REDUCING LLM HALLUCINATIONS WITH KNOWLEDGE GRAPHS: A SURVEY

Tommaso Dolci, Milos Jovanovik, and Katja Hose
Institute of Logic and Computation
TU Wien
Vienna, Austria
{tommaso.dolci, milos.jovanovik, katja.hose}@tuwien.ac.at

ABSTRACT

Large language models (LLMs) have achieved remarkable performance in natural language processing tasks, yet they still suffer from hallucinations, i.e., the generation of factually incorrect, inconsistent, or nonsensical information. Hallucinations undermine trust and limit the applicability of LLM-based systems especially in high-risk scenarios, such as personal healthcare or legal support. To address this issue, knowledge graphs (KGs) have increasingly been adopted as external sources of trustworthy knowledge for detecting and mitigating hallucinations throughout the LLM lifecycle. In recent years, the number of proposed approaches for reducing hallucinations with KG support have rapidly increased, generating a vast landscape of solutions, emerging trends, and opportunities for future research. In this survey, we present a literature review on reducing LLM hallucination with KGs, presenting a comprehensive taxonomy to organize existing methods according to the stage at which the KG intervention occurs: during model training, at inference time, or after generation. We discuss approaches for hallucination detection and mitigation, as well as KG-based benchmarks for evaluating and assessing LLM hallucinations. Finally, we highlight open challenges (e.g., KG incompleteness, cultural and language coverage, efficiency) and future research directions towards the development of more reliable, explainable, and factuality-aware KG-supported LLMs.

Keywords: Hallucination · Large Language Model · Knowledge Graph · Factuality · Survey

1 Introduction

Large language models (LLMs), such as GPT-4 [122], Gemini [158, 159], and LLaMA [165, 166], achieve state-of-the-art performance in natural language processing across a wide range of tasks, including machine translation [204], information retrieval [207], code generation [76], and others [12, 21]. However, despite these advances, LLMs still suffer from *hallucinations*, i.e., the generation of content that is factually incorrect, internally inconsistent, or entirely nonsensical, yet plausible and expressed with a confident tone [70, 199]. For instance, when asked about Einstein's Nobel Prize in 1921, an LLM may confidently state that the prize was awarded for the theory of relativity, while in fact it was awarded for Einstein's study on the photoelectric effect. Figure 1 provides an overview of the main types of LLM hallucinations, distinguishing between factuality hallucinations (i.e., inaccuracies regarding external world knowledge [70]) and faithfulness hallucinations (i.e., deviations from the internal contextual knowledge [70]).

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 reliability and safety, especially when these systems are treated as authoritative sources of knowledge [74]. Hallucinations favor misinformation, reduce trustworthiness, and severely limit the deployment of LLM-based systems in high-risk scenarios, such as personal healthcare or legal support [9]. For instance, a model that incorrectly suggests a drug dosage can directly harm the safety of individuals. The recent introduction of retrieval-augmented generation [96] sparked hope about reducing

[1] https://openai.com/index/chatgpt
[2] https://www.anthropic.com/news/introducing-claude

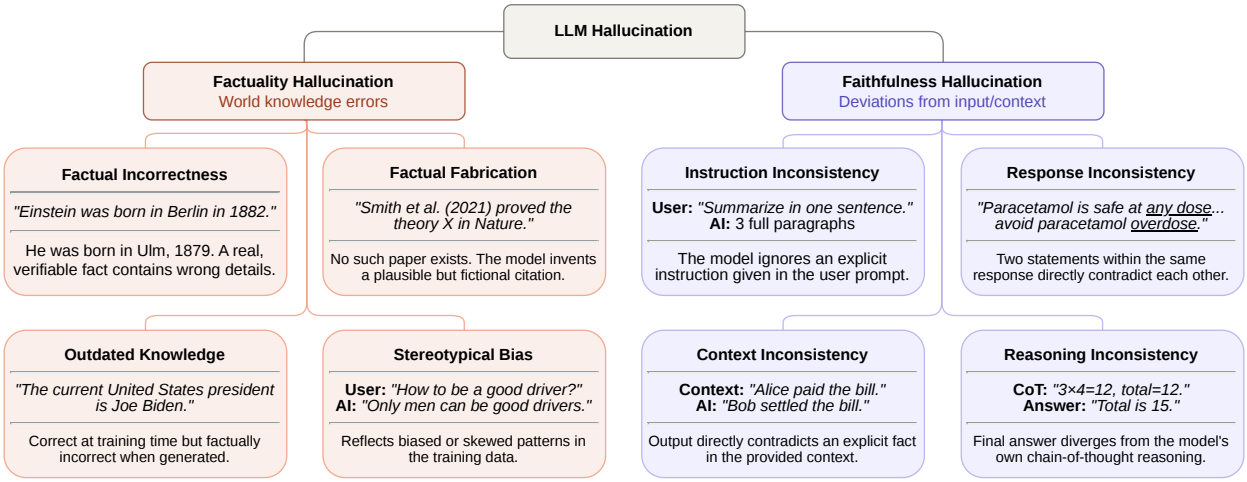

Figure 1: Overview of the main types of LLM hallucination.

hallucinations by grounding responses in external sources (e.g., web pages). However, web-based retrieval-augmented systems – such as the recently-introduced Google AI Overviews[3] – proved that hallucinations are still an open problem, as exemplified by a number of harmful suggestions, such as adding glue to pizza or eating a small rock every day [144].

In response to these challenges, numerous recent studies investigated the integration of LLMs with knowledge graphs (KGs), serving as external sources of structured and trustworthy knowledge [3, 89]. KGs represent factual knowledge as semantic triples of the form *(subject, relation, object)*, and contain machine-readable representations of real-world entities and relationships [64] curated from trustworthy sources (e.g., Wikipedia). Different types of KGs can provide different types of supporting knowledge, including encyclopedic (e.g., Wikidata [172], DBpedia [8], and YAGO [153]), commonsense (e.g., ConceptNet [151], ATOMIC [143])), and domain-specific KGs (e.g., UMLS with biomedical knowledge [17]).

In recent years, the number of proposed approaches has substantially increased, generating a vast landscape of solutions, emerging trends, and research opportunities [3, 89]. In this paper, we propose a comprehensive discussion of KG-supported approaches for detecting and mitigating LLM hallucinations. To systematically classify and compare existing solutions, we introduce an updated taxonomy of approaches inspired by previous literature [89] and grounded in the stages of the LLM lifecycle: *pre-generation* (i.e., at training-time), *in-generation* (i.e., during inference), *post-generation* (i.e., verification and mitigation of the LLM output), and *evaluation benchmarks*. To systematize our review and guarantee the reproducibility of the results, we follow a structured survey methodology based on the PRISMA guidelines [126].

In the second part of this paper, we discuss open challenges in reducing LLM hallucinations with KGs, including the reconciliation of approaches for mitigating factuality and faithfulness hallucinations, the problem of overcoming KG incompleteness, and the challenge of scalability with respect to time and cost efficiency. Moreover, we introduce a range of research directions to encourage future papers in this area, e.g., the coordination of multiple KG sources, the role of cultural bias and linguistic variety in hallucination detection, the importance of interpretability techniques to explain factual errors, and the necessity to shift LLM training goals to prioritize factual accuracy [80]. Our work is motivated by the following research questions:

RQ1: What are the current KG-supported approaches to reduce hallucinations in LLMs, and what are their strengths and limitations?

RQ2: How can KG-supported approaches for detecting and mitigating LLM hallucinations be systematically classified across the LLM lifecycle stages?

RQ3: What are open challenges and future research directions in reducing LLM hallucinations with KG support?

In response to these questions, this paper offers the following contributions: *i)* a comprehensive review of KG-supported approaches for reducing LLM hallucinations, conducted following a structured survey methodology based on the PRISMA guidelines [126]; *ii)* an updated taxonomy to classify existing approaches across the LLM lifecycle

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

Table 1: Comparison of related survey papers on reducing LLM hallucinations with KG support. Legend: LLM = Large Langue Models, PLM = Pre-trained Language Models. The symbol ◇ indicates partial content coverage.

Survey	Detection	Mitigation	Taxonomy	Reproducible	Scope	Year
Agrawal et al. [3]	✓	✓	✓	✗	PLM, LLM	2024
Yang et al. [190]	✗	✓	◇	✗	PLM	2024
Lavrinovics et al. [89]	◇	✓	◇	✗	LLM	2025
Wagner et al. [173]	✗	✓	✗	✓	LLM	2025
This Survey	✓	✓	✓	✓	LLM	2026

stages of pre-generation, in-generation, post-generation, and evaluation; *iii)* a discussion of open challenges and future research directions at the intersection of KGs and LLM hallucination reduction. To the best of our knowledge, this is the first survey that systematically classifies recent advances in reducing hallucinations with KGs, following a structured methodological approach to review the literature.

Related Surveys Previous surveys focused on the general interplay of LLMs and KGs [72, 127, 128], on specific KG-based techniques (e.g., retrieval-augmented generation from graphs [58, 130]), or specific natural language tasks supported by KGs (e.g., knowledge graph question-answering [112]). Several works presented a comprehensive and general overview of LLM hallucinations and mitigation methods [70, 74, 199]. However, only few survey papers have focused specifically on the role of KGs in mitigating LLM hallucinations. Two papers [3, 190] present an overview on reducing hallucinations with KG support. Both have been published in early 2024, primarily focusing on smaller pre-trained language models (e.g., BERT [36]), rather than today's multi-billion-parameter LLMs (e.g., GPT-4 [122]). Another recent survey adopts a systematic approach to survey the literature, but limiting the analysis to KG-based interventions during LLM inference, rather than the entire LLM lifecycle [173]. Finally, [89] provides a clear and detailed discussion on KG-based mitigation solutions, although lacking both a systematic methodological framework and a comprehensive taxonomy of the discussed approaches. Table 1 provides a comparative overview of the related surveys.

Paper Outline The rest of the paper is organized as follows. Section 2 provides an overview of background definitions on LLMs, KGs, and hallucinations. Section 3 presents the survey methodology adopted. Section 4 introduces our new taxonomy and reviews the selected KG-supported solutions for reducing LLM hallucinations. Section 5 discusses the main open challenges and future directions in the research area. Finally, Section 6 concludes the paper.

2 Preliminaries

In this section, we introduce important preliminary notions and terminology about LLMs, knowledge graphs, and hallucinations.

2.1 Large Language Models

Pre-trained language models (PLMs) are deep-learning models trained on large unlabeled corpora to learn general language representations, primarily based on the transformer architecture [170]. Depending on their architecture, PLMs are commonly categorized into: *i)* encoder-only models for representation learning and discriminative tasks, for general purpose and specific domains (e.g., BERT [36], RoBERTa [105], BioBERT [92]), *ii)* decoder-only models, or generative models, specialized for text generation (e.g., GPT [135], Llama [165], Gemini [158]), and *iii)* encoder-decoder models for sequence-to-sequence learning (e.g., T5 [136], BART [95]). Generative PLMs can be further categorized into *large* language models (LLMs) and *small* language models (SLMs) according to their size. LLMs contain billions of parameters and show strong general-purpose capabilities, whereas SLMs contain less parameters to prioritize computational efficiency [14]. In this paper, *PLM* refers to encoder-only models (e.g., BERT), while *LLM* refers to large-scale generative decoders. The knowledge acquired by LLMs is commonly referred to as *parametric*, *implicit*, or *internal knowledge*, as it is encoded within the parameters rather than through explicit symbolic representations. LLMs are trained on large-scale multilingual web corpora, encyclopedic resources, and code repositories [102]. However, these data often contain copyright issues [88], social bias [119], and factual inaccuracies [70], raising concerns about LLM trustworthiness and security. These issues are a main driver for enhancing LLMs with KGs, which contain more curated and trustworthy knowledge [190].

A fundamental characteristic of modern LLMs is their ability to solve downstream tasks through prompting rather than task-specific fine-tuning [20]. LLMs exhibit so-called *emergent abilities* that appear only in sufficiently large models [179], including instruction following [125, 129], in-context learning [39], and chain-of-thought (CoT) reasoning [180]. In-context learning enables models to infer tasks directly from prompt without parameter updates, e.g., by few-shot [20] and zero-shot learning [83], while CoT fosters multi-step reasoning [180]. Variations of CoT, such as self-consistent CoT [178], tree-of-thought [191] and graph-of-thought [192], extend reasoning via multiple candidate paths, branching, and graph-structures.

2.2 Knowledge Graphs

While the term *knowledge graph* (KG) dates back to at least the 1970s, its modern usage stems from Google's 2012 announcement of its Knowledge Graph [150], after which the concept was rapidly adopted across both industry and academia. Despite its widespread adoption, there is still no single, universally accepted definition, with interpretations ranging from narrow technical descriptions to broader conceptual definitions. In this work, we adopt the inclusive definition of Hogan et al. [64], who describe a KG as a graph of data intended to accumulate and convey knowledge of the real world, in which nodes denote entities of interest and edges denote relations between those entities. The underlying *data graph* conforms to a graph-based data model, most commonly a directed edge-labeled graph (as in RDF) or a property graph, and is typically enriched with representations of schema, identity, and context, together with ontologies or rules that allow further knowledge to be entailed deductively or extracted inductively.

A wide range of KGs has emerged in practice, which can be grouped by their scope. *Encyclopedic* KGs aim for broad, cross-domain coverage and are often derived from collaborative or semi-structured sources; prominent examples include DBpedia [93] and YAGO [153], both extracted largely from Wikipedia, Wikidata [172], built and curated by a community of volunteers, and Freebase [18], an early collaborative effort whose contents were ultimately migrated into Wikidata. *Commonsense* KGs instead encode the implicit, everyday knowledge that humans take for granted, such as ConceptNet [151], ATOMIC [143], and the long-standing Cyc project [94]. Finally, *domain-specific* KGs capture specialized knowledge within a particular field, such as biomedicine (e.g., UMLS [17], SNOMED CT [40], the Gene Ontology [7], Hetionet [63]), geography (e.g., GeoNames [53], UrbanKG [103]), and scholarly works (e.g., OpenAlex [133]). For the remainder of this paper, the knowledge encoded in KGs is referred to as *symbolic*, *explicit*, or *external* knowledge, in contrast to the knowledge implicitly captured within the parameters of LLMs.

2.3 Hallucinations

Hallucination is widely recognized as a fundamental limitation of LLMs [9]. Hallucinations are defined as the generation of outputs that are linguistically plausible but factually incorrect, internally inconsistent, or entirely nonsensical [74, 199]. In this paper, we distinguish between *factuality hallucinations* and *faithfulness hallucinations* (see Figure 1). Factuality hallucinations [70], also referred to as *extrinsic* or *open-domain* hallucinations, cannot be verified against any content internal to the inference process (e.g., the prompt), thus requiring external world knowledge for identification, e.g., datasets, web pages, or knowledge bases. For example, the claim *Marie Curie was born in France* cannot be assessed from the prompt alone, requiring external knowledge sources to identify it as incorrect. Faithfulness hallucinations [70], also referred to as *intrinsic* or *closed-domain* hallucinations, directly contradict the input content (e.g., instructions, prompt) or the internal reasoning of the LLM. For example, in a summarization task, if the source document states that a study involved 200 participants, but the generated summary reports 2000, the error is directly verifiable by comparison with the input. While factuality hallucinations are a phenomenon shared across multiple tasks, faithfulness hallucinations are particularly prominent in specific downstream tasks (e.g., summarization) or in LLM-specific frameworks such as RAG [96], where the response may contradict with the retrieved knowledge, or CoT [180], where successive reasoning steps may contradict one another.

Finally, hallucinations are frequently associated with the notion of *consistency*, such as logic-related hallucinations [54]. Consistency is defined as the invariance of a model's output under semantically equivalent inputs [44, 54]. Due to the stochastic nature of auto-regressive generation, inconsistent outputs can indicate the presence of hallucinations. Their occurrence is further increased by the complexity of maintaining up-to-date knowledge in LLMs, leading to factuality hallucinations from outdated parametric knowledge (see Figure 1). Addressing this issue with instruction training requires costly high-quality data [23], while continuous fine-tuning risks catastrophic forgetting [139]. These limitations highlight the need to re-direct the LLM training goals towards factual accuracy [80].

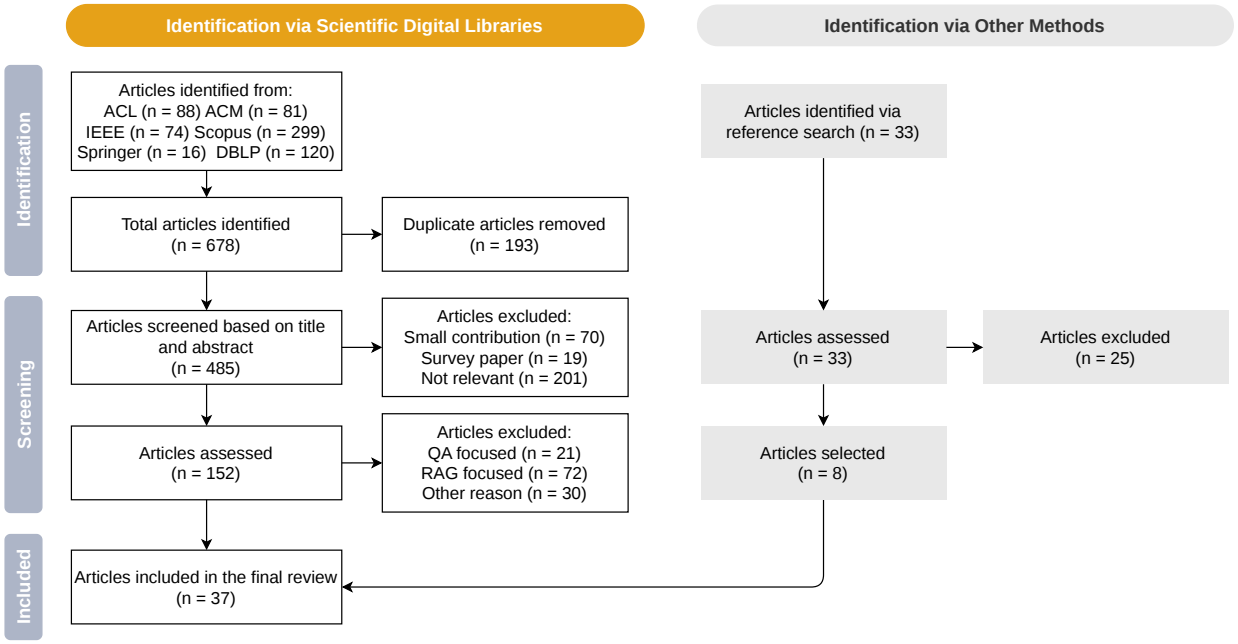

Figure 2: Flow diagram of the adopted survey methodology.

3 Review Methodology

The survey follows the PRISMA methodology guidelines [126].[4] We identified relevant studies from six scientific digital libraries, namely ACL Anthology, ACM Digital Library, IEEE Xplore, Scopus, SpringerLink, and DBLP. Table 2 summarizes the search strings, the search parameters, and the number of retrieved articles for each digital library. The overall selection process is illustrated in the PRISMA flow diagram in Figure 2. To identify relevant literature, we designed a search strategy based on combinations of keywords at the intersection of LLMs, hallucinations, and KGs. Search was performed within the abstracts when supported by the library, and within the title otherwise (i.e., for SpringerLink and DBLP). After removing duplicate records, we screened the retrieved articles according to the following inclusion and exclusion criteria:

C1: **Relevance**: The article addresses the problem of reducing LLM hallucinations with the support of KGs, including hallucination detection, mitigation, or evaluation techniques (e.g., benchmarks). Solutions that focus on LLMs to augment KGs are excluded.

C2: **Scientific Quality**: Only peer-reviewed papers formally accepted for publication are included, thus pre-prints and unpublished manuscripts are excluded.

C3: **Accessibility**: Only articles written in English are considered.

C4: **Recency**: Only papers published from 2020 onwards are included, to provide an up-to-date overview of current approaches in the literature.

C5: **Content**: Papers with limited technical contributions, such as demo papers, tutorials, and surveys, are excluded.

For each paper, we extracted information regarding the target LLMs, KGs used as support, KG integration methods, evaluation datasets, reported metrics, and more. Preliminary analysis of the identified literature[5] highlights a growing interest in this research area. The number of publications increased significantly in 2024 and continued to grow steadily in 2025, as shown in Figure 3.

Table 2: Summary of the scientific libraries included in the literature search.

Library	Field	Source	#Articles	Search String
ACL	Abstract	ACL Anthology list	88	(language model* OR llm*) AND (knowledge graph* OR kg*) AND hallucinat*
ACM	Abstract	Web search interface	81	(language model* OR llm*) AND (knowledge graph* OR kg*) AND hallucinat*
IEEE	Abstract	Web search interface	74	(language model* OR llm*) AND (knowledge graph* OR kg*) AND hallucinat*
Scopus	Abstract	API call	299	(language model* OR llm*) AND (knowledge graph* OR kg*) AND hallucinat*
Springer	Title	Web search interface	16	(language model* OR llm* OR encoder* OR decoder* OR gpt OR rag) AND (knowledge* OR graph* OR kg*) AND (hallucinat* OR fact* OR qa OR trust* OR coheren* OR reliab* OR explain*)
DBLP	Title	API call	120	(language model* OR llm* OR encoder* OR decoder* OR gpt OR rag) AND (knowledge* OR graph* OR kg*) AND (hallucinat* OR fact* OR qa OR trust* OR coheren* OR reliab* OR explain*)

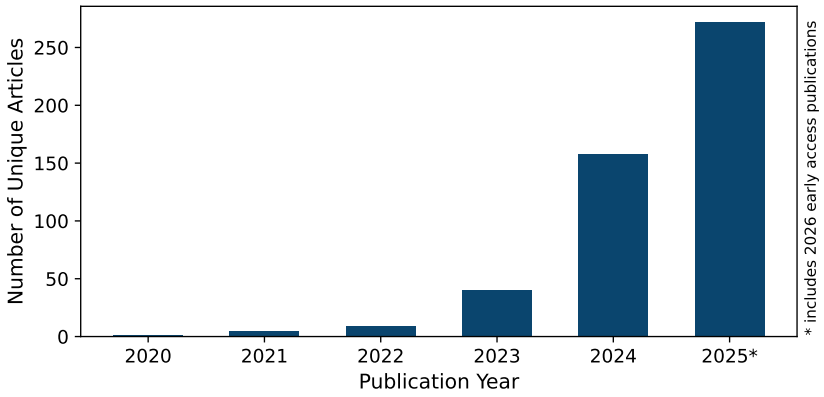

Figure 3: Distribution of unique articles extracted from the scientific libraries.

4 Reducing LLM Hallucinations with KGs

This section discusses the articles selected during the literature review. We classify the considered approaches into four categories, depending on the stage of the LLM lifecycle they address: pre-generation, in-generation, post-generation, and evaluation benchmarks. Pre-generation methods tackle hallucination before inference, e.g., during fine-tuning of the underlying model. In-generation methods intervene at runtime during the LLM inference. Post-generation methods comprise post-hoc interventions based on the LLM output. Finally, we discuss KG-derived benchmarks for detecting and measuring hallucinations. Figure 4 displays the overall taxonomy.

4.1 Pre-Generation Methods

LLM training typically consists of two stages: pre-training and fine-tuning. During pre-training the model acquires general-purpose competences by learning world knowledge and linguistic structures from massive text corpora. During fine-tuning it learns task-specific skills from smaller curated datasets (e.g., textual entailment [19] or instruction following [125]). While methods to integrate KG knowledge into language models during pre-training have been studied in early pre-trained language models [98, 148, 177, 186, 200], these solutions lack emphasis on hallucinations, which is instead a prevalent phenomenon in larger generative models due to their open-ended abilities [70]. Moreover, due to the large amount of data needed for pre-training LLMs, and the limited amount of knowledge contained in KGs, pre-training LLMs with KGs is not feasible for models with billions of parameters. On the other hand, we evidence the use of KGs for reducing hallucination during LLM *fine-tuning* (Section 4.1.1). Additionally, we describe current *probing methods* (Section 4.1.2) to identify knowledge areas more prone to hallucinations, a viable solution to guide the training objectives and consolidating the training data before the actual training process.

[4]For reproducibility purposes, the scripts used to query the digital libraries, the search strings, and the intermediate results are publicly available at: `https://github.com/dmki-tuwien/Hallucination-KG-Survey`

[5]Results as of December 2025.

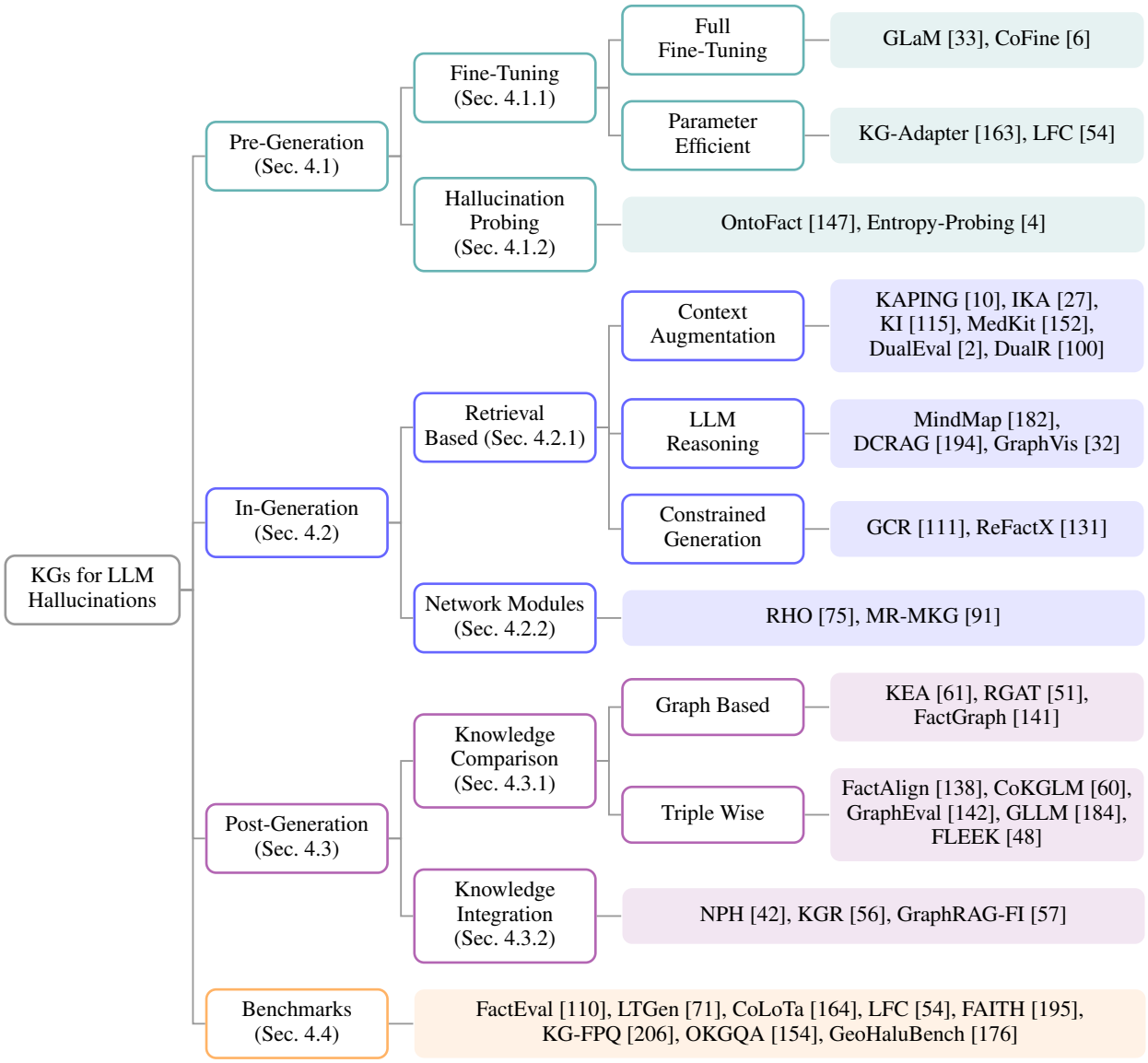

Figure 4: Taxonomy of KG-based approaches for reducing LLM hallucinations

4.1.1 Fine-Tuning

Fine-tuning methods can be categorized into full and parameter-efficient. Whereas *full fine-tuning* methods update the entire set of model parameters, *parameter-efficient fine-tuning* (PEFT) methods modify only a small fraction of the total parameters, keeping the majority of the pre-trained parameters frozen. This property makes them particularly suitable for further training very large models [37]. PEFT approaches can be categorized into additive and selective methods [59]. Additive fine-tuning introduces additional trainable components, e.g., adapter layers,[6] into the model architecture [66]. Selective fine-tuning updates only a limited subset of the original model parameters. Reparameterization (e.g., LoRA [67], QLoRA [35]) constitutes a distinct category of PEFT methods, sharing similarity with additive methods since the trainable parameters are external to the original pre-trained parameters. Figure 5 displays for an overview of KG-based fine-tuning methods.

[6]Adapters can also be interpreted as external network modules (see Section 4.2.2) due to their portability: once trained, they can be integrated into different architectures without additional modifications [66]. Nonetheless, for the sake of clarity, they are discussed in this section together with other fine-tuning methods.

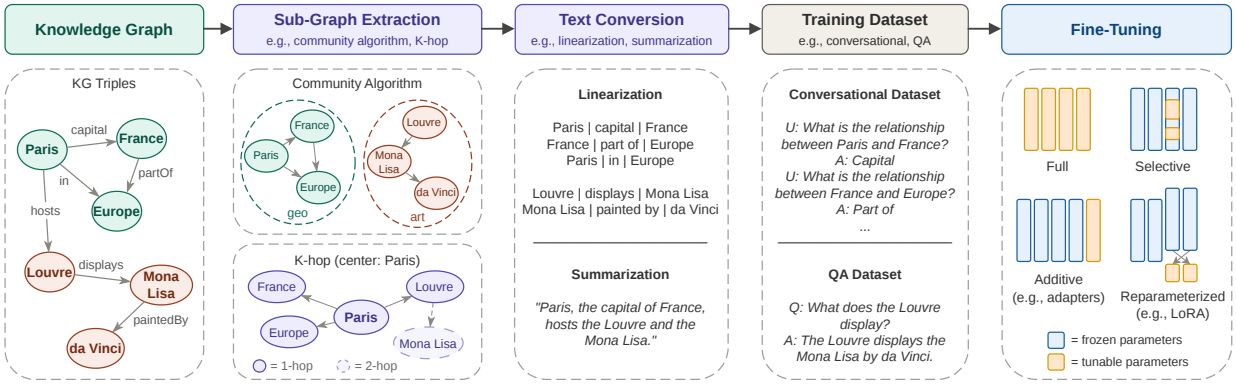

Figure 5: Overview of KG-based fine-tuning methods, illustrating the knowledge extraction and dataset creation approaches from [33] and [6].

Full Fine-Tuning GLaM [33] fine-tunes LLMs from K-hop neighborhood sub-graphs centered on KG entities, encoded into text with strategies such as triple linearization [49]. After an intermediate step of text summarization to improve semantic alignment and increase linguistic variety, the authors construct a training dataset with question-answer pairs in both open-ended and multiple-choice formats. GLaM adopts a supervised fine-tuning strategy yielding enhanced factual accuracy and multi-hop reasoning, attributed to more effective learning of relational and reasoning paths from the source KG. CoFine [6] adopt a different strategy for constructing the training dataset, dividing the KG into communities [16] before fine-tuning on conversational data by instruction-tuning. Dividing the KG into communities improves local and global knowledge balance compared to K-hop strategies [33] and ontology-based approaches [193], reducing hallucinations in downstream tasks such as link-prediction.

Parameter-Efficient Fine-Tuning Building on previous additive fine-tuning to enhance LLMs with data extracted from KGs [65, 175], KG-Adapter [163] trains adapter layers to increase factual accuracy in question-answering (QA) tasks. Training adapter layers allows direct access to the KG structure, overcoming common limitations of retrieval-based methods, such as loss of structural information and conflicts between internal and external knowledge. KG-Adapter tunes a fraction of the total model parameters (28 millions out of 7 billions), improving response accuracy with a more lightweight approach compared to full fine-tuning methods. LFC [54] expands research on KG-based PEFT by reparameterizing LLMs with QLoRA [35] to reduce logical hallucinations. LFC frames logical consistency as the task of correctly classifying KG-extracted facts under logical permutations (e.g., negation) and combinations of multiple logical operators. For instance, if a fact f_1 is true and f_2 is false, then $\neg f_1$ should be consistently classified as false, $f_1 \wedge f_2$ as false, and $f_1 \vee (f_1 \wedge f_2)$ as true. LFC is shown to achieve higher factual accuracy than naive retrieval-based approaches, generalizing to unseen logical combinations and commutative rules.

4.1.2 Hallucination Probing

Probing approaches analyze the LLM's internal representations to understand how they encode different input data. In the context of hallucinations, the goal is to identify areas of knowledge associated to factual inaccuracies. OntoFact [147] maps LLM factual inaccuracies to ontology concepts to identify overlooked domains where models are particularly error-prone. It uses a combination of encyclopedic and domain-specific KGs as a backbone to generate yes/no questions, based on transforming ontology-level triples, e.g., *(Person, birthPlace, City)*, into corresponding instance-level triples, e.g., *(LeBron James, birthPlace, Akron)*, then filling question templates, e.g., *Was [Person] born in [City]?* A reinforcement learning mechanism navigates the KG to identify triples that are more likely to elicit incorrect answers. To reduce false positives and reduce LLM overconfidence, each probe includes counterfactual variants and questions with missing entities (e.g., *Was [Person] born in N/A?*). Instead of dynamically exploring the boundary of LLM knowledge, Entropy-Probing [4] measures static entropy features on KG sub-graphs from Wikidata [172] as proxies for hallucination risk. The study confirms the intuition that the accuracy of LLM answers and entropy-related KG features – e.g., property or entity entropy – are correlated, i.e., higher accuracy is associated to higher fact-dense regions of KGs. This result suggests to reinforce knowledge areas with high entropy to mitigate hallucinations. Figure 6 displays the workflow of hallucination probing methods.

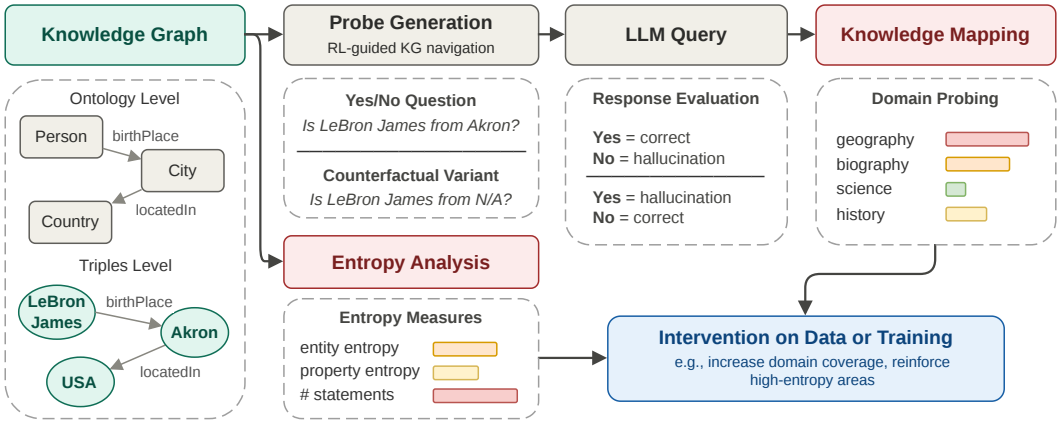

Figure 6: Overview of hallucination probing approaches, describing [147] and [4].

4.1.3 Discussion

Pre-generation approaches modify the internal knowledge representation without additional intervention at runtime, thus maintaining efficiency during inference especially when confronted with retrieval-based methods [6, 33]. While further training LLMs is computationally expensive, particular KG interventions (e.g., community partitioning [6]) and PEFT methods [54, 163] reduce computational and memory costs. Additive fine-tuning allows portability and re-usability by treating adapters as interchangeable layers of the LLM architecture [66], whereas full and selective fine-tuning requires distinct training for each model. However, additive fine-tuning achieves higher performance with medium-sized models (3 to 13 billion parameters), with lower impact on smaller and larger ones [163]. At the same time, performance in reducing hallucination by PEFT is shown to increase with a higher number of tunable parameters, opening a trade-off question between efficiency and hallucination reduction [163]. Although probing methods to identify domains of knowledge that are more prone to hallucinations can improve the creation of KG-derived training datasets [147], fine-tuning approaches are typically task- and domain-specific [6], thus requiring further training to adapt to open-domain scenarios. Moreover, re-training the model with updated data to reduce hallucination from outdated knowledge is expensive and inflexible [190]. Transparency remains an open challenge in pre-generation methods, as the internal representations learned by LLMs are still not well understood and require costly interpretability methods (e.g., mechanistic interpretability) for their analysis [201].

4.2 In-Generation Methods

In-generation methods address LLM hallucination during inference time. They can be divided into two categories: *retrieval-based* approaches, which guide the LLM generation with external KG information retrieved according to the LLM input, and *network-module* approaches, which inject KG knowledge through additional neural components (e.g., GNNs [185], GATs [171]).

4.2.1 Retrieval Based

Retrieval-based methods combine LLM implicit knowledge with explicit knowledge from KGs. They can be categorized according to how the retrieved KG information influences the LLM inference: by *context augmentation*, *LLM reasoning*, or *constrained generation*. Figure 7 provides an overview of retrieval-based methods.

Context Augmentation Context augmentation integrates external knowledge into the LLM input context, supporting factual generation without modifying the model architecture or the internal parameters [101]. Building on the principles of in-context learning [39], this approach is sometimes referred to as *context injection* or *prompt augmentation*. Contrary to RAG approaches [96], the goal is to align and integrate the LLM knowledge with external factual knowledge, rather than restricting the generation to the sole retrieved information. KAPING [10] pioneered context augmentation to support the generation of factually accurate responses by KG triples injection. After matching entities in the original prompt to entities in external encyclopedic KGs (e.g., Wikidata [172], Freebase [18]), KAPING retrieves the most relevant top-K triples, converts them into natural language by linearization, and appends them to the initial prompt. IKA [27] extends context augmentation to dialogue generation tasks, introducing a KG-grounded framework to increase both factuality and faithfulness. Alongside relevant KG-extracted facts, the approach selects

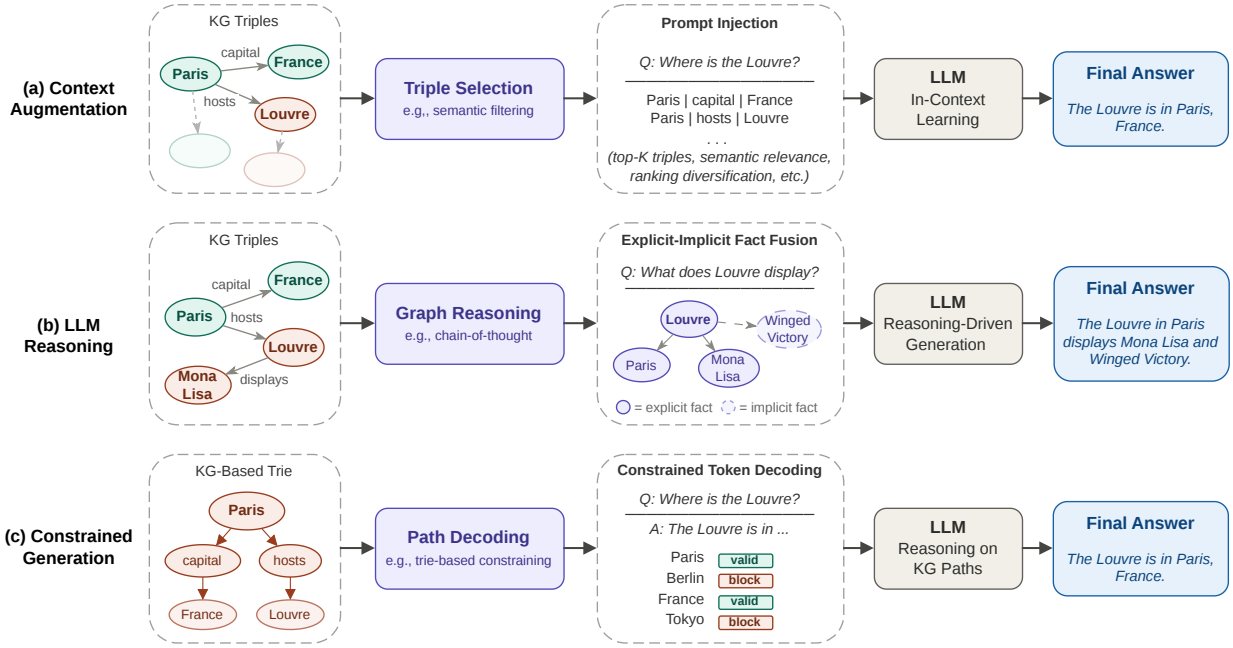

Figure 7: Overview of retrieval-based approaches, highlighting the similarities and differences between context augmentation, LLM reasoning, and constrained generation.

and injects pieces of dialogue history. IKA comprises distinct modules: the *retrieval* module gathers relevant KG knowledge and dialogue history, the *diversity* module enhances the coverage of the candidate knowledge by content diversification, and the *prompt construction* module assembles the final LLM input. External KG knowledge is incorporated as linearized triples to increase information density and reduce token consumption compared to free text. Knowledge Injection (KI) [115] adopts the principles of *controllable text generation* [132, 197] to guide the LLM by injecting factual information from business-specific KGs. Evaluation from domain experts highlights increased factual correctness and overall text quality when using small language models (e.g., BLOOM 560M [146]). Focusing on the medical domain, MedKit [152] injects knowledge from a biomedical KG [17] to reduce hallucinations in radiology report generation. MedKit consists of three steps: *i)* identification of relevant entities in the KG to derive disease-specific concepts, *ii)* prompt template generation to incorporate the extracted knowledge (e.g., disease symptoms), and *iii)* embedding fusion to inject the template directly into the LLM.

DualEval [2] and DualR [100] adapt cognitive psychology theories to support context augmentation. Inspired by *semantic priming*, i.e., the effect of increasing accuracy and speed in accessing concepts from human memory when exposed to relevant related concepts, DualEval suggests to prime LLMs by injecting factual KG triples semantically relevant to the LLM prompt. Two evaluation sets are introduced: a *TrueSet*, containing factual KG triples, and a *FalseSet*, containing fabricated entities without supporting triples. Results from injecting the two sets demonstrate increased factuality from priming LLMs with the TrueSet. DualR draws inspiration from the dual process theory [15, 79] to investigate the combination of two *systems* to improve LLM factuality: *System 1* (fast thinking process) performs implicit reasoning with a frozen LLM and *System 2* (slow thinking process) performs explicit reasoning with a GNN. System 2 explores a KG by iterative pruning and propagation, linking input queries to candidate answers. Candidate answers are passed to System 1, which selects the most appropriate candidate based on its internal implicit knowledge.

LLM Reasoning Reasoning approaches incorporates chain-of-thought (CoT) mechanisms [83, 180] to elicit reasoning in LLMs, supported by the explicit knowledge representation and interpretable paths contained in KGs. MindMap [182] enables synergistic LLM-KG inference through *graph-of-thought*, a variant of CoT where the reasoning process is performed on graph structures. After retrieving a set of sub-graphs relevant to the entities in the original prompt, a supporting LLM aggregates the sub-graphs to build a set of *reasoning graphs*, which are consolidated together to produce a comprehensive *mind map*. MindMap demonstrates reduced hallucination rates and improved generation accuracy in general and domain-specific tasks (e.g., biomedical domain), by combining explicit KG evidence with implicit LLM knowledge. This combination overcomes the limitations of stand-alone LLMs (which fail to incorporate explicit factual knowledge) and RAG approaches (which disregard the LLM internal knowledge). Similarly, DCRAG [194] integrates implicit and explicit knowledge by *thought-then-generate* reasoning, an interac-

tive collaboration between LLM and KG. DCRAG identifies knowledge demands based on the dialogue history in three stages: *i)* extraction and inference of relevant entities from the dialogue history, *ii)* selection and expansion of the entities, and *iii)* coherence revision before retrieving facts from the external KG. GraphVis [32] introduces *visual reasoning* on KGs, by transforming sub-graphs from ConceptNet [151] into visual graph representations. Contrary to other reasoning approaches, that focus mostly on the semantic layer of KGs, GraphVis leverages a visual language model to provide deeper understanding of the KG structure. Visual reasoning is shown to better captures the KG relational context compared to context augmentation by linearized triples, increasing accuracy in generating factual statements.

Constrained Generation Constrained generation, or *constrained decoding*, reduces hallucinations by restricting the LLM candidate tokens to valid facts from external sources [22, 52]. Graph-Constrained Reasoning (GCR) [111] introduces path-constrained generation to improve factuality by enhancing accuracy in traversing KGs during multi-hop reasoning, a task where LLMs typically struggle [120]. GCR incorporates the KG structure into the LLM decoding process using prefix tree structures (i.e., *trie*), constraining reasoning by disabling tokens not corresponding to valid KG facts. The approach maintains computational efficiency by leveraging a smaller fine-tuned model for path decoding, while a larger LLM (e.g., GPT4 [123]) produces the final answer by reasoning on the intermediate KG paths. Notably, the small model for path decoding generalizes on previously unseen KGs (e.g., UMLS [17]), enabling portability to new domains. ReFactX [131] scales path-constrained generation to 800 million facts from Wikidata [172], by indexing factual triples in a tree structure and storing them in a relational database. Compared to context augmentation approaches, constrained generation ensures that the generated tokens adhere strictly to known facts without relying on external retriever modules, which are computationally expensive and require ad-hoc training. However, despite performing well with open questions that require point-wise factual information (e.g., *Who directed the movie X?*), ReFactX highlights the struggle of constrained generation with other types of questions – such as count queries (e.g., *How many movies did X direct?*) – due to the LLM inability to count during inference.

4.2.2 Network Modules

Network modules are neural components added to the system architecture to support LLMs by encoding and managing external KG data (see Figure 8). RHO [75] mitigates hallucinations by modifying the encoder module of encoder-decoder architectures (e.g., BART [95]) to fuse embeddings from external KGs and from dialogue history. Two complementary mechanisms support KG embeddings: *local knowledge grounding*, which projects the embeddings of entities and relations directly into the LLM latent space, and *global knowledge grounding*, which embeds sub-graph information relevant to the conversation to increase multi-hop capabilities. The decoder module receives as input the concatenation of local, global, and dialogue history knowledge. Finally, the most faithful response from multiple candidates is selected by a re-ranking module [145]. MR-MKG [91] focuses instead on factual accuracy in multimodal reasoning tasks, combining multiple data modalities and structured to support factual LLM generation. The approach aggregates textual, visual, and KG representations, leveraging a pre-trained visual encoder [134] and a relational graph attention network [73, 174]. To improve consistency between textual, visual, and graph content prior to the LLM generation, a cross-modal alignment component fuses the resulting embeddings to ensure that all the representations adhere to a common latent space.

4.2.3 Discussion

In-generation methods integrate factual knowledge directly at runtime, without incurring additional costs during model training [10]. However, this flexibility comes with a set of limitations depending on the specific strategy employed. For instance, context augmentation techniques increase token consumption, network modules introduce additional computational overhead, and LLM reasoning can substantially amplify both, due to long traces generation [50] and multiple LLM calls [182]. While the use of small language models [1, 160] can mitigate these costs [115], it introduces a fundamental trade-off between efficiency and effectiveness: larger models consistently demonstrate stronger performance on complex reasoning tasks, partly due to emergent capabilities that do not arise at smaller scales [179]. A further limitation concerns closed-source models, including some of the most used and best-performing LLMs (e.g., GPT or Claude). While context augmentation techniques are compatible with closed-source models, approaches that require architectural modifications or manipulation of the model's predictions are not, such as network modules [75, 91] or constrained generation [111, 131]. Context augmentation also faces the problem of *knowledge conflict*: the LLM may disregard the external knowledge or exhibit inconsistent behavior when contextual information contradicts the parametric knowledge [106, 187]. Therefore, while context augmentation can reduce hallucinations on average, they can inadvertently increase the rate of faithfulness hallucinations. This distinction calls for more fine-grained evaluation methodologies to differentiate between hallucination types. Finally, while context augmentation is relatively transparent, as the injected knowledge remains in compact human-readable form, how in-context learning influences and

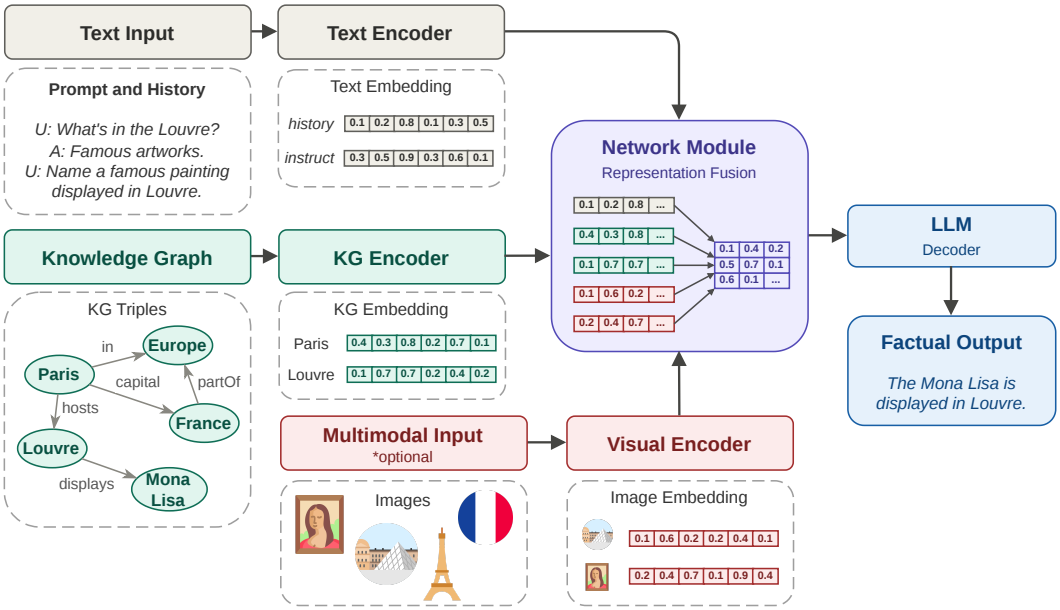

Figure 8: Overview of network modules approaches.

directs the internal reasoning process of LLMs remains an open question [29, 167]. In contrast, network modules encode KG knowledge as distributed neural representations, inheriting the general opacity of deep-learning models.

4.3 Post-Generation Methods

Post-generation methods reduce hallucinations after LLM generation by first detecting factual inaccuracies in the output, then correcting them. They can be divided into two categories: mitigation by *knowledge comparison* and *knowledge integration*.

4.3.1 Knowledge Comparison

Knowledge comparison methods perform an external comparison between retrieved KG information and the LLM output to detect and correct hallucinations. They can further be classified according to the granularity of the comparison: *graph-based* methods compare entire sub-graphs with graph-transformed LLM outputs, leveraging either symbolic (e.g., graph kernels [85]) or sub-symbolic techniques (e.g., GNNs [185]); *triple-wise* methods compare individual LLM claims with corresponding KG triples, adopting a finer granularity approach. An overview of knowledge comparison methods is shown in Figure 9.

Graph-Based Comparison KEA [61] detects hallucinations by comparing the LLM response with either an external encyclopedic KG (*open-domain scenario*) or a KG constructed from contextual knowledge (*closed-domain scenario*). An LLM-driven approach construct the response KG [142], while relevant triples from the reference KG are retrieved by entity selection through semantic similarity [140]. Hallucination detection relies on measuring structural similarity between the response KG and the reference KG, adapting the Weisfeiler-Lehman algorithm [149]. Comparing the entire graph improves detection in long-form responses, while the algorithmic-based approach improves explainability by highlighting mismatched or unsupported triples. FactGraph [141] focuses on closed-domain scenarios by comparing KG representations of input documents with KG representations of LLM summaries to identify and correct faithfulness hallucinations in text summarization. The approach encodes both semantic and structural information by leveraging fine-tuned adapters for text-to-graph transformation: a *text encoder* adapter for the semantic content and a *graph encoder* adapter for encoding structural information. FactGraph advances text-to-graph encoding [55] by capturing the full structure of text, demonstrating high correlation with human judgment on hallucination detection. Although being tested solely on closed-domain scenarios, FactGraph can be extended to open-domain scenarios to detect factuality hallucinations, e.g., by using external encyclopedic KGs similarly to KEA [61]. Furumai et al. [51] detect hallucination using a relational graph attention network (RGAT), a GNN model that incorporates an attention mechanism [73, 174]. A *graph module* constructs the response KG representing the LLM output, while a reference sub-graph is retrieved from a reliable external KG. A *matching module* encodes the response KG and the reference sub-

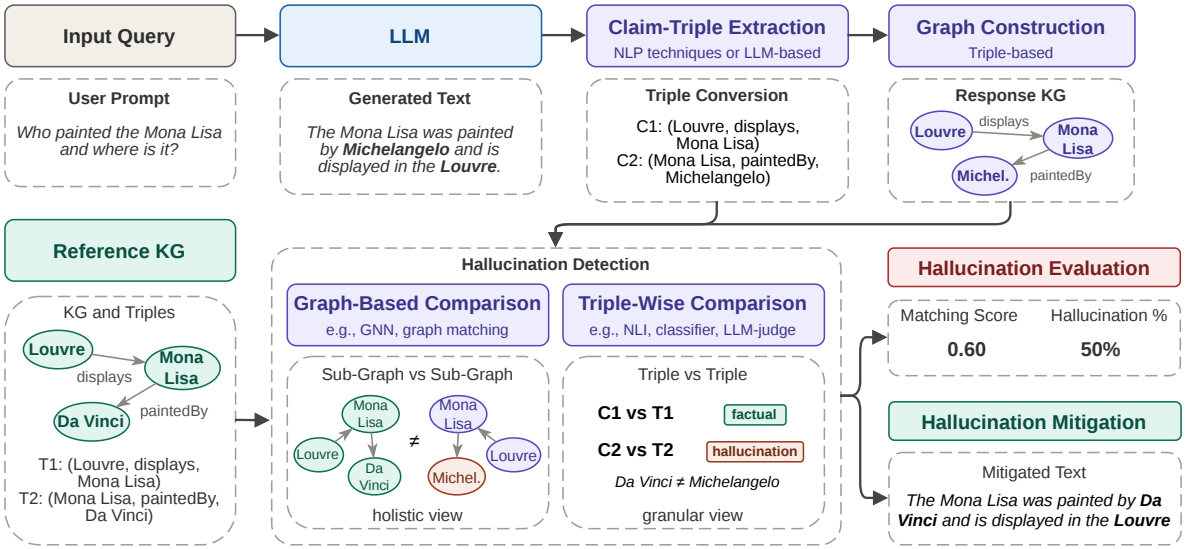

Figure 9: Overview of knowledge comparison approaches.

graph, which are then provided as input to the RGAT classifier for hallucination detection. To improve interpretability, the system provides a visualization of the entities and relations involved in the hallucination.

Triple-Wise Comparison In contrast to graph-based comparison, FactAlign [138] identifies individual hallucinated statements by estimating semantic similarity between LLM claims and the corresponding triples from an external KG. A threshold, selected through Bayesian optimization, is applied to the score to determine whether a claim should be classified as hallucination. Both factuality and faithfulness hallucinations are supported, depending on the reference KG used. Faithfulness hallucination detection enables the integration of FactAlign with information retrieval pipelines (e.g., RAG), serving as a KG-based alternative to LLM-driven frameworks such as RAGAs [46].

A number of approaches rely instead on deep-learning classifiers for hallucination detection. CoKGLM [60] pre-trains a binary cross-attention classifier [75] for factuality hallucinations. For each entity in the LLM output, CoKGLM retrieves the associated triples from an external KG and ranks them by semantic similarity with the overall LLM dialogue history. Only the top-K most relevant triples are selected to be encoded and fused with dialogue history embeddings. The resulting fused embeddings and the text embeddings of the LLM output become the input of the pre-trained classifier, which returns a probability of hallucination. GraphEval [142] focuses on faithfulness hallucinations, detected via natural language inference (NLI): an NLP task that classifies the relationship between two text fragments as entailment, contradiction, or neutral, with contradictions treated as evidence of hallucination. GraphEval builds on earlier NLI-based methods for hallucination detection, e.g., SummaC [87], FactScore [117], while providing a more structured and fine-grained analysis by transforming text into a set of triples. Hallucination correction is performed by an external LLM, prompted with the original context and the associated factual triples. GLLM [184] extends NLI-classification to factuality hallucinations, comparing LLM claims against external KGs. For each triple extracted from the LLM output, the corresponding entities and relations in the external KG are retrieved via a combination of similarity and breath-first search. On top of per-triple scores, an overall hallucination score is computed as the average of per-triple scores, combining fine- and coarse-grained evaluation.

Recently, comparison frameworks supported by LLMs have started to emerge, involving evaluation strategies inspired by LLM-as-a-judge [202]. Notably, FLEEK [48] adopts a prompt-based approach to compare LLM claims with retrieved factual evidence, classifying the comparison into three categories – supported, likely supported, or questionable – and suggesting corrections for unsupported claims. FLEEK generates validating questions for each claim in the original LLM output (e.g., *What is LeBron James's age?* for the claim *LeBron James is 40*) and retrieves supporting evidence for answering the validating questions from both an external encyclopedic KG and the Web. Mixed source retrieval enables to retrieve trustworthy evidence (from KGs) while achieving broad coverage of topics (from the Web).

4.3.2 Knowledge Integration

Knowledge integration methods perform post-hoc knowledge injection in the LLM system to retroactively correct the original hallucinated output. In contrast to knowledge comparison techniques, where the comparison and mitigation is

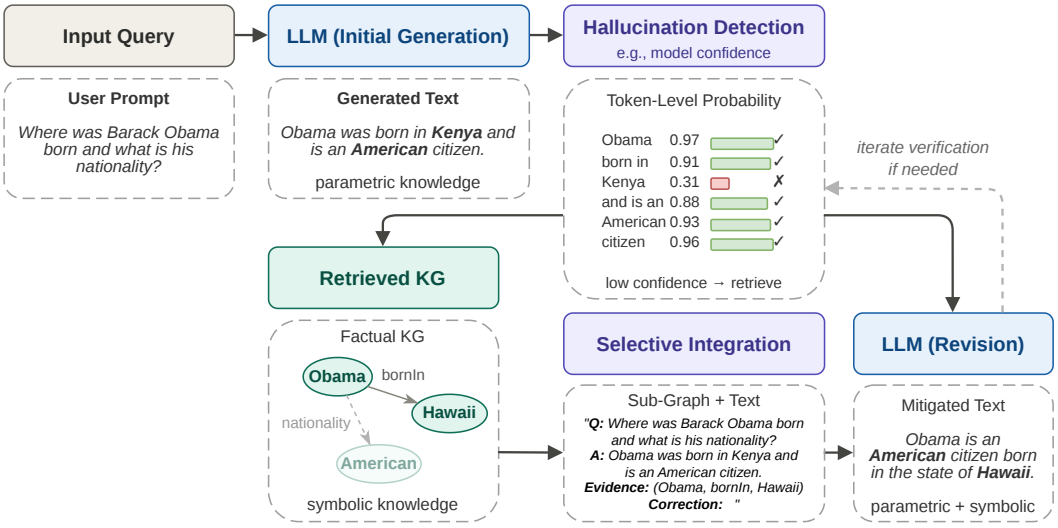

Figure 10: Overview of knowledge integration approaches.

performed externally, here the integration of KG knowledge is performed by the LLM itself, similarly to in-generation methods [27, 75]. However, hallucinations are always corrected after a first round of generation, thus applying KG intervention only when necessary, reducing the risk of faithfulness hallucinations, and fostering the integration of symbolic and parametric knowledge. An overview of knowledge integration methods is presented in Figure 10.

Neural Path Hunter (NPH) [42] pioneers KG knowledge integration to reduce factuality hallucinations in dialogue generation. NPH identifies and masks hallucinated entities in the model output, then corrects the masked entities by selecting the most appropriate completion candidate from external K-hop sub-graphs. The selection is performed by a fine-tuned language model. Knowledge Graph-based Retrofitting (KGR) [56] identifies hallucinations in LLM responses and verifies them by retrieving KG facts associated to the entities mentioned in the response. When the validation fails, the response is retrofitted by introducing factual knowledge, while preserving correct LLM-generated content. The approach leverages multiple external LLMs to support entity recognition and sub-graph retrieval, generating a *chain-of-verification*. KGR achieves high accuracy in open-domain scenarios, addressing KG incompleteness by combining (and not disregarding) LLM parametric knowledge and explicit KG information. GraphRAG-FI [57] addresses the tendency of GraphRAG approaches to over-rely on the retrieved information [58], by integrating and merging standard GraphRAG outputs and LLM-only outputs. GraphRAG-FI leverages two intuitions: *i)* attention scores approximate information relevance [183] and *ii)* token-level probability indicates model confidence [41, 47]. Thus, this method filters out retrieved graph paths below an attention-based relevance threshold, and introduces graph knowledge when the LLM response exhibits low confidence.

4.3.3 Discussion

Post-generation methods perform external verifications for hallucination detection, making them highly portable between models and compatible with both open-source and proprietary LLMs. Computational costs depend primarily on the efficiency of the KG retrieval process and the extent to which LLMs are involved across the pipeline. However, LLMs have been increasingly adopted for graph retrieval [56], claim extraction [48, 57], and automated KG construction [142, 184]. KG retrieval remains the primary challenge, as hallucination detection is only as reliable as the factual evidence retrieved, while claim extraction from LLM responses presents the challenge of decomposing long text into complete, verifiable atomic claims [48]. Hallucinations can be detected by comparison of external and internal knowledge with language models, e.g., [62, 105], fine-tuned on specific tasks such as NLI [142, 184] or binary classification [138]. On the other hand, classifiers based on LLM-as-a-judge frameworks [48, 56] do not require further training, but are generally more costly to run and risk introducing biases during the verification step [202]. All post-generation methods share common limitations: the LLM response is always produced before the system intervenes and – whenever the detection fails – the correction is not performed. On the other hand, correction is triggered only when hallucinations are actually detected, thus limiting the computational overhead compared to in-generation methods.

Table 3: Summary of KG-derived hallucination benchmarks. Task: QA = question-answering, Conv-QA = conversational QA, Open-QA = open-ended QA, MC-QA = multiple-choice QA, CV = claim verification, LC = logical consistency. Type: E = encyclopedic KG, C = commonsense KG, D = domain-specific KG.

Benchmark	Year	Scope	Size	Split	Task	Repo	KG	Type	Evaluation
FactEval [110]	2023	Factuality	N/A	N	QA		Google-RE, T-REx, UMLS, WikiBio	E, D	Accuracy
LTGen [71]	2025	Conversation, Long-Tail	1.2K	N	QA + Conv-QA		Wikidata	E	Accuracy + NLI
CoLoTa [164]	2025	Reasoning, Long-Tail	3.3K	N	QA + CV		Wikidata	E	Accuracy + Factuality + Reasoning
LFC [54]	2025	Logical Consistency	184K	Y	LC		Wikidata, Freebase, NELL	E	Accuracy + Consistency
FAITH [195]	2024	False-Premise	6K	N	QA		Wikidata	E	Uncertainty
KG-FPQ [206]	2025	False-Premise	178K	N	QA + Open-QA		Wikidata	E	Factuality
OKGQA [154]	2025	Robustness	2K	N	Open-QA		DBpedia	E	Factuality
GeoHaluBench [176]	2025	Robustness, Domain-Driven	2.1K	N	MC-QA		SpatialKG	D	Accuracy

4.4 KG-Based Benchmarks for Hallucination Evaluation

KGs are a rich source of data for constructing benchmark datasets due to their large-scale, structured, and accurate representation of knowledge. Numerous benchmarks have been derived from KGs to evaluate hallucinations in LLMs, including tasks such as claim verification [164], logical consistency assessment [54], false-premise questions [195, 206], and long-tail QA [71, 164]. Table 3 provides an overview of KG-based benchmarks for evaluating factuality and hallucination.

The most common hallucination task is factual recall. FactEval [110] introduces a factual evaluation framework by generating QA pairs from both general-domain KGs (Google-RE [124], T-REx [45]) and domain-specific ones (WikiBio [156], UMLS [17]), spanning true/false, multiple-choice, and short-answer formats. LTGen [71] complements this by targeting long-tail knowledge, in both standard and conversational QA settings. LTGen covers Wikidata-extracted entities at different levels of popularity to obtain a varying degree of question complexity.

Other benchmarks move from simple factual recall to assessing complex and compositional reasoning. CoLoTa [164] evaluates LLM commonsense reasoning in claim verification and QA over long-tail knowledge, targeting both domain-independent skills (e.g., temporal reasoning, numerical comparison) and domain-specific ones (e.g., reasoning on history, geography, or sports). LFC [54] takes a different angle by focusing on logical hallucinations: it constructs three datasets from Freebase [18], NELL [24], and Wikidata [172] respectively, where queries concatenate facts with logical operators, e.g., conjunction, disjunction, implication. The evaluation assess not only individual answer accuracy, but also the logical consistency according to related answers, e.g., if p is classified True, then $p \vee q$ should be consistently classified True.

A distinct set of benchmarks targets false-premise hallucinations, i.e., factual errors induced by misleading or incorrect information injected in the initial question [68]. FAITH [195] and KG-FPQ [206] both generate false-premise questions by systematically corrupting triples form encyclopedic KGs. KG-FPQ offers a more scalable benchmark construction pipeline and multiple knowledge-editing strategies, including entity substitution at varying hop distances and across semantically similar or unrelated entity types (e.g., substituting a city with another city, or a city with a person). KG-FPQ further distinguishes between two QA formats (discriminative "yes/no" and generative open-ended questions) and three domains of knowledge (art, people, and places).

Finally, OKGQA [154] and GeoHaluBench [176] examine hallucination robustness under noisy or perturbed knowledge. OKGQA considers open-ended QA for KG-augmented LLMs, pairing its main benchmark with a variant (OKGQA-P) that simulates incomplete and noisy retrieval of KG information. GeoHaluBench extends the focus on robustness to the geo-spatial domain, constructing a domain-specific KG (adapted from [103]) and a benchmark that targets entity-, relation-, and attribute-level hallucinations in the form of a multiple-choice QA task, where one choice is factual and the others represent different types of hallucination, such as factual fabrication (e.g., non-existing locations) or incorrectness (e.g., negating the existence of a real location).

Table 4: Comparison overview of KG-based approaches for hallucination reduction (pre-, in-, and post-generation). Legend: ● ● ● = high, ● ● ○ = medium, ● ○ ○ = low.

Approach	Development Efficiency	Runtime Efficiency	Transparency	Scalability	Portability
Full Fine-Tuning	● ○ ○	● ● ●	● ○ ○	● ○ ○	● ○ ○
PEFT	● ● ○	● ● ●	● ○ ○	● ● ○	● ● ○
Probing	● ● ○	● ● ●	● ● ○	● ● ○	● ● ○
Context Augmentation	● ● ●	● ● ○	● ● ○	● ● ○	● ● ●
LLM Reasoning	● ● ●	● ○ ○	● ● ○	● ○ ○	● ● ○
Constrained Generation	● ● ○	● ● ●	● ● ○	● ● ○	● ○ ○
Network Modules	● ● ○	● ● ○	● ● ○	● ● ○	● ● ○
Knowledge Comparison (Graph Based)	● ● ○	● ○ ○	● ● ○	● ● ○	● ● ●
Knowledge Comparison (Triple Wise)	● ● ○	● ○ ○	● ● ●	● ● ○	● ● ●
Knowledge Integration	● ● ●	● ○ ○	● ● ○	● ● ○	● ● ○

Discussion Hallucination benchmarks from KG highlight a landscape of solutions: from explicit KG facts coverage [110] to targeted assessment of long-tail knowledge [71,164], specific hallucination types [54,176], and misleading scenarios [154,176,195,206]. LLM factuality is reduced in specialized domains [110,176], for questions with multiple valid answers [110], and when adversarial or irrelevant context is injected [110, 154, 195, 206]. Factual inaccuracies are more frequent when irrelevant context is semantically close to the original, suggesting that LLMs are particularly susceptible to plausible falsehoods [206]. LLMs also struggle with consistency on complex logical queries [54]. Overall, hallucination arises not by random errors, but due to systematic weaknesses tied to knowledge gaps, semantic uncertainty, or reasoning complexity. The prevalence of inaccuracies by omission over fabrication in specific domains suggests that hallucinations often reflect knowledge gaps and biases in training data rather than conflicting parametric knowledge (e.g., geo-spatial hallucinations more common in Beijing than New York [176]).

A recurring challenge across KG-derived benchmarks concerns scalability during construction and evaluation. The size of KGs enables automated large-scale construction with broad knowledge coverage, but processing large KGs remains computationally demanding, particularly when LLMs are involved in triple selection or corruption [110, 154, 206]. Regarding evaluation, standard statistical measures (e.g., accuracy, F1) are straightforward to apply but provide limited insight. Ad-hoc factuality metrics (e.g., FActScore [117] or SAFE [181]) offer specialized assessment with additional complexity, while LLM-as-a-judge [202] adds flexibility at the expense of increased computational cost and potential reliability issues.

4.5 Summary

Table 4 summarizes the categories of approaches discussed in this section, highlighting their key characteristics, advantages, and limitations across five properties: *i) development efficiency* describes the cost associated with training the LLM or developing the system components; *ii) runtime efficiency* describes the computational overhead incurred during inference; *iii) transparency* determines the interpretability of the approach; *iv) scalability* represents the ability to scale with larger LLMs and graph sizes; and *v) portability* represents the ease of adapting the solution to different LLMs or integrating it into different system architectures. The comparison reveals distinct trade-offs, highlighting existing challenges and limitations. For a consolidated overview of the literature, Table 5 in appendix summarizes all pre-, in-, and post-generation approaches.

5 Open Challenges and Future Directions

Based on the literature discussed previously, this section highlights open challenges and future research directions in reducing LLM hallucinations with KGs. Figure 11 presents an overview of the challenges discussed in this section across the stages of the LLM lifecycle.

5.1 KG Incompleteness

While KGs are trustworthy sources of structured factual knowledge, they are also limited and often incomplete [31]. As a result, outside of closed-domain scenarios – where the source knowledge is assumed to be complete [112] –

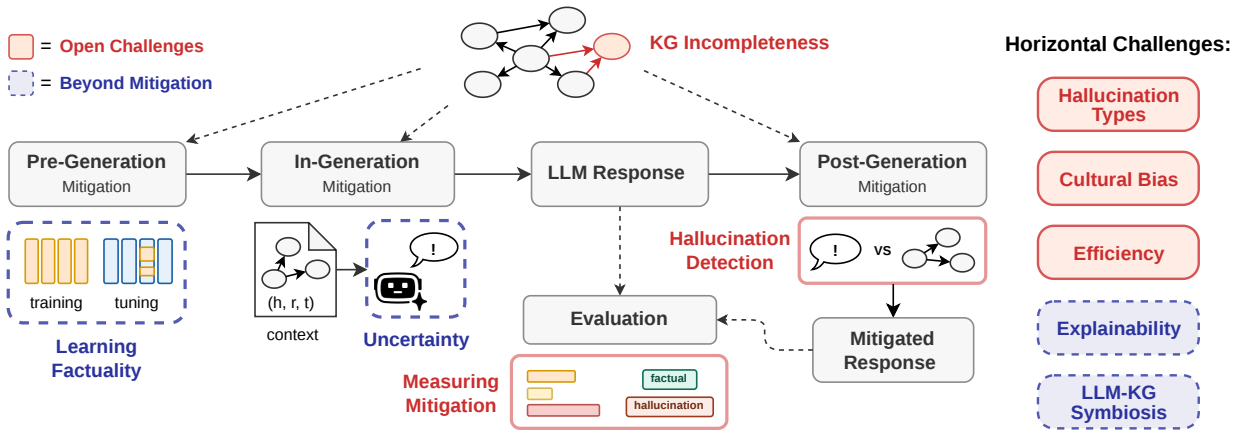

Figure 11: Mitigation approaches within the LLM lifecycle and position of open challenges.

incomplete or inconsistent retrieval can introduce noise and knowledge conflicts. This affects both in-generation and post-generation methods, which rely on retrieval to inject facts in the context [91] and for verification [48]. A recent evaluation of KG retrieval approaches under incompleteness confirms the lack of robustness in this setting [203]. In this context, the following challenges emerge.

1. **Improve quality of retrieval**: Moving from traditional NLP-based retrieval towards LLM-assisted retrieval [48, 56] improves KG navigation and expands the available reference knowledge, but also risks introducing additional hallucinations. This challenge requires balancing LLM and traditional processes to achieve higher efficiency, transparency, and accuracy.
2. **Overcome incompleteness**: LLM reasoning can predict missing links [43] or generate knowledge from incomplete KGs by combining explicit and implicit knowledge, inferring logically-entailed facts, and deriving semantically-related knowledge [188]. Agentic approaches augment KG retrieval with iterative LLM reasoning [77, 78, 113, 155, 205], offering a promising direction to combine symbolic and sub-symbolic retrieval.
3. **Overcome limited knowledge**: No single KG provides universal coverage, hence integrating multiple sources can extend the knowledge coverage to multiple domains, e.g., by combining evidence from encyclopedic and domain-specific KGs. However, how to manage schema misalignment, ad-hoc entity resolution, potential inconsistencies across the sources, and conflict resolution remains an open question.

5.2 Hallucination Detection

Before mitigation, post-generation methods detect hallucinations by comparing LLM-generated text against KG facts (see examples in Figure 12). Comparison strategies span two dimensions: granularity and abstraction. Fine-grained strategies perform triple-wise comparison, while coarse-grained strategies consider entire graphs or text passages. Abstraction describes the underlying model: from symbolic approaches such as triple matching or graph kernels [61], to sub-symbolic assessment involving GNNs [51], NLI classifiers [142, 184], or LLMs [48, 56] – see Figure 13. Each strategy offers distinct advantages, e.g, fine-grained strategies offer precision but lack a holistic view, while coarse-grained strategies provide overview at the cost of interpretability and detailed diagnostic. However, a systematic evaluation is currently missing, and how to combine complementary techniques remains an open question that raises the following challenges.

1. **Convert KG to text**: Fine-grained strategies based on LLMs and NLI classifiers require converting KG triples into natural language. Triple linearization is a widespread approach due to its simplicity, but disregards graph structure entirely. Semi-structured formats such as JSON, XML, or YAML can preserve more relational information [30], though at the cost of increasing verbosity and token consumption. Identifying the optimal representation to balance structural fidelity, token efficiency, and interpretability remains an open challenge, with direct implications also for context augmentation methods (see Section 4.2.1).
2. **Convert text to KG**: Conversely, graph-based comparison requires converting the LLM output into structured representations: individual triples [142, 184] or full graphs [61]. The use of LLMs supports automated KG construction, but risks introducing additional hallucinations before the verification step. Reasoning models

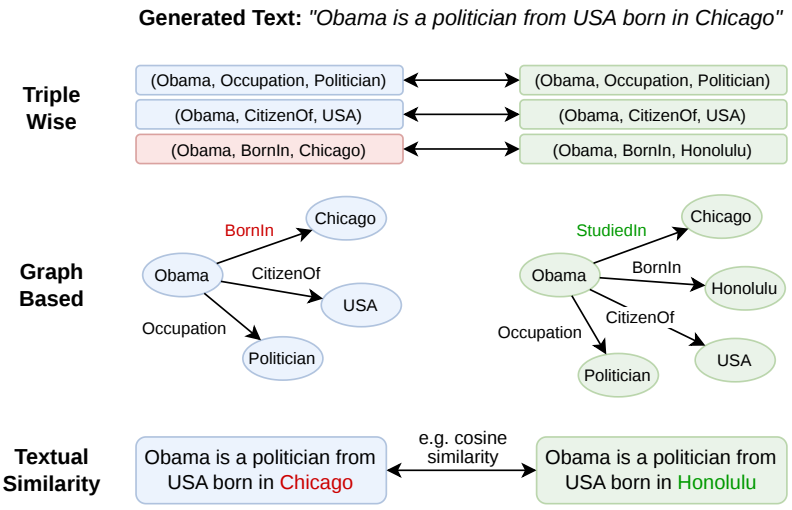

Figure 12: Examples of comparison strategies between LLM generated text and KG knowledge.

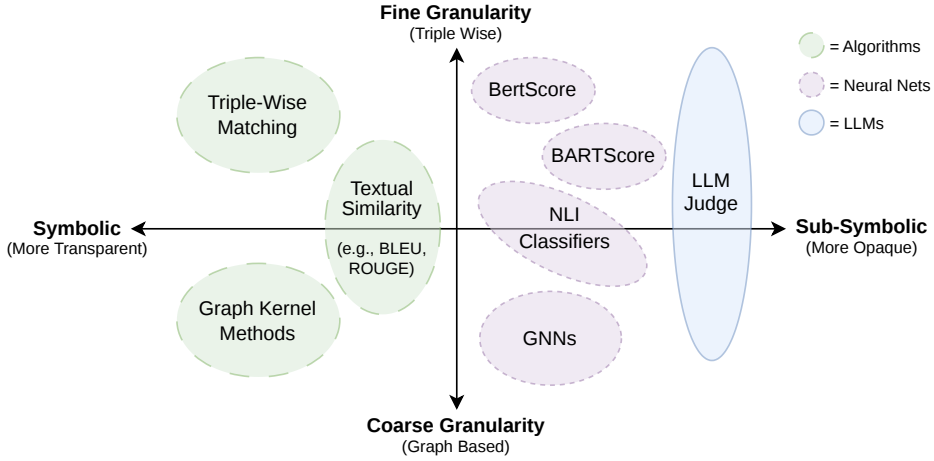

Figure 13: Comparison strategies along the granularity and abstraction dimensions.

can improve LLM consistency in KG construction, albeit increasing computational cost and token consumption. Finally, agentic frameworks offer a promising direction to combine stochastic LLM processes with more transparent and efficient deterministic tools [108].

3. **Continuous scores**: Current detection approaches frequently reduce comparison to binary classification [51, 138, 141]. Continuous scores would provide richer and more informative outcomes by quantifying the degree of factual divergence from the reference KG, distinguishing between omissions, contradictions, fabrications, and plausible statements.

5.3 Measuring the Effect of Mitigation

Evaluating the effectiveness of mitigation techniques requires standardized benchmarks and metrics. However, many approaches rely instead on proxy evaluation strategies such as QA datasets, which assess the LLM ability in answering a set of questions extracted from a KG. However, QA benchmarks rarely account for model abstention, favoring guessing over uncertainty [80] by relying on global metrics such as accuracy or Hit@K. For instance, considering the question *How many Nobel Prizes was Marie Curie awarded?*, the answer *I am not confident enough to respond* would be considered inaccurate despite not containing any hallucination.

On the other hand, a number of hallucination-specific benchmarks have been proposed, involving QA tasks [99], fact-checking scenarios [5, 82, 161], multi-domain settings [28], and multi-task evaluation [97]. However, their adoption

is inconsistent in the context of KG-based hallucination reduction: different solutions tend to adopt benchmarks tailored to a specific task or domain, limiting comparison between solutions and trade-off evaluations across different tasks (e.g., QA, summarization), question formats (open-ended, multiple-choice, adversarial), context properties (e.g., multi-hop, multi-modal), knowledge domains (e.g., biomedicine, geography, finance), and languages (see Table 5 for an overview of benchmarks and metrics adopted in the surveyed literature). Within this broad context, the following challenges can be identified.

1. **Establish a unified benchmark**: There is a need for a comprehensive hallucination benchmarks for KG-based solutions, covering multiple settings, domains, and reference KGs. Recent efforts such as [69] improved task variety, but without including reference sub-graphs or triples as ground truth, nor integrating multiple KG sources for increased coverage.
2. **Multilingual hallucination**: The relationship between hallucination, language, and domain of knowledge remains largely unexplored. Only a small number of current approaches evaluate hallucination reduction in languages other than English, with Chinese being a notable exception [147, 182, 194]. Existing benchmarks are predominantly English-centric, with only recent efforts addressing multilingual evaluation [69,90]. Moreover, translating the KG reference ground truth into multiple languages represents an open challenge, with LLM-based translation risking introducing factual inaccuracies.
3. **Long-form text**: Longer LLM responses contain multiple claims, requiring metrics to aggregate claim-level assessments into coherent text-level scores beyond simple proportion of truthful claims [117, 181]. Semantic similarity evaluators [196, 198] and LLM-as-a-judge [104, 202] offer more detailed diagnosis, although reducing interpretability and lacking deterministic behavior. Moreover, decomposing text into atomic claims represents an additional challenge [48].
4. **Internal assessment**: Current hallucination assessment is often external, e.g., measuring factuality through QA accuracy. Therefore, it is often unclear why and how a given mitigation technique reduces hallucinations *internally*, nor how robust is the intervention. This limitation calls for probing the internal model behavior and advancing research on KG-derived robustness benchmarks, such as [154, 176].
5. **Quality of metrics**: Current metrics adopted in benchmarks exhibit weak inter-metric correlation and inconsistent alignment with human judgment [86]. Introducing hallucination metrics that respect these two properties would provide more informative and reliable evaluations, supporting the overall mitigation effort.

5.4 Hallucination Types

Most approaches target either factuality or faithfulness hallucination in isolation. Furthermore, as illustrated in Figure 1, specific hallucination types demand dedicated investigation due to their unique characteristics, thus opening a number of challenges.

1. **Factuality and faithfulness**: Developing benchmarks to cover both categories would provide a more detailed overview on hallucinations. This effort requires combining intrinsic hallucination evaluation [13, 121, 157] with factuality assessment [28, 97, 99].
2. **Hallucination from outdated knowledge**: KGs can be updated more efficiently than LLMs [24, 118], making them a natural solution for addressing hallucination from outdated knowledge. However, how to update KGs effectively, which sources to use to guarantee trustworthiness, and the frequency of updates are open questions.
3. **Reasoning inconsistency**: The widespread adoption of CoT and reasoning models introduces hallucinations arising from inconsistencies across intermediate reasoning steps. Although KGs offer a natural framework for representing reasoning traces to identify intrinsic hallucinations, their use for evaluating reasoning faithfulness remains largely unexplored. This gap persists in the context of both few-shot prompting and specialized reasoning LLMs.

5.5 Cultural and Social Bias

Hallucinations from stereotypical bias derive from erroneous associations learned during training [74]. Considering external knowledge sources from different cultures is fundamental to mitigate factual hallucinations while reducing the risk of cultural bias and partiality. For instance, relying on a commonsense KG with a Western-centric perspective may lead to classifying an LLM suggestion to eat with hands as a hallucinated practice, despite its cultural relevance in many regions of the world. Addressing hallucinations from bias by introducing multi-cultural knowledge sources represents a step towards introducing KGs in responsible AI practice [168], yet includes the following challenges.

1. **KG bias reduction**: KGs can themselves contain biases [84]. Finding and addressing social and cultural bias in KGs before their use as factual references is fundamental to improve assessment and reduce hallucination effectively.
2. **Cultural coverage**: Encyclopedic KGs lack multi-cultural coverage and frequently adopt Western-centric perspectives [84], with few existing KGs featuring cultural and stereotype awareness [34]. Capturing subjective, belief-based, and context-dependent cultural knowledge (e.g., how many wives can a man have or whether God exists) requires dedicated effort.
3. **Robustness to stereotypical hallucination**: Beyond serving as reference sources for reducing hallucination, multi-cultural KGs would provide structured resources to support the creation of adversarial datasets to assess robustness against hallucination from stereotypes, covering ethnic, religious, and social dimensions.

5.6 Time and Cost Efficiency

Emergent LLM abilities, such as few-shot learning and CoT, carry significant computational and token costs [179]. Similarly, KG-based approaches involving LLM reasoning [182] severely amplify time and computational complexity, such as token consumption, memory required, number of LLM invocations. Moreover, large-scale graphs require substantial resources to store and traverse. Balancing mitigation effectiveness against time and resource efficiency remains an open challenge across three dimensions.

1. **Measure efficiency**: KG-based solutions should be evaluated not only in terms of accuracy: time complexity, computational cost, memory requirements, and token consumption are also important. However, comprehensive efficiency assessment is frequently overlooked, implying that many proposed solutions may be impractical for real-world deployment due to their high computational and resource costs. KGs intrinsically offer more compact and concise knowledge than formats like text [71] but their large sizes also hinder efficient fact retrieval.
2. **Agentic solutions**: Reducing the involvement of expensive LLMs can be achieved by integrating graph mining processes as external tools within agentic pipelines. Hybrid architectures would delegate graph traversal to more efficient and interpretable symbolic techniques, while preserving the reasoning and generative strengths of LLMs.
3. **Small language models**: Reducing LLMs size greatly reduces computational costs and inference latency. While small language models (SLMs) have proven effective for a number of simple tasks, including agentic AI applications [14], they typically exhibit weaker performance in complex reasoning and in handling externally injected knowledge [71]. Leveraging SLMs for graph navigation, claim decomposition, KG construction, and other processes involved in reducing hallucinations with KGs remains an open research direction.

5.7 Beyond Hallucination Mitigation

While mitigation is the most investigated approach towards reducing hallucinations, some studies highlight the inevitability of this phenomenon under current LLM training and deployment conditions [11, 81, 189]. Given the persistence of hallucination risks in increasingly advanced and capable LLMs [116], the following research directions propose a shift towards non-conventional perspectives for reducing hallucinations involving KGs.

Explainability for Hallucinations Explainable AI supports the interpretability of internal LLM processes and inference steps, opening a path towards explaining hallucinations as an alternative to mitigation. Combining the internal view of LLM reasoning with the structured, trustworthy knowledge of KGs enables to expose both the decision-making process and the pieces of knowledge that the LLM draws upon. KGs offer source-traceable representations of entities and relations to explain LLM outputs through human-understandable knowledge structures [137]. Moreover, structuring LLM outputs with triple or graph representations improves verifiability [107]. Therefore, explainable AI can take an intermediate role to improve model transparency and favor the detection of factual inaccuracies and reasoning inconsistencies.

Knowledge Uncertainty A systemic challenge to overcome hallucination is understanding when LLMs are confident or uncertain about their response [80]. Indirect confidence metrics, e.g., based on token probabilities [169] or self-consistency from multiple model invocations [114], are often adopted as proxies for LLM hallucination. However, such indicators can easily fail when LLMs are over-confident about wrong notions or under-confident about correct claims. KGs can support uncertainty estimation by grounding confidence in the properties of their knowledge space: sparse regions of a KG, domains with loosely consolidated knowledge, or frequently updated regions may indicate

higher epistemic uncertainty and lower reliability of the associated LLM output [4, 147]. This perspective shifts confidence assessment from a model-centric paradigm to a knowledge-centric one, by treating the structural properties of the KG as indicators of general epistemic uncertainty.

Learning Factuality Current LLM training and evaluation procedures encourage hallucination by rewarding guessing over acknowledging uncertainty [80]: on the contrary, LLMs should be trained to include uncertainty and abstention as training goals alongside semantic learning, learning to avoid sycophancy, fabrication of facts, and plausible-sounding but incorrect outputs. KGs provide a curated source of verifiable information to support the construction of factuality-aware training datasets and benchmarks. While the scale of contemporary LLM pre-training exceeds the coverage of existing KGs, fine-tuning models for factuality is time- and resource-efficient [162]. Additionally, considering that instruction-tuning have a positive impact on reducing hallucinations [86], reinforcement learning approaches for factuality can incorporate rewards and penalties based on factual consistency, evidence alignment, and faithful reasoning [25].

LLM-KG Symbiosis Unifying LLMs and KGs offers a promising path toward neuro-symbolic systems that combine the strengths of statistical learning and symbolic reasoning [128]. LLM-KG symbiosis is a recent approach to strengthen bi-directional synergy between LLMs and KGs to mitigate hallucinations and misinformation, where KGs provide verifiable knowledge to improve factuality, while LLMs compensate for KG incompleteness [38]. The use of LLMs and reasoning models for knowledge retrieval over KGs has demonstrated strong performance in question-answering tasks [26, 109, 155]. Conversely, recent research addressed the challenge of KG incompleteness through the integration of LLMs to generate and infer missing knowledge [188]. Bridging the two research directions would create systems where LLMs and KGs coexist and support each other: KGs would provide structured, verifiable knowledge and reasoning constraints, while LLMs would contribute by adding flexibility, generalization, and knowledge completion capabilities, resulting in more accurate, reliable, and trustworthy AI systems.

6 Conclusions

KGs have been increasingly adopted to detect and mitigate LLM hallucinations, due to their curated and structured knowledge, compact representation, and efficient navigability. In this survey, we presented a comprehensive classification and analysis of the main approaches to reduce LLM hallucinations with KGs, divided into pre-generation, in-generation, post-generation methods, and evaluation benchmarks. Finally, we identified and discussed a number of open challenges and future research directions that we believe will drive future advances in integrating KGs with LLMs towards more reliable and trustworthy AI systems.

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.

References

[1] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al. Phi-4 Technical Report. *arXiv*, abs:2412.08905, 2024.

[2] C. K. Agnes, M. R. Rahman, and W. Maass. Semantic Priming via Knowledge graphs to analyze and treat language model's Honest Lies. In *ICIS*. Association for Information Systems, 2024.

[3] G. Agrawal, T. Kumarage, Z. Alghamdi, and H. Liu. 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. ACL, 2024.

[4] U. Ali, S. Lynden, A. Matono, and T. Amagasa. Entropy-Guided Probing for Predicting LLM Hallucinations with Knowledge Graph Features. In *Database and Expert Systems Applications*, pages 68–82. Springer Nature Switzerland, 2026.

[5] R. Aly, Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, and A. Mittal. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[6] A. Amelio, C. Buratti, M. Marchetti, D. Traini, D. Ursino, and L. Virgili. Exploiting knowledge graph communities to fine-tune large language models. *Expert Systems with Applications*, 298:129816, 2026.

[7] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[8] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*, pages 722–735. Springer, 2007.

[9] I. Augenstein, T. Baldwin, M. Cha, T. Chakraborty, G. L. Ciampaglia, D. Corney, R. DiResta, E. Ferrara, S. Hale, A. Halevy, et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863, 2024.

[10] J. Baek, A. F. Aji, and A. Saffari. Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 78–106. ACL, 2023.

[11] S. Banerjee, A. Agarwal, and S. Singla. LLMs Will Always Hallucinate, and We Need to Live with This. In *Intelligent Systems and Applications*, pages 624–648. Springer Nature Switzerland, 2025.

[12] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718. ACL, 2023.

[13] F. S. Bao, M. Li, R. Qu, G. Luo, E. Wan, Y. Tang, W. Fan, M. S. Tamber, S. Kazi, V. Sourabh, et al. FaithBench: A Diverse Hallucination Benchmark for Summarization by Modern LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 448–461. ACL, 2025.

[14] P. Belcak, G. Heinrich, S. Diao, Y. Fu, X. Dong, S. Muralidharan, Y. C. Lin, and P. Molchanov. Small language models are the future of agentic ai. *arXiv*, abs.2506.02153, 2025.

[15] Y. Bengio. From System 1 Deep Learning to System 2 Deep Learning. In *Neural Information Processing Systems – Invited Talk (Posner Lecture)*, 2019.

[16] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[17] O. Bodenreider. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004.

[18] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. 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. ACM, 2008.

[19] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. ACL, 2015.

[20] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language Models are Few-Shot Learners. In *NeurIPS*, 2020.

[21] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv*, abs:2303.12712, 2023.

[22] N. D. Cao, G. Izacard, S. Riedel, and F. Petroni. Autoregressive Entity Retrieval. In *International Conference on Learning Representations*, 2020.

[23] Y. Cao, Y. Kang, C. Wang, and L. Sun. Instruction Mining: Instruction Data Selection for Tuning Large Language Models. In *First Conference on Language Modeling*, 2024.

[24] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka, and T. Mitchell. Toward an Architecture for Never-Ending Language Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1):1306–1313, 2010.

[25] S. Chaudhari, P. Aggarwal, V. Murahari, T. Rajpurohit, A. Kalyan, K. Narasimhan, A. Deshpande, and B. Castro da Silva. RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. *ACM Computing Surveys*, 58(2):53:1–53:37, 2025.

[26] L. Chen, P. Tong, Z. Jin, Y. Sun, J. Ye, and H. Xiong. Plan-on-Graph: Self-Correcting Adaptive Planning of Large Language Model on Knowledge Graphs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[27] Q. Chen, W. Wu, and S. Li. Exploring In-Context Learning for Knowledge Grounded Dialog Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10071–10081. ACL, 2023.

[28] S. Chen, Y. Zhao, J. Zhang, I.-C. Chern, S. Gao, P. Liu, and J. He. FELM: Benchmarking factuality evaluation of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 44502–44523. Curran Associates Inc., 2023.

[29] Y. Chen, J. Benton, A. Radhakrishnan, J. Uesato, C. Denison, J. Schulman, A. Somani, P. Hase, M. Wagner, F. Roger, et al. Reasoning Models Don't Always Say What They Think. *arXiv*, abs:2505.05410, 2025.

[30] Z. Chen, H. Mao, H. Li, W. Jin, H. Wen, X. Wei, S. Wang, D. Yin, W. Fan, H. Liu, et al. Exploring the Potential of Large Language Models (LLMs)in Learning on Graphs. *SIGKDD Explor. Newsl.*, 25(2):42–61, 2024.

[31] Z. Chen, Y. Wang, B. Zhao, J. Cheng, X. Zhao, and Z. Duan. Knowledge Graph Completion: A Review. *IEEE Access*, 8:192435–192456, 2020.

[32] Y. Deng, C. Ye, Z. Huang, M. D. Ma, Y. Kou, and W. Wang. GraphVis: Boosting LLMs with Visual Knowledge Graph Integration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[33] S. Dernbach, K. Agarwal, A. Zuniga, M. Henry, and S. Choudhury. GLaM: Fine-Tuning Large Language Models for Domain Knowledge Graph Alignment via Neighborhood Partitioning and Generative Subgraph Encoding. *Proceedings of the AAAI Symposium Series*, 3(1):82–89, 2024.

[34] A. Deshpande, D. Ruiter, M. Mosbach, and D. Klakow. StereoKG: Data-Driven Knowledge Graph Construction For Cultural Knowledge and Stereotypes. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 67–78. ACL, 2022.

[35] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186. ACL, 2019.

[37] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

[38] T. Dolci, M. Jovanovik, and K. Hose. Towards LLM-KG Symbiosis for Reducing Factual Hallucinations. In *Proceedings of the Workshops of the EDBT/ICDT Joint Conference*, volume 4192. CEUR-WS, 2026.

[39] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, et al. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128. ACL, 2024.

[40] K. Donnelly. SNOMED-CT: The advanced terminology and coding system for ehealth. *Studies in Health Technology and Informatics*, 121:279–290, 2006.

[41] J. Duan, H. Cheng, S. Wang, A. Zavalny, C. Wang, R. Xu, B. Kailkhura, and K. Xu. Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063. ACL, 2024.

[42] N. Dziri, A. Madotto, O. Zaïane, and A. J. Bose. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214. ACL, 2021.

[43] O. El Khatib. Reasoning over Incomplete Knowledge Graphs. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pages 6785–6788. ACM, 2025.

[44] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, and Y. Goldberg. Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021.

[45] H. Elsahar, P. Vougiouklis, A. Remaci, C. Gravier, J. Hare, F. Laforest, and E. Simperl. 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). European Language Resources Association (ELRA), 2018.

[46] S. Es, J. James, L. Espinosa Anke, and S. Schockaert. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158. ACL, 2024.

[47] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[48] F. Fatahi Bayat, K. Qian, B. Han, Y. Sang, A. Belyy, S. Khorshidi, F. Wu, I. Ilyas, and Y. Li. FLEEK: Factual Error Detection and Correction with Evidence Retrieved from External Knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 124–130. ACL, 2023.

[49] B. Fatemi, J. Halcrow, and B. Perozzi. Talk like a Graph: Encoding Graphs for Large Language Models. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024.

[50] S. Feng, G. Fang, X. Ma, and X. Wang. Efficient Reasoning Models: A Survey. *Transactions on Machine Learning Research*, 2025.

[51] K. Furumai, Y. Wang, M. Shinohara, K. Ikeda, Y. Yu, and T. Kato. Detecting Dialogue Hallucination Using Graph Neural Networks. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 871–877, 2023.

[52] S. Geng, M. Josifoski, M. Peyrard, and R. West. Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952. ACL, 2023.

[53] GeoNames. The GeoNames geographical database. `https://www.geonames.org/`. Accessed: June 22, 2026.

[54] B. Ghosh, S. Hasan, N. A. Arafat, and A. Khan. Logical Consistency of Large Language Models in Fact-Checking. In *The Thirteenth International Conference on Learning Representations*. OpenReview.net, 2025.

[55] T. Goyal and G. Durrett. Annotating and Modeling Fine-grained Factuality in Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462. ACL, 2021.

[56] X. Guan, Y. Liu, H. Lin, Y. Lu, B. He, X. Han, and L. Sun. Mitigating Large Language Model Hallucinations via Autonomous Knowledge Graph-Based Retrofitting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18126–18134, 2024.

[57] K. Guo, H. Shomer, S. Zeng, H. Han, Y. Wang, and J. Tang. Empowering GraphRAG with Knowledge Filtering and Integration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25450–25464. ACL, 2025.

[58] H. Han, Y. Wang, H. Shomer, K. Guo, J. Ding, Y. Lei, M. Halappanavar, R. A. Rossi, S. Mukherjee, X. Tang, et al. Retrieval-Augmented Generation with Graphs (GraphRAG). *arXiv*, abs:2501.00309, 2025.

[59] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. *Transactions on Machine Learning Research*, 2024.

[60] R. Hasegawa and R. Ichise. CoKGLM: Detecting Hallucinations Generated by Large Language Models via Knowledge Graph Verification. In *Knowledge Graphs and Semantic Web*, pages 212–224. Springer Nature Switzerland, 2025.

[61] R. Haskins and B. Adams. KEA Explain: Explanations of Hallucinations using Graph Kernel Analysis. In *Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning*, pages 1041–1058. PMLR, 2025.

[62] P. He, X. Liu, J. Gao, and W. Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*, 2020.

[63] D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian, and S. E. Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726, 2017.

[64] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al. Knowledge Graphs. *ACM Comput. Surv.*, 54(4):71:1–71:37, 2021.

[65] Y. Hou, W. Jiao, M. Liu, C. Allen, Z. Tu, and M. Sachan. Adapters for Enhanced Modeling of Multilingual Knowledge and Text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3902–3917. ACL, 2022.

[66] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. D. Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[67] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2021.

[68] S. Hu, Y. Luo, H. Wang, X. Cheng, Z. Liu, and M. Sun. Won't Get Fooled Again: Answering Questions with False Premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5626–5643. ACL, 2023.

[69] X. Hu, J. Chen, X. Li, Y. Guo, L. Wen, P. S. Yu, and Z. Guo. Towards Understanding Factual Knowledge of Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2023.

[70] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55, 2025.

[71] W. Huang, G. Zhou, M. Lapata, P. Vougiouklis, S. Montella, and J. Z. Pan. Prompting large language models with knowledge graphs for question answering involving long-tail facts. *Knowledge-Based Systems*, 324:113648, 2025.

[72] N. Ibrahim, S. Aboulela, A. Ibrahim, and R. Kashef. A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): Models, evaluation metrics, benchmarks, and challenges. *Discover Artificial Intelligence*, 4(1):76, 2024.

[73] T. Ishiwatari, Y. Yasuda, T. Miyazaki, and J. Goto. Relation-aware Graph Attention Networks with Relational Position Encodings for Emotion Recognition in Conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370. ACL, 2020.

[74] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, 2023.

[75] Z. Ji, Z. Liu, N. Lee, T. Yu, B. Wilie, M. Zeng, and P. Fung. RHO: Reducing Hallucination in Open-domain Dialogues with Knowledge Grounding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4504–4522. ACL, 2023.

[76] J. Jiang, F. Wang, J. Shen, S. Kim, and S. Kim. A Survey on Large Language Models for Code Generation. *ACM Trans. Softw. Eng. Methodol.*, 35(2):58:1–58:72, 2026.

[77] J. Jiang, K. Zhou, Z. Dong, K. Ye, X. Zhao, and J.-R. Wen. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251. ACL, 2023.

[78] J. Jiang, K. Zhou, W. X. Zhao, Y. Song, C. Zhu, H. Zhu, and J.-R. Wen. KG-Agent: An Efficient Autonomous Agent Framework for Complex Reasoning over Knowledge Graph. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9505–9523. ACL, 2025.

[79] D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[80] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang. Why Language Models Hallucinate. *arXiv*, abs:2509.04664, 2025.

[81] M. P. Karpowicz. On the Fundamental Impossibility of Hallucination Control in Large Language Models. *arXiv*, abs:2506.06382, 2025.

[82] J. Kim, S. Park, Y. Kwon, Y. Jo, J. Thorne, and E. Choi. FactKG: Fact Verification via Reasoning on Knowledge Graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206. ACL, 2023.

[83] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, 2022.

[84] A. Kraft and R. Usbeck. The Lifecycle of "Facts": A Survey of Social Bias in Knowledge Graphs. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 639–652. ACL, 2022.

[85] N. M. Kriege, F. D. Johansson, and C. Morris. A survey on graph kernels. *Applied Network Science*, 5(1):6, 2020.

[86] A. Kulkarni, Y. Zhang, J. R. A. Moniz, X. Ge, B.-H. Tseng, D. Piraviperumal, S. Swayamdipta, and H. Yu. Evaluating Evaluation Metrics – The Mirage of Hallucination Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19013–19032. ACL, 2025.

[87] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022.

[88] P.-C. Langlais, P. Chizhov, C. Arnett, C. R. Hinostroza, M. Nee, E. K. Jones, I. Girard, D. Mach, A. Stasenko, and I. P. Yamshchikov. Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training. In *The Fourteenth International Conference on Learning Representations*, 2025.

[89] E. Lavrinovics, R. Biswas, J. Bjerva, and K. Hose. Knowledge Graphs, Large Language Models, and Hallucinations: An NLP Perspective. *Journal of Web Semantics*, 85:100844, 2025.

[90] E. Lavrinovics, R. Biswas, K. Hose, and J. Bjerva. MultiHal: Multilingual Dataset for Knowledge-Graph Grounded Evaluation of LLM Hallucinations. *arXiv*, abs:2505.14101, 2025.

[91] J. Lee, Y. Wang, J. Li, and M. Zhang. Multimodal Reasoning with Multimodal Knowledge Graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10767–10782. ACL, 2024.

[92] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[93] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, et al. DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195, 2015.

[94] D. B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[95] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. ACL, 2020.

[96] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 9459–9474. Curran Associates Inc., 2020.

[97] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464. ACL, 2023.

[98] S. Li, X. Li, L. Shang, C. Sun, B. Liu, Z. Ji, X. Jiang, and Q. Liu. Pre-training Language Models with Deterministic Factual Knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11118–11131. ACL, 2022.

[99] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252. ACL, 2022.

[100] G. Liu, Y. Zhang, Y. Li, and Q. Yao. Dual Reasoning: A GNN-LLM Collaborative Framework for Knowledge Graph Question Answering. In *The Second Conference on Parsimony and Learning (Proceedings Track)*, 2025.

[101] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, 55(9):195:1–195:35, 2023.

[102] Y. Liu, J. Cao, C. Liu, K. Ding, and L. Jin. Datasets for large language models: A comprehensive survey. *Artificial Intelligence Review*, 58(12):403, 2025.

[103] Y. Liu, J. Ding, Y. Fu, and Y. Li. UrbanKG: An Urban Knowledge Graph System. *ACM Trans. Intell. Syst. Technol.*, 14(4):60:1–60:25, 2023.

[104] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. ACL, 2023.

[105] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, abs:1907.11692, 2019.

[106] S. Longpre, K. Perisetla, A. Chen, N. Ramesh, C. DuBois, and S. Singh. Entity-Based Knowledge Conflicts in Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063. ACL, 2021.

[107] H. Lu, S. Bao, H. He, F. Wang, H. Wu, and H. Wang. Towards Boosting the Open-Domain Chatbot with Human Feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4060–4078. ACL, 2023.

[108] Y. Lu, W. Wu, X. Zhao, R. Peng, and J. Wang. KARMA: Leveraging Multi-Agent LLMs for Automated Knowledge Graph Enrichment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[109] 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*, 2023.

[110] L. Luo, T. Vu, D. Phung, and R. Haf. Systematic Assessment of Factual Knowledge in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13272–13286. ACL, 2023.

[111] L. Luo, Z. Zhao, G. Haffari, Y.-F. Li, C. Gong, and S. Pan. Graph-constrained Reasoning: Faithful Reasoning on Knowledge Graphs with Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 41540–41565. PMLR, 2025.

[112] C. Ma, Y. Chen, T. Wu, A. Khan, and H. Wang. Large Language Models Meet Knowledge Graphs for Question Answering: Synthesis and Opportunities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24589–24608. ACL, 2025.

[113] J. Ma, Z. Gao, Q. Chai, W. Sun, P. Wang, H. Pei, J. Tao, L. Song, J. Liu, C. Zhang, et al. Debate on Graph: A Flexible and Reliable Reasoning Framework for Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24768–24776, 2025.

[114] P. Manakul, A. Liusie, and M. Gales. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017. ACL, 2023.

[115] A. Martino, M. Iannelli, and C. Truong. Knowledge Injection to Counter Large Language Model (LLM) Hallucination. In *The Semantic Web: ESWC 2023 Satellite Events*, pages 182–185. Springer Nature Switzerland, 2023.

[116] C. Metz and K. Weise. A.I. Is Getting More Powerful, but Its Hallucinations Are Getting Worse. *The New York Times*, 2025. Accessed: June 22, 2026.

[117] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100. ACL, 2023.

[118] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, et al. Never-ending learning. *Communications of the ACM*, 61(5):103–115, 2018.

[119] R. Navigli, S. Conia, and B. Ross. Biases in Large Language Models: Origins, Inventory, and Discussion. *ACM Journal of Data and Information Quality*, 15(2):10:1–10:21, 2023.

[120] M.-V. Nguyen, L. Luo, F. Shiri, D. Phung, Y.-F. Li, T.-T. Vu, and G. Haffari. Direct Evaluation of Chain-of-Thought in Multi-hop Reasoning with Knowledge Graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2862–2883. ACL, 2024.

[121] C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, and T. Zhang. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878. ACL, 2024.

[122] OpenAI. GPT-4 Technical Report. *arXiv*, abs:2303.08774, 2023.

[123] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, et al. GPT-4 Technical Report. *arXiv*, abs:2303.08774, 2024.

[124] D. Orr. 50,000 Lessons on How to Read: A Relation Extraction Corpus. `https://research.google/blog/50000-lessons-on-how-to-read-a-relation-extraction-corpus/`, Apr 2013.

[125] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 27730–27744. Curran Associates Inc., 2022.

[126] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021.

[127] J. Z. Pan, S. Razniewski, J.-C. Kalo, S. Singhania, J. Chen, S. Dietze, H. Jabeen, J. Omeliyanenko, W. Zhang, M. Lissandrini, et al. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Transactions on Graph Data and Knowledge*, 1(1):2:1–2:38, 2023.

[128] 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.

[129] B. Peng, C. Li, P. He, M. Galley, and J. Gao. Instruction Tuning with GPT-4. *arXiv*, abs:2304.03277, 2023.

[130] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, and S. Tang. Graph Retrieval-Augmented Generation: A Survey. *ACM Trans. Inf. Syst.*, 44(2):35:1–35:52, 2025.

[131] R. Pozzi, M. Palmonari, A. Coletta, L. Bellomarini, J. Lehmann, and S. Vahdati. ReFactX: Scalable Reasoning with Reliable Facts via Constrained Generation. In *The Semantic Web – ISWC 2025*, pages 290–308. Springer Nature Switzerland, 2026.

[132] S. Prabhumoye, A. W. Black, and R. Salakhutdinov. Exploring Controllable Text Generation Techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14. ICCL, 2020.

[133] J. Priem, H. Piwowar, and R. Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv*, abs:2205.01833, 2022.

[134] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[135] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[136] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[137] E. Rajabi and K. Etminani. Knowledge-graph-based explainable AI: A systematic review. *Journal of Information Science*, 50(4):1019–1029, 2024.

[138] M. Rashad, A. Zahran, A. Amin, A. Abdelaal, and 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)*, pages 79–84. ACL, 2024.

[139] A. Razdaibiedina, Y. Mao, R. Hou, M. Khabsa, M. Lewis, and A. Almahairi. Progressive Prompts: Continual Learning for Language Models. In *The Eleventh International Conference on Learning Representations*, 2022.

[140] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. ACL, 2019.

[141] L. F. R. Ribeiro, M. Liu, I. Gurevych, M. Dreyer, and M. Bansal. FactGraph: Evaluating Factuality in Summarization with Semantic Graph Representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253. ACL, 2022.

[142] H. Sansford, N. Richardson, H. P. Maretic, and J. N. Saada. GraphEval: A Knowledge-Graph Based LLM Hallucination Evaluation Framework. In *KiL@KDD*, volume 3894 of *CEUR Workshop Proceedings*, pages 20–31. CEUR-WS.org, 2024.

[143] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035, 2019.

[144] A. Sarkar. Gluing Pizza, Eating Rocks, and Counting Rs in Strawberry: The Discursive Social Function of Stupid AI Answers. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, CHIWORK '25, pages 1–12. ACM, 2025.

[145] R. Sarkar, M. Arcan, and J. P. McCrae. KG-CRuSE: Recurrent Walks over Knowledge Graph for Explainable Conversation Reasoning using Semantic Embeddings. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 98–107. ACL, 2022.

[146] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilic, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *CoRR*, abs/2211.05100, 2022.

[147] Z. Shang, W. Ke, N. Xiu, P. Wang, J. Liu, Y. Li, Z. Luo, and K. Ji. OntoFact: Unveiling Fantastic Fact-Skeleton of LLMs via Ontology-Driven Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18934–18943, 2024.

[148] T. Shen, Y. Mao, P. He, G. Long, A. Trischler, and W. Chen. Exploiting Structured Knowledge in Text via Graph-Guided Representation Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8980–8994. ACL, 2020.

[149] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt. Weisfeiler-Lehman Graph Kernels. *Journal of Machine Learning Research*, 12(77):2539–2561, 2011.

[150] A. Singhal. Introducing the Knowledge Graph: Things, not strings. Google Official Blog, 2012.

[151] R. Speer, J. Chin, and C. Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), 2017.

[152] Z. Su, H. Song, Y. Lin, Y. Wu, X. Weng, Z. Mao, B. Liu, H. Yin, and J. Yang. MedKit: Multi-level feature distillation with knowledge injection for radiology report generation. *Expert Systems with Applications*, 296:129003, 2026.

[153] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706. ACM, 2007.

[154] Y. Sui, Y. He, Z. Ding, and B. Hooi. Can Knowledge Graphs Make Large Language Models More Trustworthy? An Empirical Study Over Open-ended Question Answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12685–12701. ACL, 2025.

[155] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, L. Ni, H.-Y. Shum, and J. Guo. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. In *The Twelfth International Conference on Learning Representations*, 2023.

[156] M. Sung, J. Lee, S. Yi, M. Jeon, S. Kim, and J. Kang. Can Language Models be Biomedical Knowledge Bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734. ACL, 2021.

[157] L. Tang, I. Shalyminov, A. Wong, J. Burnsky, J. Vincent, Y. Yang, S. Singh, S. Feng, H. Song, H. Su, et al. TofuEval: Evaluating Hallucinations of LLMs on Topic-Focused Dialogue Summarization. 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 4455–4480. ACL, 2024.

[158] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: A Family of Highly Capable Multimodal Models. *arXiv*, abs:2312.11805, 2025.

[159] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv*, abs:2403.05530, 2024.

[160] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al. Gemma: Open Models Based on Gemini Research and Technology. *arXiv*, abs:2403.08295, 2024.

[161] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. FEVER: A Large-scale Dataset for Fact Extraction and VERification. 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 809–819. ACL, 2018.

[162] K. Tian, E. Mitchell, H. Yao, C. D. Manning, and C. Finn. Fine-Tuning Language Models for Factuality. In *The Twelfth International Conference on Learning Representations*, 2023.

[163] S. Tian, Y. Luo, T. Xu, C. Yuan, H. Jiang, C. Wei, and X. Wang. KG-Adapter: Enabling Knowledge Graph Integration in Large Language Models through Parameter-Efficient Fine-Tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3813–3828. ACL, 2024.

[164] A. Toroghi, W. Guo, and S. Sanner. CoLoTa: A Dataset for Entity-based Commonsense Reasoning over Long-Tail Knowledge. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pages 3444–3454. ACM, 2025.

[165] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv*, abs:2302.13971, 2023.

[166] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv*, abs:2307.09288, 2023.

[167] M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

[168] E. Vakaj, N. Mihindukulasooriya, M. Gaur, and A. Khan. Knowledge Graphs for Responsible AI. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pages 5596–5598. ACM, 2024.

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

[170] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.

[171] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph Attention Networks. In *International Conference on Learning Representations*, 2018.

[172] D. Vrandečić and M. Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.

[173] R. Wagner, E. Kitzelmann, and I. Boersch. Mitigating Hallucination by Integrating Knowledge Graphs into LLM Inference – a Systematic Literature Review. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 795–805. ACL, 2025.

[174] K. Wang, W. Shen, Y. Yang, X. Quan, and R. Wang. Relational Graph Attention Network for Aspect-based Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238. ACL, 2020.

[175] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, J. Ji, G. Cao, D. Jiang, and M. Zhou. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418. ACL, 2021.

[176] S. Wang, J. Feng, T. Liu, D. Pei, and Y. Li. Mitigating Geospatial Knowledge Hallucination in Large Language Models: Benchmarking and Dynamic Factuality Aligning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 870–888. ACL, 2025.

[177] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021.

[178] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*, 2022.

[179] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, 2022.

[180] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, 2022.

[181] J. Wei, C. Yang, X. Song, Y. Lu, N. Z. Hu, J. Huang, D. Tran, D. Peng, R. Liu, D. Huang, et al. Long-form factuality in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[182] Y. Wen, Z. Wang, and J. Sun. MindMap: Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10370–10388. ACL, 2024.

[183] S. Wiegreffe and Y. Pinter. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20. ACL, 2019.

[184] T. S. Winter and J. H. van Vuuren. A Knowledge Graph Approach Towards Detecting Large Language Model Hallucination. In *Modelling, Computation and Optimization in Information Systems and Management Sciences*, pages 224–237. Springer Nature Switzerland, 2026.

[185] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021.

[186] W. Xiong, J. Du, W. Y. Wang, and V. Stoyanov. Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model. In *International Conference on Learning Representations*, 2019.

[187] R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, and W. Xu. Knowledge Conflicts for LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565. ACL, 2024.

[188] Y. Xu, S. He, J. Chen, Z. Wang, Y. Song, H. Tong, G. Liu, J. Zhao, and K. Liu. Generate-on-Graph: Treat LLM as both Agent and KG for Incomplete Knowledge Graph Question Answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18410–18430. ACL, 2024.

[189] Z. Xu, S. Jain, and M. Kankanhalli. Hallucination is Inevitable: An Innate Limitation of Large Language Models. *arXiv*, abs:2401.11817, 2025.

[190] L. Yang, H. Chen, Z. Li, X. Ding, and X. Wu. Give us the Facts: Enhancing Large Language Models With Knowledge Graphs for Fact-Aware Language Modeling. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3091–3110, 2024.

[191] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. R. Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

[192] Y. Yao, Z. Li, and H. Zhao. GoT: Effective Graph-of-Thought Reasoning in Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2901–2921. ACL, 2024.

[193] H. Ye, N. Zhang, S. Deng, X. Chen, H. Chen, F. Xiong, X. Chen, and H. Chen. Ontology-enhanced Prompt-tuning for Few-shot Learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 778–787. ACM, 2022.

[194] J. Yu, S. Wu, J. Chen, and W. Zhou. LLMs as Collaborator: Demands-Guided Collaborative Retrieval-Augmented Generation for Commonsense Knowledge-Grounded Open-Domain Dialogue Systems. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13586–13612. ACL, 2024.

[195] H. Yuan, P. Cao, Z. Jin, Y. Chen, D. Zeng, K. Liu, and J. Zhao. Whispers that Shake Foundations: Analyzing and Mitigating False Premise Hallucinations in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2670–2683. ACL, 2024.

[196] W. Yuan, G. Neubig, and P. Liu. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems*, 2021.

[197] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song. A Survey of Controllable Text Generation Using Transformer-based Pre-trained Language Models. *ACM Computing Surveys*, 56(3):1–37, 2024.

[198] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2019.

[199] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *Computational Linguistics*, pages 1–46, 2025.

[200] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451. ACL, 2019.

[201] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du. Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.*, 15(2):20:1–20:38, 2024.

[202] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[203] D. Zhou, Y. Zhu, X. Wang, Y. He, J. Chen, S. Staab, and E. Kharlamov. Evaluating Knowledge Graph Based Retrieval Augmented Generation Methods under Knowledge Incompleteness. *arXiv*, abs:2504.05163, 2025.

[204] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781. ACL, 2024.

[205] Y. Zhu, S. Qiao, Y. Ou, S. Deng, S. Lyu, Y. Shen, L. Liang, J. Gu, H. Chen, and N. Zhang. KnowAgent: Knowledge-Augmented Planning for LLM-Based Agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3709–3732. ACL, 2025.

[206] Y. Zhu, J. Xiao, Y. Wang, and J. Sang. KG-FPQ: Evaluating Factuality Hallucination in LLMs with Knowledge Graph-based False Premise Questions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10472–10490. ACL, 2025.

[207] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, H. Chen, Z. Liu, Z. Dou, and J.-R. Wen. Large Language Models for Information Retrieval: A Survey. *ACM Trans. Inf. Syst.*, 44(1):12:1–12:54, 2025.

Literature Review Summary

Table 5 provides a summary of the analyzed KG-based methods. Figure 14 and Figure 15 present an overview of statistics related to the selected articles.

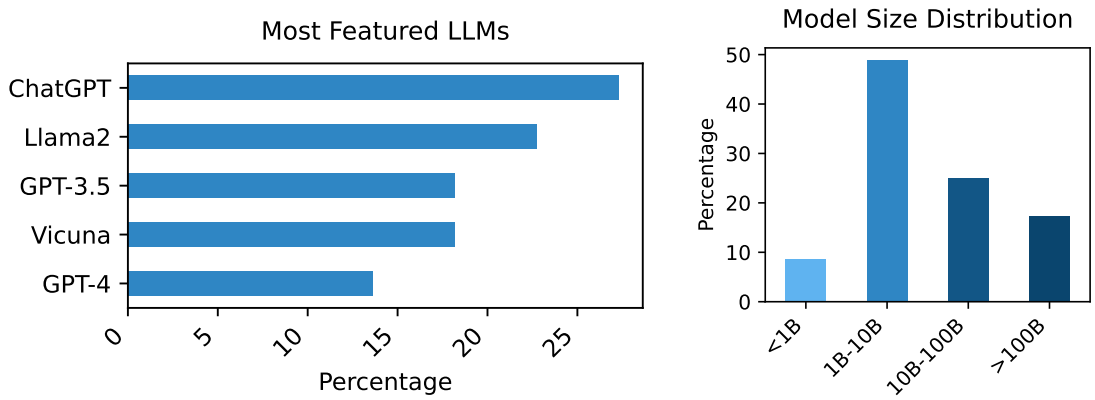

Figure 14: Most featured LLMs and distribution of LLM size across the selected articles.

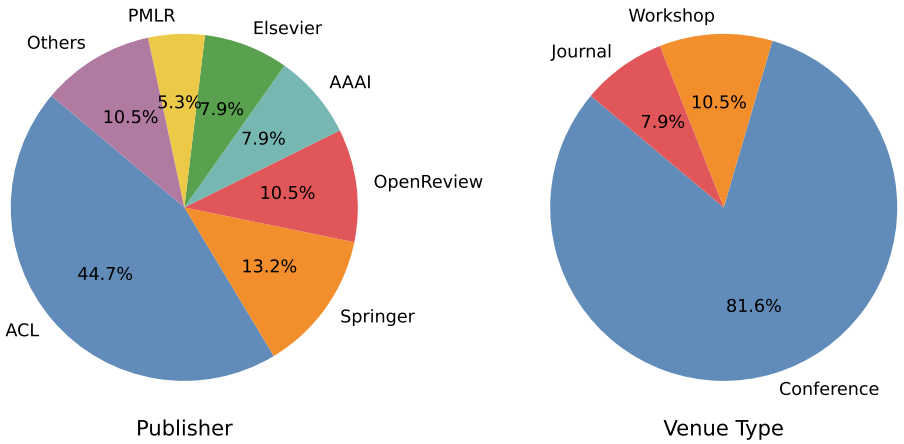

Figure 15: Distribution of selected articles by publisher and venue type.

Table 5: Summary of KG-based methods for hallucination detection and mitigation.

Solution	Year	Venue	Type	Goal	Stage	Hal. Type	Repo	LLM Tested	Benchmarks	Metrics	Supporting KG	KG Type	Eff. Eval
GLaM [33]	2024	AAAI-SS	W	M	Pre	Fact	-	Llama	UMLS-QA*, DBLP-QA*	Acc, Prec, Recall, F1, BertScore	UMLS, DBLP	D	N
CoFine [6]	2026	ESWA	J	M	Pre	Fact	C	Minerva, Ministral	Link Prediction*	Hits@K, MRR, Prec, Recall, F1	YAGO3, PrimeKG, WN18RR	E, D	Y
KG-Adapter [163]	2024	ACL	C	M	Pre	Fact	-	Llama2, Zephyr	OpenBookQA, WebQSP, CWQ, CommonsenseQA	Acc, Hits@1	ConceptNet, Freebase	E, C	N
LFC [54]	2025	ICLR	C	M	Pre	Fact	C	LLama2, Gemma	NELLLFC*, WikiLFC*, FreebaseLFC*	Acc, LogConsistency*	Freebase, NELL, Wikidata	E	Y
OntoFact [147]	2024	AAAI	C	P	Pre	Fact	C	Llama, Vicuna, Alpaca, T0pp, OPT, Bloom, GPT2, FLAN-T5, ChatGPT, GPTNeo, GPT-J	-	ErrorProportion*	DBPedia, BIOS 2.2, CN-DBPedia, YAGO	E, D	N
Entropy-Probe [4]	2025	DEXA	C	P	Pre	Fact	-	LLama3.1, Stable-LM, Mistral, Qwen2.5, QwQ	WikiWebQuestions	Acc	Wikidata	E	N
KAPING [10]	2023	NLRSE	W	M	In	Fact	-	T5, T0, OPT, GPT3	WebQSP, Mintaka	Acc	Wikidata, Freebase	E	Y
IKA [27]	2023	EMNLP	C	M	In	Fact + Faith	-	GPT2, GPT3.5, GPT4, Vicuna, GPT-J	OpenDialKG	EM, BLEU, ROUGE-L, BARTScore, FeQA	Freebase	E	N
KI [115]	2023	ESWC	C	M	In	Fact	-	BLOOM	-	% Factual Claims*	private	D	N
MedKit [152]	2026	ESWA	J	M	In	Fact	C	LLama2, Qwen2, MedLLama	MIMIC-CXR, IU-Xray, Liver-CT	BLEU, ROUGE-L, METEOR, CIDEr, BertScore, RadGraph F1	private	D	N
DualEval [2]	2024	ICIS	C	D, M	In	Fact	-	GPT4	DualSet*	DualMatrix*	DBPedia	E	N
DualR [100]	2025	CPAL	C	M	In	Fact	C	Llama2, ChatGPT, GPT4	WebQSP, CWQ, MetaQA	Hits@1	Freebase, WikiMovies	E, D	Y
MindMap [182]	2024	ACL	C	M	In	Fact	C	GPT3.5	GenMedGPT-5k, CMCQA, ExplainCPE	BertScore, LLM-Judge, UHGEval	EMCKG*, CMCKG*	D	N
DCRAG [194]	2024	EMNLP	C	M	In	Fact + Faith	-	GPT3.5, LLama3	DailyDialog, Diamante	DI-2, CDP*, CDF*, LLM-Judge	ConceptNet	C	N
GraphVis [32]	2024	NeurIPS	C	M	In	Fact	C	LLaVA-v1.6-Mistral	OpenBookQA, MMBench, OpenScienceQA, POPE, CommonsenseQA	Acc, F1	ConceptNet	C	N
GCR [111]	2026	ICML	C	M	In	Fact	C	ChatGPT, GPT4o-mini	WebQSP, CWQ, FreebaseQA, CSQA, MedQA	Acc, Hits@K, F1	Freebase, UMLS, ConceptNet	E, D, C	Y
ReFactX [131]	2025	ISWC	C	M	In	Fact	C	Llama3.3, Phi-4, Qwen2.5	Mintaka, 2WikiMultiHopQA, WebQSP, private	EM, LLM-Judge	Wikidata	E	Y
RHO [75]	2023	ACL	C	M	In	Fact + Faith	C	BART	OpenDialKG	EM, BLEU, ROUGE-L, FeQA, QuestEval	Freebase	E	N
MR-MKG [91]	2024	ACL	C	M	In	Fact	-	LLama2, FLAN-T5, FLAN-UL2	ScienceQA, MARS	Acc, Hits@K, MRR	MMKG, MarKG	E, M	N
KEA [61]	2025	NeSy	C	D	Post	Fact + Faith	C	-	SummEval, QAGS-C, WikiBio-GPT3	Acc, Prec, Recall, F1	Wikidata	E	N
RGAT [51]	2023	ICML A	C	D	Post	Fact	-	GODEL, T5, ChatGPT, RoBERTa Enc-Dec	OpenDialKG test set	Acc, Prec, Recall, F1	Freebase	E	N
FactGraph [141]	2022	NAACL	C	D	Post	Faith	C	-	FactCollect*	BACC, Micro F1	Runtime generated, real KG usable	-	N
FactAlign [138]	2024	TrustNLP	W	D	Post	Fact + Faith	-	-	WikiBio-GPT3	Prec, Recall, F1	Runtime generated, real KG usable	-	N
CoKGLM [60]	2024	KGSWC	C	D	Post	Fact	-	-	OpenDialKG	Prec, Recall, F1	Freebase	E	N
GraphEval [142]	2024	KiL	W	M	Post	Faith	-	-	SummEval, QAGS-C, QAGS-X	Acc	Runtime generated, real KG usable	-	N
GLLM [184]	2025	MCO	C	D	Post	Fact	-	Phi-4	-	NLI-Acc*	DBPedia50k	E	N
FLEEK [48]	2023	EMNLP	C	D, M	Post	Fact	-	GPT3, Vicuna	BenchLLM*, BenchText*	Prec, Recall, F1	Wikidata	E	N
NPH [42]	2021	EMNLP	C	M	Post	Fact	C	GPT2-KG, GPT2-KE, DialGPT	OpenDialKG	BLEU, FeQA, Hallucination Critic*	Freebase	E	N
KGR [56]	2024	AAAI	C	M	Post	Fact	-	ChatGPT, GPT3.5, Vicuna	Simple Question, Mintaka, HotpotQA	EM, F1	Wikidata	E	N
GraphRAG-FI [57]	2025	EMNLP	C	M	Post	Fact	C	Alpaca, LLama2, Llama3.1, ChatGPT, Qwen2.5	WebQSP, CWQ	Hits@1, F1	Freebase	E	N

* newly introduced. **Legend.** Venue Type: C = conference, J = journal, W = workshop. Goal: D = detection, M = mitigation. Hallucination Type: Fact = factuality hallucination, Faith = faithfulness hallucination. KG Type: E = encyclopedic KG, C = commonsense KG, D = domain-specific KG. Metrics: EM = Exact Match, Acc = Accuracy, Prec = Precision, MRR = Mean Reciprocal Rank.