is a primary source (no surveys etc.). iii) The publication is peer reviewed or is cited more than 50 times. iv) The integration of KGs in LLM inference is a main topic. These preselected publications were assessed according to their relevance. For this purpose, several questions were asked for each publication and assigned a score (see Tab. 2). The nine publications with the highest score were included for in-depth analysis and synthesis. The number of results after each step of this literature search and selection process is shown in Fig. 3.

In order to obtain a complete overview of the selected literature and thus recognize patterns, relevant information was extracted from each publication using a data extraction scheme (see Tab. 3).

Name	Search string	Date	Result
IEEE Xplore	(“llm*” OR “large language model*”) AND “knowledge graph*” AND (“infer*” OR “reason*” OR “retriev*”) AND “hallucinate*”	16.12.2024	18
ACM Digital Library	(“llm” OR “large language model”) AND “knowledge graph” AND (“inference” OR “reasoning” OR “retrieval”) AND “hallucination” AND NOT (“completion” OR “construction”)	29.12.2024	35
Google Scholar	(“llm” OR “large language model”) AND “knowledge graph” AND (“inference” OR “reasoning” OR “retrieval”) AND “hallucination”	30.12.2024	Top 50

Table 1: Search queries on LLMs, knowledge graphs and hallucination

ID	Question	Points
1	Is the interaction between LLM inference and KGs comprehensible and described in detail?	3
2	Are the source and structure of the KG clearly presented?	1
3	Is the goal of integrating KGs clearly stated?	1
4	Is the specific language model mentioned?	0.5
5	Is the approach presented as generally applicable?	1
6	Can the approach be understood in concrete terms?	1
7	Is the approach evaluated quantitatively?	1
8	Is the approach compared with similar procedures with or without KGs?	1
9	Are limitations or disadvantages of the approach discussed?	1

Table 2: Criteria to select papers on LLMs and knowledge graphs for analysis

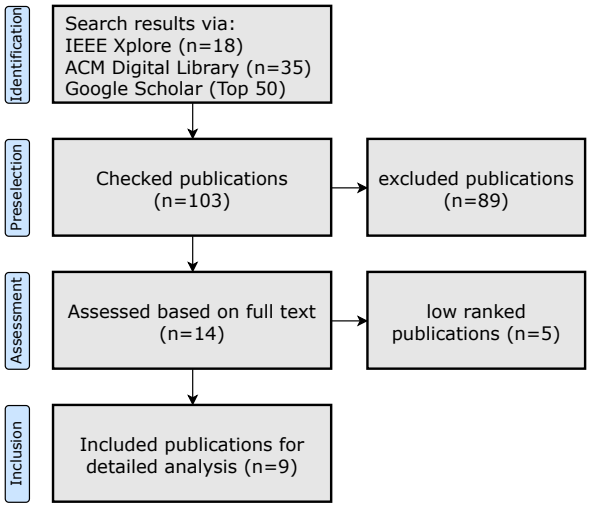

Figure 3: Selection process.

The resulting synthesis is presented in Section 5.

4 Analyzed Publications

In this section we summarize the nine analyzed publications.

(Fang et al., 2024) propose a 1-hop question answering system to integrate domain-specific knowledge using vector-based similarity for entity and relation matching. Based on a template, an LLM extracts a central entity and relation of a query which is matched to KG embeddings. The answer (target entity) is derived from the central entity via the central relation. (Luo et al., 2023) (Reasoning on Graphs) combine fine-tuned (for adapting to the KG and better utilizing the derived reasoning paths) LLMs and KGs in inference. For the retrieval, the LLM generates promising relation paths which are then instantiated based on a central entity extracted from the query. (Guo et al., 2024)(Knowledge-Navigator) navigate the KG, based on a central entity extracted from the query and semantically identical variations of the question, up to a predicted hop depth. In each step, top k relations are selected to follow. The selected triples are converted into natural language using a simple template and added as context to the prompt. (Sun et al., 2023) (Think-on-Graph) traverse the KG step by step starting from up to N entities extracted from the query. SPARQL is used to identify adjacent relations to the corresponing nodes in the KG. This process is iterated until the LLM can answer

Information	Example
Purpose of KG integration	Reduce hallucinations
Language models used	GPT-4, e5-base (Embedder)
Origin and structure of the KG	Freebase
Interaction between LLM inference and KG	1. Extract relevant entities 2. Search for entities in the KG
Evaluation methodology	Benchmarks: CWQ, WebQSP Metric: Exact-Match @1 Comparison: LLM-only, RAG
Results	Performs significantly better than...

Table 3: Exemplary extracted information from a paper on KG integration in LLMs

the question with the collected reasoning paths as context. (Kim et al., 2024) (Causal Reasoning) traverse the KG randomly starting from a certain KG node that is identified by semantic similarity to an additionally provided question concept. Collected reasoning paths are added as context to answer the question. (Zhu et al., 2024) (EMERGE) use LLMs and KGs to generate a summarized patient report from patient data in the form of structured time series and unstructured clinical notes. Therefore, a sophisticated extraction method of entities and relations from patient data including time series information is applied. Suitable context from the KG is retrieved by semantic similartiy. (Xu et al., 2024) (ChatTf) uses special KGs to answer questions about traditional Chinese folklore. An LLM extracts key folklore entities from the question. For each central entity, the semantically most similar folklore entity in the KG is determined. Then all triples in the KG that contain these entities are extracted. Triples are verbalized, ranked, and the best triples added as context. (Ye et al., 2024) (Correcting Factual Errors via Inference Paths) use KGs to detect and correct hallucinations in an LLM answer. Therefore, subquestions are derived and reasoning paths in the KG are tried to be found to prove the generated answer. Depending on the path's verdict, the answer is kept or corrected. (Kang et al., 2024) (Correcting Hallucination in Complaint LLM) use a special layered KG to provide the LLM with the necessary information to respond to complaints. For each question, a subgraph is created. This is extended by information from the KG and finally serves as context to answer the complaint.

5 Synthesis

5.1 Methods of Integrating KGs

Entry into the Knowledge Graph. In order to recognize patterns in the approaches, we first investigated which data is extracted from the input query and how this data is used to identify suitable entities in the KG as entry points. The results are shown in Tab. 4.

Most approaches start with the extraction of one or more entities from the input with an LLM. EMERGE is the only investigated approach that proposes an additional way for entity extraction without LLM. (Ye et al., 2024) uses an LLM to generate a naïve answer from which atomic facts and, in turn, sub-questions are generated. They form the basis for extracting the entities. (Kim et al., 2024) is the only approach that does not generate any initial entities but directly finds the node in the KG that has the highest semantic similarity to a provided question *concept*. Some approaches extract further information: (Fang et al., 2024) apply prompt engineering to extract a relation. (Luo et al., 2023) uses a fine-tuned LLM to extract a complete relation path from the question. (Guo et al., 2024) uses a special language model to estimate the number of hops required from the question and to generate semantically identical variants of the question.

It can happen that extracted entities do not appear verbatim in the KG. Most of the approaches ignore this problem, three approaches, however, use semantic similarity to match extracted entities with entities in the KG: (Fang et al., 2024), (Zhu et al., 2024) and (Xu et al., 2024). In (Fang et al., 2024), the principle of semantic similarity is also applied to the selection of an adjacency relation.

Approach	Extraction from Input	Entry into KG
(Fang et al., 2024)	Entity, Relation	Semantic similarity with central entity and relation
(Luo et al., 2023)	Entity, Relation paths	Directly via entity
(Guo et al., 2024)	Entity, Question variants, Number of hops	Directly via entity
(Sun et al., 2023)	Entities	Directly via entities
(Kim et al., 2024)	N/A	Semantic similarity with question concept
(Zhu et al., 2024)	Patient features, Diseases	Semantic similarity with extracted patient features and diseases
(Xu et al., 2024)	Entities	Semantic similarity with central entities
(Ye et al., 2024)	Two entities	Directly via one of the two entities
(Kang et al., 2024)	Entities	Directly via entities

Table 4: Overview of approaches to enter the KG based on input information

Querying the Knowledge Graph. Once the entry points have been defined, different methods to traverse the KG are proposed to collect knowledge that is made available to the LLM as context for generating the answer. The procedures of the approaches vary greatly (Tab. 5).

Three general approaches can be observed: First, (Fang et al., 2024), (Luo et al., 2023) and (Ye et al., 2024) apply a previously defined relation path directly to the entry node. This creates paths with specific instances. For example, the relation path "? $-$Party$\rightarrow$? $-$founded$\rightarrow$?" applied to the entity "Olaf Scholz" could lead to the reasoning path "Olaf Scholz $-$Party$\rightarrow$ SPD $-$founded$\rightarrow$ 1863". Second, KnowledgeNavigator (Guo et al., 2024) and Think-on-Graph (Sun et al., 2023) traverse the KG iteratively. Starting from the initial nodes, reasoning paths are created, which are gradually extended by relations and entities evaluated by an LLM. (Kang et al., 2024) iteratively add nodes to the subgraph representation of the problem. No LLM is used for this, but simple formulas for calculating information gain and importance of potential nodes. Third, CR (Kim et al., 2024) and ChatTf (Xu et al., 2024) consider all relations and entities adjacent to the entry node. CR then selects the best triple according to semantic similarity. ChatTf uses a special reranker language model to select the most relevant triples. EMERGE (Zhu et al., 2024) uses the entry nodes (can be disease, symptom or other feature) to identify related disease nodes in the KG. All adjacency relations and entities are extracted from these disease nodes.

The approaches are similar in providing the derived knowledge for the LLM. All approaches use prompt engineering to insert derived triples or reasoning paths as context for answering the query in the LLM prompt. An exception is (Fang et al., 2024), where the entity derived from the KG is directly output as answer. KN (Guo et al., 2024) and ChatTf (Xu et al., 2024) verbalize the triples. EMERGE (Zhu et al., 2024) uses a comprehensive prompt to generate a patient report.

The majority of the approaches are based on popular, publicly accessible KGs: Freebase (Bollacker et al., 2008) provides factual knowledge, collaboratively created by an online community. Discontinued in 2016 and migrated to WikiData. WikiData (Vrandečić and Krötzsch, 2014) provides comprehensive multilingual factual knowledge. Like other wiki projects, it is added to and updated collaboratively by users. ConceptNet (Speer et al., 2017) provides semantic relationships between words. Different sources and multilingual. PrimeKG (Chandak et al., 2023). provides a holistic view of 17080 diseases. Classification of entities and limitation to a few relations. Extracted from high quality medical sources. FB15k-237 (Toutanova et al., 2015) is a subgraph from Freebase.

Some approaches constructed their own domain-specific KG (Fang et al., 2024) parse source material to automatically construct a KG. The result is a KG with entities some of which consist of several sentences. ChatTf (Xu et al., 2024) defines a detailed schema "TFOnto" for modeling Chinese folklore as a KG. (Kang et al., 2024) use a four-layer KG generated from complaint texts and official information on competent authorities. KGs tend to have a simple structure. Some use classes (such as PrimeKG, TFOnto) or specify constraints for certain relations (e.g., WikiData), but none are based on formal, e.g., description logics.

Approach	Traversing the KG	Final Context
(Fang et al., 2024)	Relation	N/A
(Luo et al., 2023)	By relation path	Reasoning paths
(Guo et al., 2024)	Iterative selection of the most relevant relation up to the predicted hop depth	Verbalized triples
(Sun et al., 2023)	Iterative selection of the most relevant relation until LLM terminates	Reasoning paths
(Kim et al., 2024)	All adjacency relations	Reasoning paths
(Zhu et al., 2024)	Identification of disease from entry node, then all adjacency relations of diseases mentioned	Patient features, Diseases mentioned, Diseases found with definition, description, Info triplet on the disease
(Xu et al., 2024)	All adjacency relations	Verbalized triples
(Ye et al., 2024)	By relation path	Naive answer, Reasoning path
(Kang et al., 2024)	Iterative inclusion of entities with high information gain in subgraph	Classification, Subgraph

Table 5: Strategies for traversing the KG and construction of final context

5.2 Advantages of Integrating KGs

In addition to the mitigation of hallucinations, other problems of LLMs that are improved by the integration of KGs are mentioned in the reviewed publications (Tab. 6): *Reasoning*: Complex questions with multiple logical connections pose a challenge for LLMs. The structured representation of relationships in KGs can be used to simplify the modeling of complex questions as a chain of triples. *New domain-specific knowledge*: An external knowledge base such as a KG enables access to new knowledge without having to retrain the LLM. This enables state-of-the-art LLMs such as ChatGPT 4o from OpenAI to access up-to-date and domain-specific knowledge. *Explainability*: LLMs are black boxes. Their internal decision-making processes are difficult for humans to understand. The use of an external knowledge source that explicitly presents facts ensures the explainability of the answers.

Benchmarks. The examined publications use various benchmarks to evaluate the performance of their approaches. The respective results are shown in Tab. 7. Most benchmarks are so-called "Knowledge Base Question Answering" benchmarks (KBQA). They are used to evaluate systems that answer questions in natural language using a knowledge base. They specify the knowledge base, questions, expected answers and evaluation metrics. These include WebQuestions (WebQ) (Berant et al., 2013), WebQuestionsSP (WebQSP) (Yih et al., 2016), ComplexWebQuestions (CWQ) (Talmor and Berant, 2018), SimpleQuestions (SimpleQ) (Gu et al., 2021), 10th Question Answering over Linked Data Challenge (QALD10-en) (Usbeck et al., 2024), MetaQA (Zhang et al., 2018), and Mintaka (Sen et al., 2022).

ToG (Sun et al., 2023) also uses T-Rex (Elsahar et al., 2018) and Zero-Shot RE (Petroni et al., 2021) to quantify the performance of extracting relations from questions. In addition, the fact-checking performance is quantified with Creak (Onoe et al., 2021). (Kim et al., 2024) use CommonsenseQA (Talmor et al., 2019) as a benchmark. It is not based on a knowledge base, but is suitable for testing reasoning capacities.

Three studies created their own benchmarks to evaluate their approaches. In (Fang et al., 2024), test subjects were commissioned to formulate questions for a car handbook, from which the KG was generated. For ChatTf (Xu et al., 2024), questions were derived from official sources such as the "China Intangible Cultural Heritage" database and the "China Folklore Society" website. (Kang et al., 2024) derived a test dataset from official responses to complaints. The papers mainly use the following metrics, but do not describe in detail how they are derived from the outputs: *Exact match, Hits@1*: Percentage of outputs that exactly match the expected response (Ye et al., 2024), (Luo et al., 2023). (Sun et al., 2023) implies that the two metrics are used synonymously. *Acc@1*: Percentage of outputs that are correct, regardless of the output form (Kim et al., 2024).

Approach	Hallucinations	Reasoning	New Knowledge	Explainability
(Fang et al., 2024)	Yes	no	no	Yes
(Luo et al., 2023)	Yes	Yes	Yes	Yes
(Guo et al., 2024)	Yes	Yes	Yes	Yes
(Sun et al., 2023)	Yes	Yes	Yes	Yes
(Kim et al., 2024)	Yes	Yes	no	no
(Zhu et al., 2024)	Yes	no	Yes	Yes
(Xu et al., 2024)	Yes	no	Yes	no
(Ye et al., 2024)	Yes	no	no	no
(Kang et al., 2024)	Yes	no	Yes	no

Table 6: Functional aspects of the approaches w.r.t. hallucinations, reasoning, new knowledge, and explainability

The benchmark scores show that the integration of KGs improves the performance of LLMs for different types of questions. For KBQA-benchmarks, performance improvements range from 4% to 320%. It can be concluded that the use of explicit knowledge from KGs reduces the likelihood of hallucinations. Correctly answering complex questions proves that LLMs gain an improved understanding of complex questions by reasoning paths from KGs. ChatTf (Xu et al., 2024) and (Kang et al., 2024) show that knowledge of LLMs can be effectively extended by domain-specific knowledge through the integration of KGs. Only the approach (Fang et al., 2024) led to unsatisfactory results, which according to the authors is due to complex user-generated queries, a difficult use case (manual with similar information on different models) and domain-specific abbreviations.

5.3 Weaknesses and Limits

The following challenges with the integration of KGs into LLM inference can be concluded from the evaluation of the papers: *Incorrect traversal*: With iterative traversal of the KG, the LLM can have problems selecting the correct next relation in certain cases. One problem are complex questions that require a longer sub-graph as context for the LLM to answer the question correctly (Guo et al., 2024). The LLM has to select one relation after the other without knowing which other relations lie behind the one currently under consideration. Another problem are large, dense KGs such as WikiData, as the LLM has to evaluate hundreds of relations at once in the worst case when evaluating the adjacency relations of a node (Sun et al., 2023). *Complexity*: KG-supported LLM systems perform several LLM requests before the final response is generated. This increases the runtime and costs of the system, as each LLM request costs time and money (as energy consumption of powerful hardware or directly through API requests) (Guo et al., 2024), (Luo et al., 2023), (Sun et al., 2023). Comparison of the language models and retrieval procedures used reveals major differences in computational cost between the analyzed approaches (see Tab. 8). Lightweight approaches like (Fang et al., 2024) extract an entry entity and a relation path from the query and apply them directly to the graph. Computationally intensive approaches such as (Guo et al., 2024) use LLM agents to traverse the graph and expand adjacent relations and entities step-by-step.

6 Conclusion

In this paper, a systematic literature search was conducted on the integration of KGs into the inference processes of LLMs for mitigation of hallucinations. A systematic search on IEEE Xplore, ACM Digital Library and Google Scholar yielded 103 search results. By applying inclusion criteria and evaluating relevance with a scoring system, nine suitable papers were selected to answer the research questions. A data extraction scheme was used to extract relevant information from these papers in a standardized way.

General findings are summarized in the literature synthesis. One focus was on the collaboration between LLM and KG. Most approaches start with an entity extraction from the query that serve as entry points to the KG, some approaches use semantic similarity instead of exact match. The traversal of the KG starting from the entry node varies greatly from approach to approach. Almost all approaches use prompt engineering to provide the LLM with the extracted knowledge in the form of triples in a structured way. Most approaches use publicly

Approach	Benchmark	Metric	LLM	Performance
(Fang et al., 2024)	custom	Acc@1	GPT-3.5	34.3
(Luo et al., 2023)	CWQ	Hits@1	LLaMA 2 Chat (7B)	62.6 (+81%)
	WebQSP			85.7 (+33%)
(Guo et al., 2024)	WebQSP	Hits@1	GPT-3.5	82.3 (+35%)
	MetaQA (2H)			99.1 (+320%)
	MetaQA (3H)			95.0 (+220%)
(Sun et al., 2023)	CWQ	Hits@1	GPT-3.5	57.1 (+52%)
	WebQSP			76.2 (+20%)
	GrailQA			68.7 (+134%)
	QALD10-en			50.2 (+20%)
	SimpleQ			53.6 (+168%)
	WebQ			54.5 (+12%)
	T-REx			76.8 (+29%)
	Zero-Shot RE			88.0 (+218%)
	Creak			91.2 (+2%)
(Kim et al., 2024)	CQA	Acc@1	LLaMA 2 Chat	0.59 (+4%)
(Zhu et al., 2024)	MIMIC-III M	AUROC	Qwen (7B),	86.25
	MIMIC-III R		DeepSeek-V2 Chat	79.06
	MIMIC-IV M			89.50
	MIMIC-IV R			80.61
(Xu et al., 2024)	custom	Acc@1	GPT-3.5	0.91 (+81%)
(Ye et al., 2024)	CWQ	Exact-Match	GPT-3.5	64.0 (+68%)
	WebQSP			94.0 (+24%)
(Kang et al., 2024)	SimpleQ	Exact-Match	GPT-3.5	58.1 (+254%)
	Mintaka			53.9 (+131%)
	HotpotQA			27.3 (+34%)
	custom	Acc@1		0.85 (+47%)

Table 7: Performance improvements of approaches integrating KGs into LLMs across various benchmarks. Performance of the approaches is shown with relative improvement compared to baseline LLM performance in parentheses.

available general KGs, such as Freebase or WikiData. Some use domain-specific KGs (medicine) or constructed their own domain-specific KGs (car manual, Chinese folklore, complaints). In addition to mitigating hallucination, the papers cited further advantages of integrating KGs into LLM inference: improvement of reasoning capacities for complex questions, costeffective expansion of the knowledge base of LLMs and explainability of results. To prove the improved answer quality, mostly conventional KBQA benchmarks such as WebQuestionsSP or ComplexWebQuestions were used. Some approaches constructed their own test data sets manually or by interviewing test takers. The benchmark scores consistently show that the integration of KGs achieves a higher LLM answer quality, especially with regard to complex questions and specific facts. Disadvantages of integrating KGs were hardly described in the reviewed publications: Only the increased complexity and problems with LLM-based KG traversal for complex questions or entities with many relations were mentioned.

This review provides researchers and users with an overview of current approaches to integrating KGs into the LLM inference process for mitigating hallucinations. This area of research is currently developing rapidly. While these approaches mostly rely on relatively shallow traversal methods and semantic similarity, future research should explore more expressive and principled mechanisms to query KGs. This can include the translation of natural language queries into formal query languages such as SPARQL or Cypher, which could enable more precise access to the represented knowledge. Furthermore, deeper exploitation of the graph schema, e.g. property constraints, could be tried. Finally, ontological reasoning based on logical ax-

Approach	Model	Retrieval	Notes
(Fang et al., 2024)	*	*	GPT-3.5; entity-relation retrieval with template
(Luo et al., 2023)	**	*	finetuned LLaMA 2-7B; relation path extraction
(Guo et al., 2024)	**	***	GPT-3.5, pretrained LM; adjacent expansion
(Sun et al., 2023)	*	***	GPT-3.5; adjacent expansion
(Kim et al., 2024)	*	**	LLaMA 2 Chat; similar neighbors and random walk
(Zhu et al., 2024)	*	*	Qwen-7B; all neighbors of disease entities
(Xu et al., 2024)	**	**	GPT-3.5, finetuned reranker; ranking of all triples
(Ye et al., 2024)	**	**	GPT-3.5, policy network; paths between entities
(Kang et al., 2024)	*	***	GPT-3.5; query to subgraph, subgraph expansion

Table 8: Comparative analysis of computational costs of approaches integrating KGs into LLMs. More stars mean higher complexity because of the used language models (size, finetuning) or retrieval strategy. The valuation is based on the descriptions of the approaches in the referenced papers.

ioms (e.g., transitivity, subclass inference) could further improve inference quality, consistency, and explainability. We advocate for integrating LLMs with symbolic reasoners for a more principled differentiation between LLM as language interface and structured knowledge bases and reasoners as knowledge sources to developing reliable systems with better and more explicit explainability. Additionally, future research could focus on exploring automated KG construction from domain-specific corpora, optimizing task-specific prompting strategies that utilize KG context (Prompt Engineering) and developing continual learning frameworks that allow LLMs to adapt to evolving KGs without retraining. These directions will help guide the next generation of intelligent, knowledge-aware AI systems.

References

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3947–3960, Mexico City, Mexico. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA. Association for Computing Machinery.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yunfei Fang, Yong Chen, Zhonglin Jiang, Jun Xiao, and Yanli Ge. 2024. Effective and Reliable Domain-Specific Knowledge Question Answering. In *2024 IEEE International Conference on E-Business Engineering (ICEBE)*, pages 238–243.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond I.I.D.: Three Levels of Generalization for Question Answering on Knowledge Bases. In *Proceedings of the Web Conference 2021*, WWW '21, pages 3477–3488, New York, NY, USA. Association for Computing Machinery.

Tiezheng Guo, Qingwen Yang, Chen Wang, Yanyi Liu, Pan Li, Jiawei Tang, Dapeng Li, and Yingyou Wen. 2024. KnowledgeNavigator: Leveraging large language models for enhanced reasoning over knowledge graph. *Complex & Intelligent Systems*, 10(5):7063–7076.

Muhammad Usman Hadi, Qasem Al Tashi, Rizwan Qureshi, Abbas Shah, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, and Seyedali Mirjalili. 2023. A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. *TechRxiv*.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez,

Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge Graphs. *ACM Comput. Surv.*, 54(4):71:1–71:37.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*

Nourhan Ibrahim, 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).

Jiaju Kang, Weichao Pan, Tian Zhang, Ziming Wang, Shuqin Yang, Zhiqin Wang, Jian Wang, and Xiaofei Niu. 2024. Correcting Factuality Hallucination in Complaint Large Language Model via Entity-Augmented. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Staffs Keele and 1 others. 2007. Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, ver. 2.3 ebse technical report. ebse.

Yejin Kim, Eojin Kang, Juae Kim, and H. Howie Huang. 2024. Causal Reasoning in Large Language Models: A Knowledge Graph Approach. In *Causality and Large Models @NeurIPS 2024*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023. Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning. In *The Twelfth International Conference on Learning Representations*.

Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi, and Greg Durrett. 2021. CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: A Benchmark for Knowledge Intensive Language Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A Complex, Natural, and Multilingual Dataset for End-to-End Question Answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. In *The Twelfth International Conference on Learning Representations*.

Alon Talmor and Jonathan Berant. 2018. The Web as a Knowledge-Base for Answering Complex Questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing Text for Joint Embedding of Text and Knowledge Bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Usbeck, Xi Yan, Aleksandr Perevalov, Longquan Jiang, Julius Schulz, Angelie Kraft, Cedric Möller, Junbo Huang, Jan Reineke, Axel-Cyrille Ngonga Ngomo, Muhammad Saleem, and Andreas Both. 2024. QALD-10 – The 10th challenge on question answering over linked data: Shifting from DBpedia to Wikidata as a KG for KGQA. *Semantic Web*, 15(6):2193–2207.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Zichong Wang, Zhibo Chu, Thang Viet Doan, Shiwen Ni, Min Yang, and Wenbin Zhang. 2024. History, development, and principles of large language models: An introductory survey. *AI and Ethics*.

Jun Xu, Hao Zhang, Haijing Zhang, Jiawei Lu, and Gang Xiao. 2024. ChatTf: A Knowledge Graph-Enhanced Intelligent Q&A System for Mitigating Factuality Hallucinations in Traditional Folklore. *IEEE Access*, 12:162638–162650.

Weiqi Ye, Qiang Zhang, Xian Zhou, Wenpeng Hu, Changhai Tian, and Jiajun Cheng. 2024. Correcting Factual Errors in LLMs via Inference Paths Based on Knowledge Graph. In *2024 International Conference on Computational Linguistics and Natural Language Processing (CLNLP)*, pages 12–16.

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. 2016. The Value of Semantic Parse Labeling for Knowledge Base Question Answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206, Berlin, Germany. Association for Computational Linguistics.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander Smola, and Le Song. 2018. Variational Reasoning for Question Answering With Knowledge Graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2024. A Survey of Large Language Models. *Preprint*, arXiv:2303.18223.

Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. 2024. EMERGE: Enhancing Multimodal Electronic Health Records Predictive Modeling with Retrieval-Augmented Generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pages 3549–3559, New York, NY, USA. Association for Computing Machinery.