



## OPEN ACCESS

## EDITED BY

Xinqing Xiao,  
China Agricultural University, China

## REVIEWED BY

Leonardo Nascimento,  
Federal Institute of Rio Grande do Sul, Brazil  
Joel Oduro-Afriyie,  
OriGen AI, United States

## \*CORRESPONDENCE

Yong Zhao  
✉ [yong.zhao@scupi.cn](mailto:yong.zhao@scupi.cn)

RECEIVED 10 March 2025  
ACCEPTED 25 June 2025  
PUBLISHED 16 July 2025

## CITATION

Cai L, Yu C, Kang Y, Fu Y, Zhang H and Zhao Y (2025) Practices, opportunities and challenges in the fusion of knowledge graphs and large language models. *Front. Comput. Sci.* 7:1590632.  
doi: 10.3389/fcomp.2025.1590632

## COPYRIGHT

© 2025 Cai, Yu, Kang, Fu, Zhang and Zhao. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Practices, opportunities and challenges in the fusion of knowledge graphs and large language models

Linyue Cai<sup>1</sup>, Chaojia Yu<sup>1</sup>, Yongqi Kang<sup>1</sup>, Yu Fu<sup>1</sup>, Heng Zhang<sup>2</sup> and Yong Zhao<sup>1\*</sup>

<sup>1</sup>Department of Computer Science, Pittsburgh Institute, Sichuan University, Chengdu, China,

<sup>2</sup>Department of Computer Science, Information and Artificial Intelligence Institute, Zhejiang University of Finance & Economics Dongfang College, Jiaxing, China

The fusion of Knowledge Graphs (KGs) and Large Language Models (LLMs) leverages their complementary strengths to address limitations of both technologies. This paper explores integration practices, opportunities, and challenges, focusing on three strategies: KG-enhanced LLMs (KEL), LLM-enhanced KGs (LEK), and collaborative LLMs and KGs (LKC). The study reviews these methodologies, highlighting their potential to enhance knowledge representation, reasoning, and question answering. We comprehensively compile and categorize key challenges such as knowledge acquisition and real-time updates, providing valuable directions for future research. The paper also discusses emerging techniques and applications to advance the synergy between KGs and LLMs. Overall, this work offers a comprehensive overview of the current landscape and the transformative potential of KG-LLM fusion across various domains.

## KEYWORDS

large language model, knowledge graph, KEL, LEK, LKC

## 1 Introduction

Language Models (LLMs), which are trained on extensive datasets, have demonstrated impressive advances in a wide range of natural language processing (NLP) tasks. The exponential growth in model size has endowed LLMs with emergent capabilities, enabling them to handle increasingly complex problems. Highly sophisticated LLMs equipped with billions of parameters have shown significant promise in handling complex, real-world tasks, including educational assistance, code generation, and recommendation systems. Despite their growing success, LLMs continue to face considerable criticism, particularly for their shortcomings in handling factual information.

Language Models (LLMs) rely heavily on memorizing facts from the vast amount of data they are trained on, but research has shown that they frequently struggle to retrieve these facts accurately, leading to what is commonly known as hallucination. This phenomenon involves LLMs generating responses that, while sounding plausible, are factually incorrect. [Zhang et al. \(2024b\)](#) conducted experiments on six main LLMs on the CoderEval dataset, elaborated on the hallucination phenomena, and analyzed the distribution of these phenomena in different models. This issue of factual inconsistency is especially problematic in sensitive applications. Moreover, LLMs, being black-box models, are often criticized for their lack of transparency ([Liao and Vaughan, 2023](#)). The knowledge they encode within their massive parameters is implicit and difficult to interpret or validate.

To mitigate these problems, a promising strategy is the integration of Knowledge Graphs (KGs) with LLMs. KGs store factual knowledge in a structured manner, typically in the form of a 3-tuple which contains head entity, relation, tail entity, and have long been valued for their precise and interpretable nature. By incorporating KGs, LLMs can benefit from a solid foundation of explicit knowledge that is both reliable and easily understood. Additionally, KGs excel at symbolic reasoning and evolve as new knowledge is discovered, making them well suited to providing the domain-specific information.

In recent years, increasing attention has been paid to unifying LLMs and KGs, as researchers and practitioners recognize their complementary strengths. On one side, KGs can be used to inject external knowledge during both the pre-training and inference phases of LLMs, offering an additional layer of factual grounding and improving interpretability. On the other side, LLMs have shown their utility in performing key tasks for KGs, such as KG embedding, completion, construction, and question answering, thereby enhancing the overall quality and applicability of KGs. A collaborative approach, wherein LLMs and KGs mutually reinforce each other, holds great potential for advancing knowledge representation and reasoning, combining the advantages of data-driven learning and structured knowledge. We observed that most existing surveys focus primarily on the use of KGs to enhance LLMs (KEL). Therefore, we aim to explore other potential possibilities of integrating the two, including how LLMs can contribute to KG-related tasks and their collaboration.

Our main contributions are summarized as follows:

1. Categorization and review. We present a detailed categorization and novel taxonomies of research on unifying LLMs and KGs. In each category, we review the research from the perspectives of different integration strategies and tasks, which provides more insights into each framework.
2. Coverage of emerging advancements. We cover the advanced techniques in both LLMs and KGs.
3. Summary of challenges and future directions: We highlight the challenges in existing research and present several promising future research directions.

The rest of this article is organized as follows. Section II first explains the background of LLMs and KGs. Section III presents the categorization and challenges of LLM-enhanced KGs. Section IV presents the categorization and challenges of KGs-enhanced LLM approaches. Section V presents the categorization of collaborative LLMs and KGs. Section VII discusses several applications. Finally, Section VIII concludes this paper.

## 2 Background

### 2.1 Large Language Models

Large Language Models (LLMs) represent a significant leap in the field of Natural Language Processing (NLP), primarily due to deep learning techniques. These models are trained in vast amounts of textual data, enabling them to understand, generate, and manipulate human language across various tasks.

LLMs use architectures like transformers (Vaswani, 2017), which handle context and capture long-range dependencies, facilitating the generation of human-like text. The development of LLMs has evolved from traditional rule-based and statistical models such as n-grams (Brown et al., 1992) and Hidden Markov Models (HMMs) (Rabiner and Juang, 1986), progressing through Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks (Sherstinsky, 2020). While RNNs and LSTMs helped handle sequential data, their limitations in managing long-range dependencies led to the creation of the transformer architecture, which now forms the basis of most modern LLMs.

#### 2.1.1 Encoder-only models

Encoder-only models focus primarily on understanding the input using bidirectional attention, making them ideal for tasks that require deep comprehension of text, such as classification, entity recognition, and reading comprehension. For instance, models like BERT, RoBERTa, and ALBERT, relying on masked language modeling and next sentence prediction, are widely used for a variety of NLP tasks, including question answering, sentiment analysis, and named entity recognition.

#### 2.1.2 Decoder-only models

Decoder-only models excel in generating sequences, such as sentences or paragraphs, by using unidirectional attention. These models are often auto-regressive, predicting the next token in a sequence based on previously generated tokens. Transformer models like GPT, OPT, and LLaMA employ decoder-only architectures to achieve high performance in text generation tasks such as chatbots, text summarization, and code generation.

#### 2.1.3 Encoder-decoder models

Encoder-decoder models (also called sequence-to-sequence models) are designed to transform one sequence into another, making them particularly effective for tasks like translation, summarization, and paraphrasing. These models use an encoder to process the input sequence and a decoder to generate the output sequence, often employing cross-attention to connect the two. Encoder-decoder models, like T5 and BART, are widely used in tasks like machine translation, text summarization, question answering, and dialoguesystems.

## 2.2 Knowledge graphs

A Knowledge Graph (KG) is a structured representation of knowledge that organizes information to highlight relationships between entities. This structure makes it easier for machines to understand and leverage the connections between data. KGs are pivotal in enabling better semantic search, data integration, and AI applications like question answering and recommendation systems.

### 2.2.1 Classification of knowledge graphs

KGs can be categorized into different types based on their content patterns, including encyclopedic, commonsense, domain-specific, and multi-modal KGs.

**Encyclopedic knowledge graphs** capture general knowledge across multiple domains, similar to encyclopedias.

**Commonsense knowledge graphs** capture everyday knowledge and reasoning, essential for enhancing AI's understanding of human-like reasoning.

**Domain-specific knowledge graphs** focus on specialized knowledge from specific domains like medicine, finance, or law.

**Multi-modal knowledge graphs** incorporate diverse data types such as text, images, and videos, provide a holistic understanding of knowledge across multiple forms of media.

### 2.2.2 Mechanism of knowledge graphs

Key Concepts in Knowledge Graphs:

1. **Entities (Nodes):** Primary objects or concepts, such as people, places, organizations, or events, represented as nodes.
2. **Relationships (Edges):** Connections between entities, specifying interactions.
3. **Attributes:** Properties or characteristics of entities, providing additional information.
4. **Triples:** Facts within a KG, represented as subject-predicate-object triples (e.g., "Barack Obama was born in Hawaii").
5. **Ontology:** The schema or structure of the KG, organizing entities, relationships, and attributes to ensure consistency.

## 2.3 The pros and cons of large language models and knowledge graphs

### Large Language Models (LLMs): Pros:

- Versatile across tasks (e.g., text generation, summarization, question-answering).
- Strong contextual understanding for coherent and nuanced language generation.
- Scalable and capable of handling diverse inputs.
- Zero-shot and few-shot learning capabilities.

### Cons:

- Lack of explicit knowledge structure, leading to hallucinations and factual inaccuracies.
- High data and computational intensity, making them expensive and environmentally taxing.
- Limited interpretability, often considered "black boxes."
- Struggles with complex reasoning tasks that require multi-step logic.
- Potential for bias and ethical concerns in generated content.

### Knowledge Graphs (KGs): Pros:

- Structured and explicit knowledge representation for machine understanding.

- Enhanced reasoning and querying, supporting multi-hop queries and logical inferences.
- Domain-specific precision with high accuracy in specialized fields.
- Consistency and reusability across applications.
- High explainability, making them ideal for transparent decision-making.

### Cons:

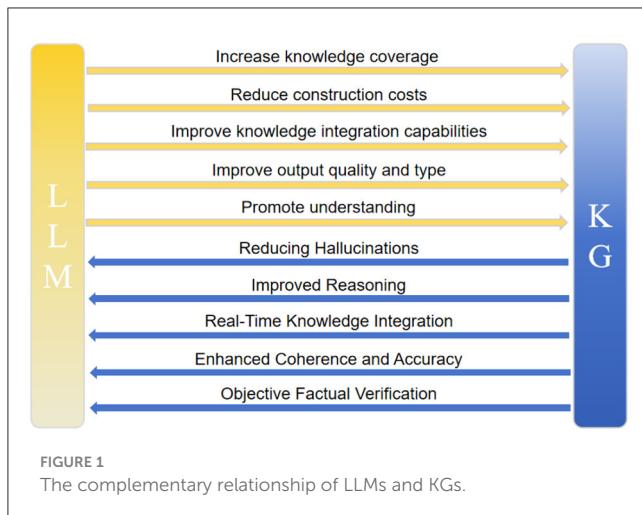
- Labor-intensive construction, requiring manual curation and domain expertise.
- Scalability challenges as KGs grow.
- Difficulty integrating with unstructured data sources.
- Limited coverage of knowledge.

## 2.4 How LLM helps reduce KG limitations

1. Increase knowledge coverage: Large models use semantic understanding, generation and other capabilities to extract knowledge and improve the accuracy and coverage of knowledge extraction.
2. Reduce construction costs: Large models extract implicit, complex, and multimodal knowledge with a better understanding of text and basic knowledge, which can reduce the cost of graph construction.
3. Improve output quality and type: Large models help improve the output of KGs and generate more reasonable, coherent, and innovative content.
4. Promote understanding: Large models help the output of KGs to better integrate and classify unstructured data and information.

## 2.5 How knowledge graph-based retrofitting corrects LLM limitations

1. Reducing Hallucinations: KGR incorporates KGs to verify and retrofit LLM-generated responses. Cross-referencing LLM output with KG data ensures the alignment of response with verified knowledge.
2. Improved Reasoning: KGR extracts claims from initial LLM drafts and performs a chain of verification, enabling the LLM to validate its reasoning processes using structured knowledge.
3. Real-Time Knowledge Integration: KGR autonomously integrates real-time knowledge from KGs, enabling LLMs to access up-to-date factual information, improving the reliability and relevance of responses.
4. Enhanced Coherence and Accuracy: KGR ensures contextually relevant facts in generating responses through structured verification processes, improving the overall coherence and accuracy of output.
5. Objective Factual Verification: KGR relies on KGs to provide a more objective source of factual information, helping to counterbalance biases in the training data of LLMs.



The complementary relationship of LLMs and KGs is illustrated in Figure 1, which highlights how their integration can enhance AI systems.

## 2.6 The roadmap of the fusion of KGs and LLMs

The technical approach outlined in Figure 2 demonstrates the integration of KGs with LLMs, which enhances the overall system by combining structured information with the reasoning and language generation capabilities of LLMs. This integration results in improved accuracy, context-awareness, and reasoning efficiency across various tasks. As the system evolves, it can handle complex, multimodal applications more effectively by continuously improving through the synergy of KGs and LLMs.

To take this integration further, there are three possible fusion strategies: LLM-Enhanced KGs (LEK), KG-Enhanced LLMs (KEL), and Collaborative LLMs and KGs (LKC).

## 3 LLM-enhanced KGs

KGs link entities and relationships in a structured format, supporting applications like question answering, recommendation systems, and web search. However, traditional KGs face challenges such as incompleteness and under-utilization of textual data. Recent research integrates LLMs to address these issues by incorporating text data and improving KG performance across various tasks shown in the Figure 3.

Building upon these task-specific advancements, the literature surveyed in this section reflects the rapid evolution of LLM-enhanced KG techniques since 2019, with particular emphasis on breakthroughs from 2023–2025. We prioritize studies that address fundamental challenges in KG construction, embedding, and reasoning through innovative integration of LLMs, evaluating papers based on their methodological novelty, such as the introduction of hybrid architectures or prompt-based techniques, and their demonstrated impact through benchmarks on standard

datasets. Open-source availability further informed our selection to facilitate future research. Figure 4 provides a structured overview of key studies with different tasks, organized by their publication year.

### 3.1 LLMs for KGs construction

Knowledge Graph Construction refers to the process of extracting entities, relations, and events from structured or unstructured data to form a structured knowledge network.

Figure 5 illustrates the role of LLMs in KGs construction, which involves the extraction and generation of entities, relations, and events, as well as tasks like entity linking and coreference resolution. In the extraction and generation processes, LLMs act as both prompts and generators, aiding in the creation of structured knowledge. For entity linking, LLMs serve as prompts and categorizers, linking entities to external knowledge sources. In coreference resolution, LLMs function as selectors and summarizers, resolving references to the same entity across different contexts. This integration enhances the accuracy and efficiency of KG construction, allowing for the development of comprehensive, interconnected KGs.

#### 3.1.1 Named entity recognition

Named Entity Recognition (NER) identifies and classifies entities in unstructured data. LLMs improve NER by utilizing their deep understanding of language and context, enhancing entity recognition and classification in challenging scenarios. GPT-NER (Wang S. et al., 2023) bridges the gap between LLMs and NER by converting the sequence labeling task into a text generation task, using special markers to identify entities. TOPT (Zhang et al., 2024a) is a task-oriented pre-training model that uses LLMs to generate task-specific knowledge corpora, enhancing domain adaptability and NER sensitivity. Graphusion (Yang et al., 2024d) combines entity merging, conflict resolution, and novel triple discovery to provide a global perspective for entity extraction, addressing the challenge of using free text inputs. SF-GPT (Sun et al., 2025) uses three modules for knowledge triple extraction: Entity Extraction Filter to filter results, Entity Alignment Generator to enhance semantic richness, and Self-Fusion Subgraph strategy to reduce noise.

#### 3.1.2 Relation extraction and generation

Relation extraction is the process of identifying and classifying semantic relationships between entities in text. Many studies have already used LLMs to enhance this process. BertNet (Hao et al., 2022) built a search and re-scoring mechanism that effectively searches a wide entity pair space with minimal relationship definitions, improving both efficiency and accuracy. CDLM (Caciularu et al., 2021) employs a dynamic global attention mechanism to improve long-range transformers, enabling them to access the entire input for predicting masked tokens. DREAM (Ma et al., 2023) introduces a memory-efficient approach that uses evidence information as a supervisory signal to guide the attention module in assigning high weights to evidence.

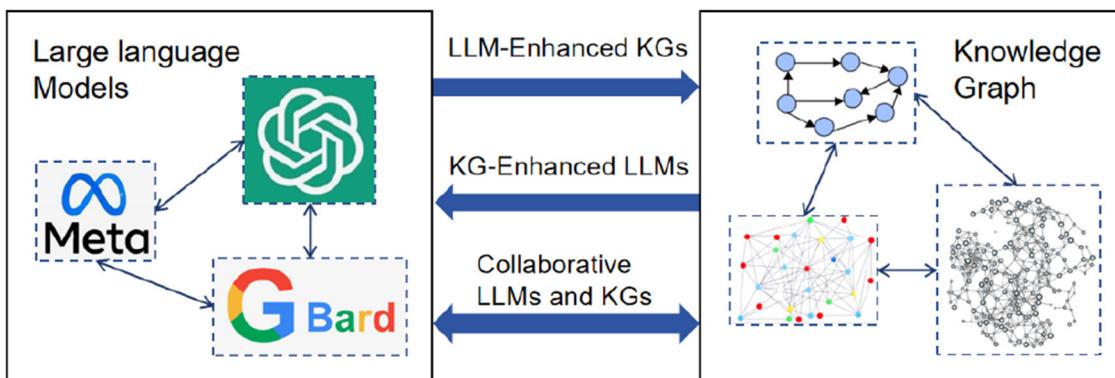


FIGURE 2  
The roadmap of the fusion of KGs and LLMs.

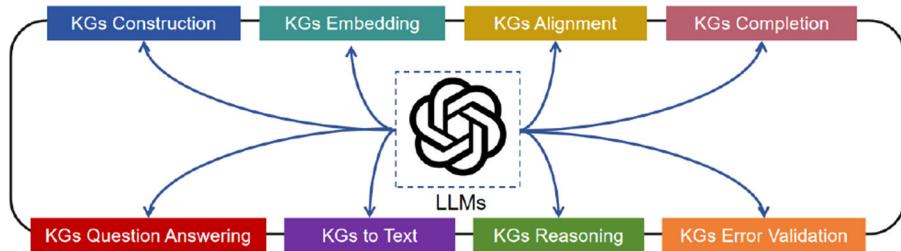


FIGURE 3  
The roadmap of LLM-enhanced KGs (LEK).

### 3.1.3 Event extraction and generation

Event extraction involves automatically identifying events and related information from text, including triggers, entities, and key details like time and location. This forms the basis for constructing event KGs, such as EventKG (Gottschalk and Demidova, 2018). EvIT (Tao et al., 2024) trains LLMs through event-oriented instruction tuning and uses a heuristic unsupervised method to mine event quadruples from large-scale corpora. Chen R. et al. (2024) uses LLMs as expert annotators to extract event information from sentences and generate augmented datasets aligned with baseline distributions. STAR (Ma M. D. et al., 2024) proposes a data generation method that uses LLMs to synthesize data with minimal seed examples. It generates target structures (Y) and paragraphs (X) through detailed instructions, followed by error identification and iterative improvements to enhance data quality.

### 3.1.4 Entity linking

Entity Linking (EL) matches text mentions to specific entities in a knowledge base to enhance text understanding and information retrieval. ReFinED (Ayoola et al., 2022) uses fine-grained entity types and entity descriptions to construct an efficient end-to-end entity linking model, which can be generalized to other large-scale knowledge bases. ChatEL (Ding Y. et al., 2024) proposes a three-step framework that leverages LLMs for efficient entity linking by generating candidate entities, enhancing contextual information, and incorporating a multiple-choice

format. UniMEL (Liu Q. et al., 2024) proposes a multimodal entity linking framework using LLMs. It integrates textual and visual information to enhance mention and entity representations, and improves linking performance through embedding-based retrieval and candidate re-ranking.

### 3.1.5 Coreference resolution

Coreference resolution is a NLP task that aims to identify and link different expressions in a text that refer to the same entity. Zheng L. et al. (2024) proposes an adaptive multimodal data augmentation framework to tackle data scarcity and under-utilization in multimodal coreference resolution (MCR). Min et al. (2024) presents a collaborative approach for Cross-Document Event Co-reference Resolution(CDECR), combining a general-purpose LLM to summarize events and a task-specific small language model to further improve its event representation learning. Nath et al. (2024) propose a principle-based method for event clustering and knowledge refinement, utilizing Free Text Reasoning (FTR) generated by modern auto-regressive LLMs to improve event co-reference resolution.

## 3.2 LLMs for KGs embedding

KGs embedding (KGE) involves learning low-dimensional representations of entities and relations within KGs. The process

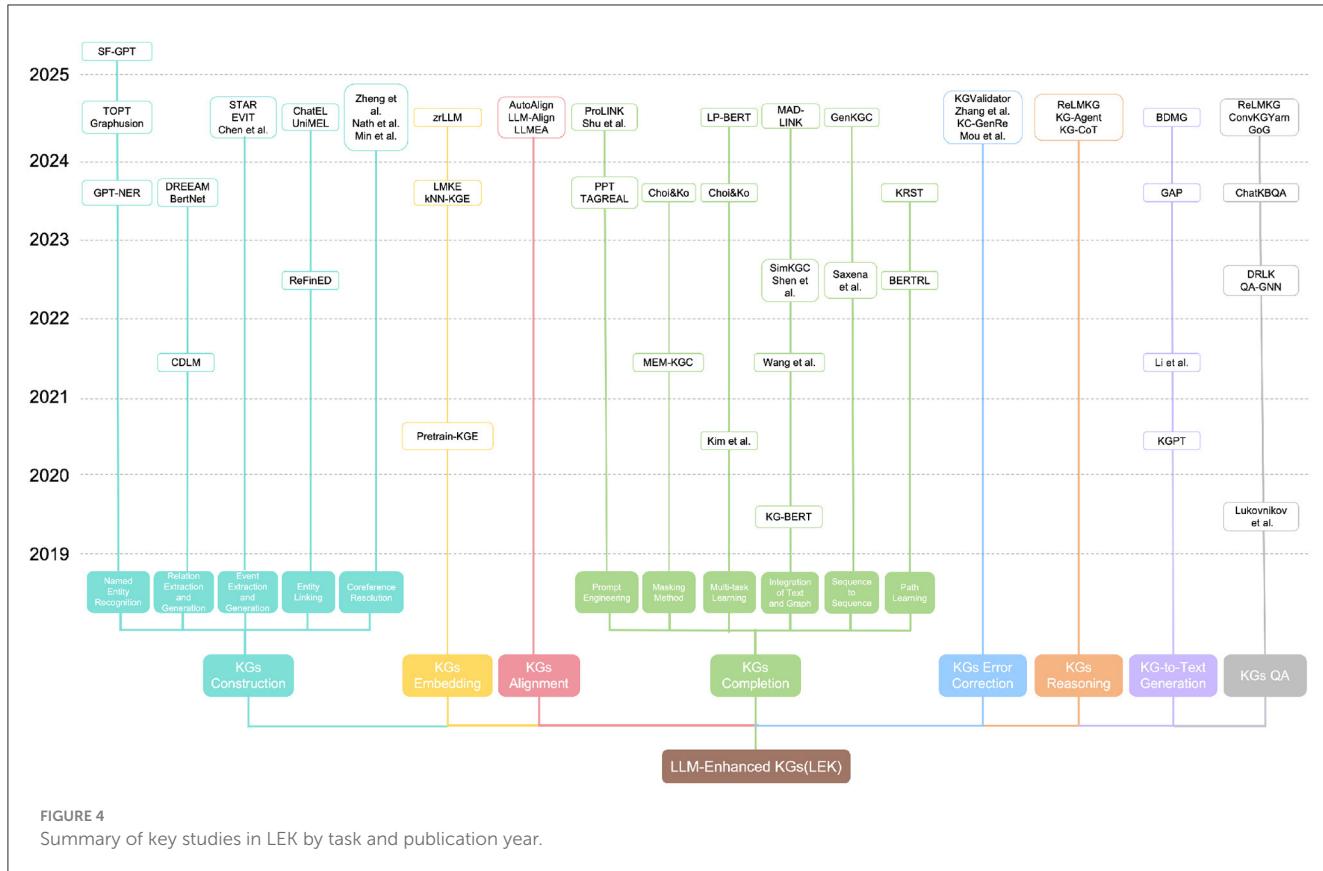


FIGURE 4  
Summary of key studies in LEK by task and publication year.

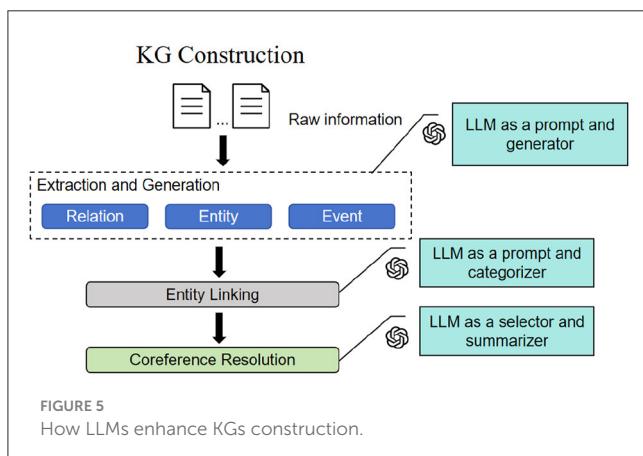


FIGURE 5  
How LLMs enhance KGs construction.

begins with entity and relation representation, followed by scoring function definition, and culminates in representation learning. There are two main approaches for embedding: structure-based and description-based. Below is the Figure 6 showing how LLMs enhance KGs embedding process.

Pretrain-KGE (Zhang Z. et al., 2020) is a training framework applicable to any KGE model, which incorporates world knowledge from the pre-trained model into entity and relation embeddings to enhance the performance of the KGE model. LMKE (Wang X. et al., 2022) and zrLLM (Ding Z. et al., 2024) uses a language model to derive knowledge embeddings, enriching long-tail entity representation and addressing issues in description-based methods.

kNN-KGE (Wang P. et al., 2023), with a pre-trained language model, uses k nearest neighbors for linear interpolation of entity distributions, calculated based on the distance between entity embeddings and knowledge storage.

### 3.3 LLMs for KGs alignment

Entity alignment refers to the process of matching and aligning nodes representing the same entities across different KGs. AutoAlign (Zhang R. et al., 2023) constructs a predicate proximity graph to capture the similarity of predicates between KGs and uses TransE (Bordes et al., 2013) to compute entity embeddings, aligning entities from different graphs into the same vector space. LLM-Align (Chen X. et al., 2024) selects important entity attributes and relations through heuristic methods, inputs entity triples into the LLM to infer alignment results, and employs a multi-round voting mechanism to mitigate hallucinations and positional bias. Additionally, the LLMEA (Yang et al., 2024b) method further identifies candidate alignments by combining entity embedding similarity and edit distance, optimizing alignment results through the reasoning capabilities of LLMs.

### 3.4 LLMs for KGs completion

#### 3.4.1 Prompt engineering

Prompt engineering for KGs completion involves designing input prompts to guide LLMs in inferring and filling missing

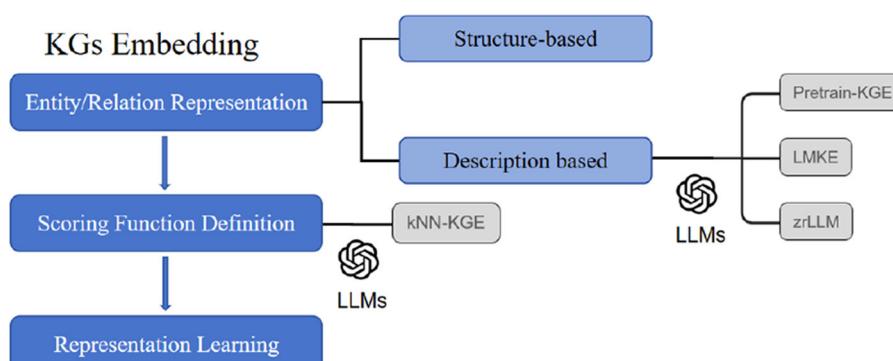


FIGURE 6  
How LLMs enhance KGs embedding.

parts of KGs. This approach enhances multi-hop link prediction and explores the potential of LLMs to handle unseen cues in zero-sample scenarios (e.g., Shu et al., 2024). On this basis, ProLINK (Wang K. et al., 2024) proposes a novel pre-training and hinting framework designed for low-resource inductive reasoning in arbitrary knowledge graph without additional training. At the same time, TAGREAL (Jiang P. et al., 2023) is able to automatically generate high-quality query hints and retrieve supporting information from large text corpora to detect knowledge in pre-trained language models (PLMs). PPT-based TKGC model (Xu et al., 2023) uses Prompt-based Pre-trained Language Models. This model is trained with a masking strategy, turning the TKGC task into masked token prediction to utilize semantic information from the pre-trained model.

### 3.4.2 Masking method

The Masked Language Model (MLM) is a pre-training task where some words in a text sequence are replaced with [MASK]. The model then predicts the most likely word to fill the [MASK] based on the context. MEM-KGC (Choi et al., 2021) adopts this process by masking the tail entity and using the head entity and relation as context to predict the missing tail entity. This is similar to MLM, where the model predicts the masked tokens based on the given context. Building on this, Choi and Ko (2023) predict the appropriate entity or relation for the masked positions. Additionally, to address the issue of new entities in open-world KGC, Choi and Ko (2023) propose a unified learning method that generates embeddings to replace token embeddings for new entities.

### 3.4.3 Multi-task learning

Multi-task learning is an effective method for improving link prediction performance and there are substantial studies have already built relevant models. Choi and Ko (2023) proposed a multi-task learning network (MT-DNN) architecture that combines Entity Description Prediction (EDP) and Entity Type Prediction (ITP) tasks, sharing the same pre-trained language model and network layers for joint training. Similarly, the LP-BERT

(Li et al., 2023) model employs a multi-task learning approach with three tasks: Masked Language Model (MLM), Masked Entity Model (MEM), and Masked Relation Model (MRM), sharing the same input format to simultaneously learn contextual and semantic information. Kim et al. (2020) integrate relation prediction and relevance ranking tasks with link prediction, enabling the model to better learn relational attributes in KGs.

### 3.4.4 Integration of text representation and graph embedding

In recent years, combining text encoding with graph embedding has emerged as a promising approach for knowledge graph completion. KG-BERT (Yao et al., 2019) treats knowledge graph triples as textual sequences, encoding them using BERT-style architectures. Similarly, SimKGC (Wang L. et al., 2022) employs contrastive learning with in-batch, pre-batch, and self-negatives to enhance entity representations. Shen et al. (2022) optimize semantic representations from language models and structural knowledge through a probabilistic loss. Another line of work integrates attention mechanisms, such as MADLINK (Biswas et al., 2024), which uses an attention-based encoder-decoder to combine KG structure with textual entity descriptions. Wang B. et al. (2021) employ Siamese networks to learn structured representations while avoiding combinatorial explosion.

### 3.4.5 Sequence-to-sequence methods

In recent advancements in KG completion tasks, leveraging sequence-to-sequence models has also shown great promise. Saxena et al. (2022) propose transforming KG link prediction into a sequence-to-sequence task and replacing the traditional triple scoring method with auto-regressive decoding. Similarly, GenKGC (Xie et al., 2022) leverages pre-trained language models to convert the KGs completion task into a sequence-to-sequence generation task.

### 3.4.6 Path learning

The core idea of path learning is to treat the connection paths between entities as the basis, thereby capturing both explicit

information and implicit relationships in structured KGs. BERTRL (Zha et al., 2022) leverages pre-trained language models and fine-tunes them with relation instances and reasoning paths as training samples. KRST (Su et al., 2023) encodes reliable paths in the KG, enabling accurate path clustering and providing multifaceted explanations for predicting inductive relations.

### 3.5 LLMs for KGs error validation

KGs error validation refers to the process of checking and confirming the data within KGs to ensure its accuracy and consistency. One common method is to use LLMs for validation against external knowledge bases. Zhang M. et al. (2024) proposed an LLM-enhanced embedding framework, which first uses the graph structure information to determine whether the triplet relations hold and selects suspicious ones, and finally combines the language model for validation. KGValidator (Boylan et al., 2024) is a consistency and validation framework for validating KGs using generative models, supporting any external knowledge source.

Some studies validate and correct errors by adjusting the model itself. KC-GenRe (Wang Y. et al., 2024) transforms the KGC re-ranking task into a candidate ranking problem solved by a generative LLM. It also tackles missing issues with a knowledge-enhanced constraint reasoning method. Mou et al. (2024) proposed a self-reflective model where GPT-4 reflects on the errors it makes in a given example and generates linguistic feedback to guide the model in avoiding similar mistakes during KGC.

### 3.6 LLMs for KGs reasoning

KGs reasoning leverages graph structures and logical rules to infer new information or relationships from existing knowledge. ReLMKG (Cao and Liu, 2023) uses the language model to encode complex questions and guides the graph neural network in message propagation and aggregation through outputs from different layers. KG-Agent (Jiang J. et al., 2024) utilizes programming languages to design multi-hop reasoning processes on KGs and synthesizes code-based instruction datasets for fine-tuning base LLMs. KG-CoT (Zhao et al., 2024), utilizes a small-scale incremental graph reasoning model for inference on KGs. It employs a method for generating inference paths to create high-confidence knowledge chains for large-scale LLMs.

### 3.7 LLMs for KGs to text

KG-to-text is a method that generates natural language text from structured KGs by leveraging models to map graph data into coherent, informative sentences. GAP (Colas et al., 2022) utilizes a masking structure to capture neighborhood information and introduces a novel type encoder that biases graph attention weights based on connection types. KGPT (Chen et al., 2020) comprises a generative model for producing knowledge-enriched text and a pre-training paradigm on a large corpus of knowledge text crawled from the web to realizing tasks. Li et al. (2021) made significant

contributions in introducing a BFS strategy with a relationship bias for KG linearization, and employing multi-task learning with KG reconstruction. BDMG (Du et al., 2024) utilizes a bi-directional multi-granularity generation framework to construct sentence-level generation multiple times based on the corresponding ternary components, and ultimately generates graph-level text.

### 3.8 LLMs for KGs question answering

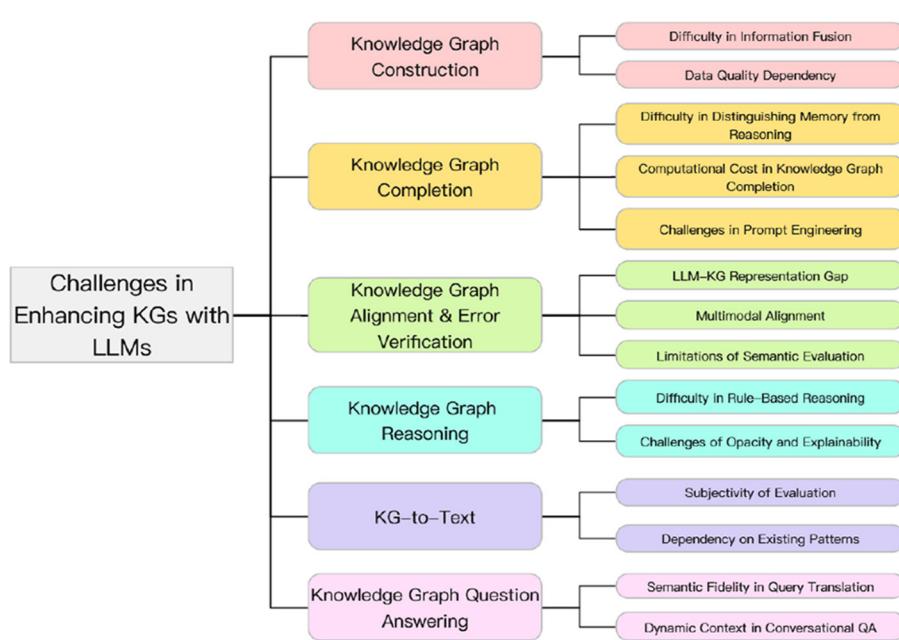
Knowledge graph question answering (KGQA) systems leverage NLP techniques to transform natural language queries into structured graph queries. Pre-trained transformer-based methods like Lukovnikov et al. (2019)'s model and ReLMKG (Cao and Liu, 2023) use language models to bridge semantic gaps between questions and KG structures, with ReLMKG (Cao and Liu, 2023) additionally employing GNNs for explicit knowledge propagation. Generation-retrieval frameworks such as ChatKBQA (Luo H. et al., 2023) and GoG (Xu et al., 2024) adopt a two-stage approach, first generating logical forms or new triples before retrieving relevant KG elements. Dynamic reasoning systems like DRLK (Zhang M. et al., 2022) extract hierarchical QA context features, while QA-GNN (Yasunaga et al., 2021) performs joint reasoning by scoring KG relevance and updating representations through GNNs. For dataset construction, ConvKGYarn (Pradeep et al., 2024) provides a scalable method to generate configurable conversational KGQA datasets using LLMs.

## 4 Challenges in enhancing KGs with LLMs

Despite the increasing research on enhancing KGs with LLMs in recent years, several challenges remain. Figure 7 summarizes these challenges and points the way for future research.

### 4.1 Challenges in knowledge graph construction

- Difficulty in information fusion:** LLM-KG fusion encounters fundamental representational conflicts between the implicit statistical patterns of LLMs and the explicit symbolic structures of KGs. This mismatch systematically disrupts entity linking consistency. Current hybrid approaches suffer three core limitations: introduce semantic noise during context augmentation (Ayoola et al., 2022; Xin et al., 2024), remain constrained by LLM training biases in candidate generation (Ding Y. et al., 2024), and create new modality-specific dependencies in multimodal fusion (Liu Q. et al., 2024). This stems from treating LLMs as peripheral tools rather than re-engineering the core symbolic-neural interface. Future solutions must move beyond augmentation paradigms to enable dynamic, runtime knowledge translation between paradigms.
- Data quality dependency:** The effectiveness of LLM-based knowledge graph construction critically depends on input data quality. Through our analysis, we identify three universal



**FIGURE 7**  
Challenges in enhancing KGs with LLMs.

limitations of LLMs in this context-inherent training data biases that propagate through knowledge extraction pipelines, fundamental domain adaptation challenges with specialized knowledge (Zhang et al., 2024a), and systematic coverage gaps for long-tail relationships, particularly in cross-document scenarios (Caciularu et al., 2021; Min et al., 2024). These issues collectively undermine the reliability of constructed knowledge graphs, especially in professional domains where precision is paramount. Current mitigation strategies, such as manual verification or domain-specific knowledge bases, often create scalability bottlenecks that limit practical implementation (Fan et al., 2007).

## 4.2 Challenges in knowledge graph completion

1. **Difficulty in distinguishing memory from reasoning:** LLMs intrinsically blend memorized knowledge with inferred predictions during KG completion. This creates evaluation challenges when benchmark datasets overlap with pre-training corpora, because LLMs generate predictions without distinguishing among: factual recall, statistical inference, or hallucination. While prompt-based methods like ProLINK (Wang K. et al., 2024) and TAGREAL (Jiang P. et al., 2023) attempt to guide LLM reasoning, they cannot fully address the fundamental ambiguity between factual recall and genuine inference—a limitation particularly problematic in healthcare applications where provenance matters (Walcock et al., 2024). This challenge persists across all LLM-based completion

paradigms (prompting, masking, seq2seq) despite their semantic richness.

2. **Computational cost in knowledge graph completion:** LLM-based completion [e.g., sequence-to-sequence GenKGC (Xie et al., 2022), text-graph hybrid MADLINK (Biswas et al., 2024)] requires exhaustive text processing and candidate scoring, which can be computationally expensive in large KGs (Wang B. et al., 2021; Ren et al., 2024). While multi-task learning approaches [MT-DNN (Choi and Ko, 2023), LP-BERT (Li et al., 2023)] attempt to share computational overhead across tasks, the fundamental scalability gap persists—especially in large-scale KGs where latency grows polynomially with graph density (Heim et al., 2025). This creates an unresolved tension between LLMs' semantic richness and traditional methods' operational efficiency.

3. **Challenges in prompt engineering:** Current prompt engineering approaches [ProLINK (Wang K. et al., 2024), TAGREAL (Jiang P. et al., 2023)] for knowledge graph completion exhibit several key limitations. When representing complex entity names, prompt methods must split long names into subword fragments, leading to information loss, whereas masking techniques like MEM-KGC (Choi et al., 2021) preserve full entity integrity using [MASK] tokens. For relationship understanding, prompt methods rely on manually crafted templates that yield inconsistent results, while path-based approaches such as BERTRL (Zha et al., 2022) automatically analyze inter-entity pathways for more reliable predictions. These constraints force prompt methods to require excessive manual maintenance when adapting to new domains or knowledge updates, severely limiting their scalability (Choi and Ko, 2023; Li H. et al., 2024).

### 4.3 Challenges in knowledge graph alignment and error verification

- LLM-KG representation gap:** The mismatch in tokenization between LLM and KG embeddings can lead to information loss during alignment. Although mixed methods such as LLM-Align (Chen X. et al., 2024) compensate for this shortcoming through multiple rounds of voting, they still have certain limitations in some complex contexts and also result in high costs. This fragmentation directly reduces the reliability of human-machine interfaces (HMIs) because it leads to inconsistent interpretations and causes ambiguity and confusion. Future research could consider more about the development of a unified, novel tokenization scheme that balances preserving semantic meaning (suitable for LLMs) with maintaining structural integrity (suitable for KGs).
- Multimodal alignment:** Effective knowledge graph alignment requires integrating structural and semantic features. Methods such as AutoAlign (Zhang R. et al., 2023) show that cross-knowledge graph alignment benefits from multi-feature fusion, but the computational overhead increases exponentially with graph size. This poses a challenge in the multimodal domain, as multimodal integration can improve accuracy but also consumes a significant amount of resources. Future work could explore dynamic feature selection strategies that prioritize high-value fusion operations while skipping redundant computations.
- Limitations of semantic evaluation:** Existing evaluation metrics for knowledge graph completion often prioritize surface-level correctness over logical consistency. For example, a generated triple like (*Einstein*, *won*, *Nobel Prize in Chemistry*) may achieve high confidence scores from embedding-based metrics, despite contradicting factual knowledge. Rule-based systems like AMIE (Galárraga et al., 2013) can catch such errors through predefined constraints, but struggle with open-domain scenarios where rules are incomplete. A more promising direction may be hierarchical evaluation frameworks, such as applying strict symbolic verification only to high-risk predictions.

### 4.4 Challenges in knowledge graph reasoning

- Difficulty in rule-based reasoning:** The core challenge in LLM-based KG reasoning stems from the inherent conflict between probabilistic inference and deterministic symbolic rules. Current methods aim to enhance LLM performance in logical tasks through various strategies, but each has core limitations: ReLMKG (Cao and Liu, 2023) struggles with dynamic multi-hop reasoning and lacks interpretability; KG-Agent (Jiang J. et al., 2024) relies on predefined rules, resulting in limited generalization and high maintenance costs; KG-CoT (Zhao et al., 2024) is constrained by the completeness of knowledge graphs, and local correctness cannot guarantee global logical consistency. All three face issues of static knowledge dependency and error propagation, and

lack the modular processing capabilities of symbolic systems for complex logic. Future work should prioritize hybrid architectures that dynamically switch between neural flexibility and symbolic rigor.

- Challenges of opacity and explainability:** The probabilistic nature of LLMs creates fundamental explainability barriers in KG reasoning tasks. Unlike symbolic systems that maintain explicit inference graphs, LLMs cannot reliably reconstruct the logical chain connecting input premises to final predictions—a critical shortfall for HMI applications requiring auditability (e.g., clinical decision support where physicians must verify diagnostic pathways). This opacity persists even in advanced CoT frameworks like KG-CoT (Zhao et al., 2024), as their generated rationales often conflate genuine reasoning with post-hoc justifications. Future solutions may require “white-box” intermediate representations that simultaneously support neural computation and human-interpretable stepwise verification.

### 4.5 Challenges in KG-to-text

- Subjectivity of evaluation:** Current evaluation metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) mainly measure surface text similarity and cannot effectively capture the semantic consistency between generated text and KG content. Recent studies (Luo et al., 2024; Honovich et al., 2022) have begun to explore fact consistency evaluation based on LLM, but the computational cost has increased significantly.
- Dependency on existing patterns:** Generated text descriptions may overly rely on existing templates or syntactic structures, lacking innovation in language expression. This dependency makes it difficult for generated text to offer novel perspectives or unique linguistic styles, thus limiting its creativity in expressing knowledge graph contents.

### 4.6 Challenges in knowledge graph question answering

- Semantic fidelity in query translation:** A critical challenge in LLM-powered KGQA systems is semantic drift during natural language-to-KG-query conversion (Li H. et al., 2024). Current LLM-driven retrieval-enhanced methods [such as ChatKBQA (Luo H. et al., 2023)] face semantic fidelity issues when dealing with complex queries. When handling queries with implicit constraints, LLM-generated queries often lose critical elements such as temporal ranges or comparative logic. While GoG (Xu et al., 2024) partially mitigates this issue through a generate-retrieve framework, semantic deviations in the generation phase directly lead to retrieval results deviating from user intent. Future work requires end-to-end context-aware architectures capable of simultaneously processing language parsing and graph structure constraints.
- Dynamic context in conversational QA:** Current KGQA systems often mishandle contextual continuity in multi-turn dialogues, either dropping key constraints (e.g.,

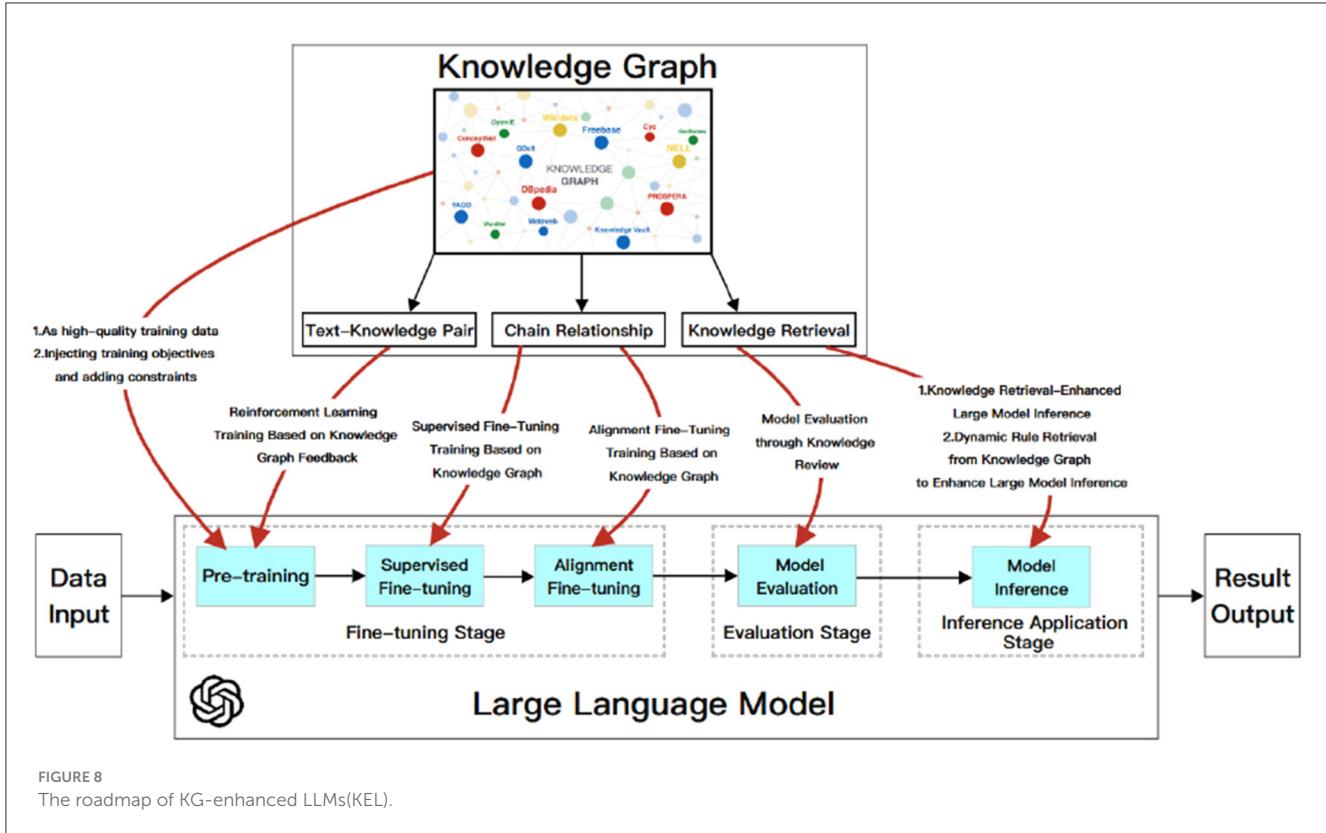


FIGURE 8  
The roadmap of KG-enhanced LLMs(KEL).

temporal filters) or misapplying them. While frameworks like ConvKGYarn (Pradeep et al., 2024) generate coherent standalone queries, they lack cross-turn KG validation, producing contradictory answers. This directly impacts HMI by generating unexplained contradictions that erode user trust. Future solutions require integrated context tracking that simultaneously monitors dialogue history and KG constraints.

## 5 KG-enhanced LLMs

In the realm of NLP, LLMs have emerged as powerful tools for understanding and generating text. However, they often struggle with tasks that require deep knowledge and complex reasoning due to the limitations of their internal knowledge base. KGs, with their structured knowledge, can bridge this gap. By integrating KGs into LLMs, we can significantly enhance their performance on a variety of NLP tasks, particularly those involving intricate knowledge and reasoning. This section explores innovative methods that leverage KGs to boost LLMs' capabilities, from pre-training objectives to inference techniques, highlighting the potential of this integration to empower LLMs to tackle more sophisticated tasks. The Figure 8 illustrates the workflow of large LLMs and shows how KGs enhance different steps of LLMs.

In surveying these advancements, we focus particularly on studies from 2019-2025 that demonstrate measurable improvements in LLM capabilities through KG integration. We can categorize this process based on the effects of KG enhancement into three types: pre-training, reasoning methods,

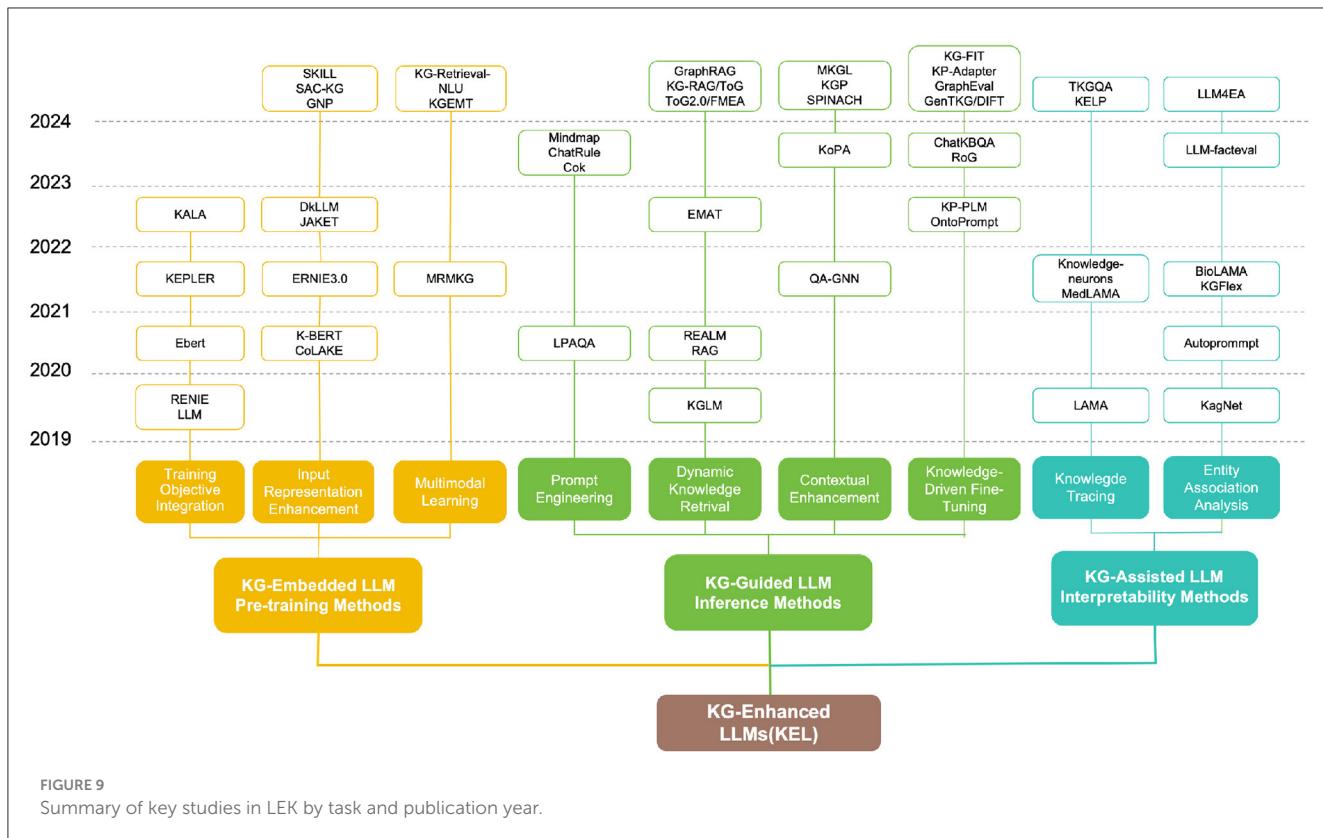
including supervised fine-tuning, alignment fine-tuning, and model interpretability. Our evaluation prioritized works that not only introduced innovative methodologies within these categories but also demonstrated consistent performance gains across knowledge-intensive tasks. Special consideration was given to techniques maintaining compatibility with mainstream LLM architectures while showing robust validation in multiple domains. These methodological developments are summarized by year and method type in Figure 9.

### 5.1 KG-embedded LLM pre-training methods

#### 5.1.1 Training objective integration

Integrating KGs into large LLMs is a crucial challenge for enhancing model performance in NLP tasks. Some models use KGs to make the pre-trained data more structured in order to enhance the performance of LLMs.

KEPLER (Wang X. et al., 2021) study generates entity embeddings by encoding text descriptions and simultaneously optimizes knowledge embedding and masked language model objectives, which performs well in knowledge graph link prediction. WKLM (Xiong et al., 2019) employs a weakly supervised pre-training objective by replacing entity mentions in documents and training the model to distinguish between true and false knowledge expressions. ERNIE (Zhang et al., 2019) enhances NLP with knowledge graphs; E-BERT (Zhang D. et al., 2020) optimizes e-commerce tasks via hybrid masking; KEPLER (Wang X. et al.,



2021) unifies knowledge embedding and language modeling for SOTA results; Multi-task QA Model (Su et al., 2019) improves generalization using XLNet; KALA (Kang et al., 2022) boosts domain adaptation with entity-aware tuning; Knowledge-enhanced Pre-training (Xiong et al., 2019) strengthens factual understanding via weak supervision. All the models leverage specialized knowledge or training strategies to outperform general models.

### 5.1.2 Input representation enhancement

When exploring how to enhance the input representations of LLMs through KGs, several key studies demonstrate various methods to improve the understanding and generation capabilities of the models.

CoLAKE (Sun et al., 2020) proposes a unified pre-training framework that jointly learns contextualized representations of language and knowledge by integrating them into a shared structure called the word-knowledge graph. ERNIE 3.0 (Sun et al., 2021b) achieves SOTA on 54 Chinese NLP tasks through hybrid architecture. DKPLM (Zhang T. et al., 2022) improves efficiency in long-tail entity processing via decomposable knowledge injection. JAKET (Yu et al., 2022) enables bidirectional enhancement between knowledge graphs and language. KG-T5 (Moiseev et al., 2022) directly pre-trains on KG triples for 3× performance gain. SAC-KG (Chen S. et al., 2024) leverages LLMs to construct million-scale high-precision knowledge graphs. GNP (Tian et al., 2024) bridges LLMs and KGs through graph neural prompting. K-BERT (Liu et al., 2020) enables efficient domain knowledge injection with noise control. Together, these models advance knowledge-enhanced

language models through approaches ranging from joint training to modular methods.

### 5.1.3 Multimodal learning

In exploring how to enhance the capabilities of language models through multimodal KGs, propose innovative methods and frameworks aimed at improving the performance of models when handling multimodal data. MRMKG (Lee et al., 2024) uses RGAT to encode MMKG and designs a cross-modal module for image-text alignment. It is pre-trained on a dataset from matching VQA instances with MMKGs. The KGEMT (Zheng J. et al., 2024) framework combines coarse and fine-grained learning to build a global semantic graph for multimodal alignment. It uses bidirectional fine-grained matching to filter image-text elements, boosting image-text retrieval performance. KG-Retrieval NLU (Huang et al., 2022) is a parameter-efficient framework leveraging multimodal KG (VisualSem) retrieval to enhance NLU.

## 5.2 KG-guided LLM inference methods

### 5.2.1 Prompt engineering

In knowledge graph-guided LLMs reasoning methods, prompt engineering plays a crucial role. The following is a summary of the applications of prompt engineering in this field. Several methods focus on enhancing LLMs' reasoning abilities by explicitly injecting structured cues from knowledge graphs (KGs) into the prompt space. LPAQA (Jiang et al., 2020) introduces label-aware prompting

by aligning KG entities with carefully designed templates, thereby guiding the model to generate accurate answers to factual questions [1]. Similarly, approaches like Mindmap (Wen et al., 2023), ChatRule (Luo et al., 2023a), and COK (Wang et al., 2023a) aim to externalize structured knowledge or human-defined rules into prompt representations, enabling LLMs to reason over complex graph-based scenarios with improved contextual grounding and reduced hallucinations. These methods exemplify how prompt design can serve as a lightweight yet powerful interface between KGs and LLMs.

### 5.2.2 Dynamic knowledge retrieval

In the era where LLMs strive to conquer complex tasks, dynamic knowledge recovery from KGs has emerged as a powerful solution, enabling LLMs to access and integrate relevant real-time knowledge for enhanced performance. This category targets real-time, query-specific information injection into LLMs through retrieval-enhanced architectures. REALM (Lewis et al., 2020) and RAG (Guu et al., 2020) pioneered the integration of neural retrievers with generative transformers, retrieving relevant documents or knowledge passages from large corpora or knowledge bases to support downstream predictions. KGLM (Youn and Tagkopoulos, 2022) extends this concept by embedding knowledge entities directly into the generation process, allowing the model to dynamically refer to entity-specific information during decoding. Further, EMAT (Mirkhani et al., 2004) improves retrieval alignment by introducing entity-matching-aware attention mechanisms [8]. Building on these foundations, newer methods such as GraphRAG (Edge et al., 2024), KG-RAG (Sanmartin, 2024), ToG (Sun et al., 2023), ToG2.0 (Ma S. et al., 2024), and FMEA-RAG (Razouk et al., 2023) incorporate structured graph reasoning and multi-hop retrieval into the RAG framework, allowing LLMs to reason over graph-structured evidence, which is particularly beneficial for technical tasks like industrial fault diagnosis, knowledge-based summarization, and domain-specific decision making.

### 5.2.3 Contextual enhancement

In the realm of NLP, Contextual Enhancement, empowered by KGs, has become an essential strategy to break through the knowledge bottlenecks of LLMs and enable them to handle intricate tasks more effectively.

QA-GNN (Yasunaga et al., 2021) combines GNN reasoning with LM-powered KG scoring to achieve state-of-the-art performance on commonsense/medical QA with interpretable reasoning. KoPA (Zhang Y. et al., 2023) enhances LLMs for KG tasks by projecting structural embeddings into virtual knowledge tokens. KGL-LLM (Guo et al., 2025) introduces a dedicated Knowledge Graph Language for precise LLM-KG integration, reducing completion errors through real-time context retrieval. KP-PLM (Wang J. et al., 2022) advances knowledge prompting with dynamic subgraph conversion and dual self-supervised tasks, excelling in both full and low-resource NLU. SPINACH (Liu S. et al., 2024) contributes an expert-annotated KBQA dataset with in-context learning, significantly outperforming GPT-4 on

complex queries. Together, these models demonstrate diverse approaches to integrating structured knowledge with language models, spanning from graph-based reasoning (QA-GNN) to prompt engineering (KP-PLM) and specialized language interfaces (KGL-LLM).

### 5.2.4 Knowledge-driven fine-tuning

With the extensive application of LLMs in the field of NLP, enhancing their performance on specific tasks has become a research focus. KGs offer external knowledge to LLMs, facilitating their understanding, reasoning, and generation.

Knowledge-Driven Fine-Tuning encompasses approaches that incorporate structured knowledge during model adaptation, leading to better generalization and knowledge-awareness. KP-LLM (Wang J. et al., 2022) and OntoPrompt (Ye et al., 2022) fine-tune LLMs with ontological paths and schema constraints, aligning model outputs with structured knowledge rules. KG-FIT (Jiang P. et al., 2024) and GraphEval (Sansford et al., 2024) provide generalizable and modular frameworks that inject KG-derived signals during fine-tuning or evaluation, enabling models to become more robust, verifiable, and explainable in knowledge-intensive tasks. Meanwhile, ChatKBQA (Luo H. et al., 2023) and RoG (Luo et al., 2023b) integrate knowledge graph reasoning into conversational QA systems, enhancing both factual accuracy and discourse coherence. GenTKG (Liao et al., 2023) and DIFT (Liu Y. et al., 2024) extend this direction into generative KG completion and domain transfer settings, allowing models to adapt and perform under sparse supervision or evolving ontologies.

## 5.3 KG-assisted LLM interpretability methods

### 5.3.1 Knowledge tracing

As LLMs strive to maintain high-quality performance across various scenarios, Knowledge Tracing empowered by KGs enables them to precisely track knowledge evolution, filling in knowledge gaps and improving the accuracy of responses. KELP (Liu H. et al., 2024) enhances the factual accuracy of LLMs outputs through a three-stage process. This process extracts and selects knowledge graph paths semantically relevant to the input text. LAMA (Petroni et al., 2019) converts knowledge into cloze-style questions to evaluate the relational knowledge and recall ability of pre-trained models. Knowledge-neurons (Dai et al., 2021) identify and activate neurons corresponding to specific facts, exploring the storage of factual knowledge in pre-trained Transformers and the editing and updating of internal knowledge. MedLAMA (Meng et al., 2021) creates a benchmark based on UML and introduces the Contrastive-Probe, a self-supervised contrastive probing method. It can adjust the representation space of the underlying pre-trained models without any task-specific data. GenTKGQA (Gao et al., 2024) presents a two-phase temporal QA framework that first retrieves relevant subgraphs using LLM-derived constraints, then generates answers through joint representation of graph and textual information.

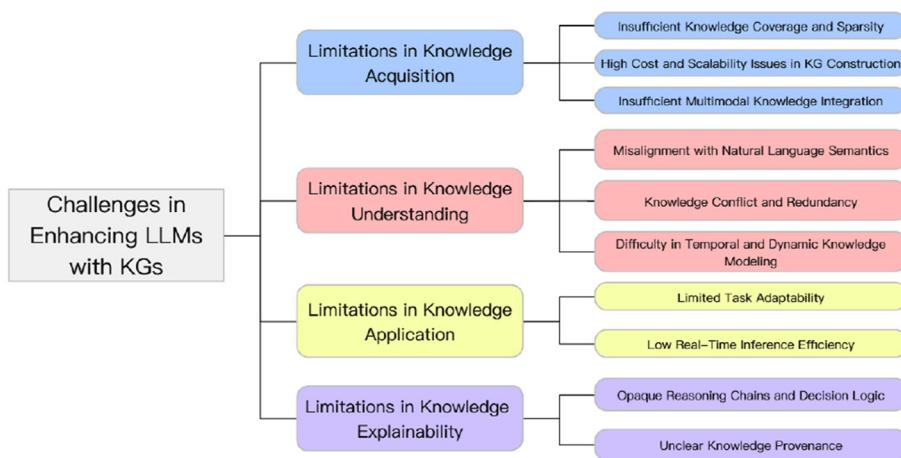


FIGURE 10

Challenges in enhancing LLMs with KGs.

### 5.3.2 Entity association analysis

As LLMs strive to handle complex knowledge-based tasks, Entity Association Analysis with the aid of KGs provides a powerful means to identify and utilize entity associations, filling knowledge gaps and promoting more accurate and intelligent responses.

KGFlex (Anelli et al., 2021) integrates KGs with a sparse factorization approach to analyze the dimensions of user decision-making and model user-item interactions. KagNet (Lin et al., 2019) constructs pattern graphs using a knowledge-aware graph network, and resorts to graph convolutional networks, LSTM, and a hierarchical path attention mechanism to solve the common-sense reasoning problems. AUTOPROMPT (Shin et al., 2020) automatically generates prompts through gradient-guided search to assist pre-trained models in performing tasks. BioLAMA (Sung et al., 2021) introduces a biomedical knowledge probing benchmark, assessing whether LMs can serve as domain-specific KBs using structured fact triples. LLM-facteval (Luo et al., 2023c) proposes a KG-based framework to systematically evaluate LLMs by generating questions from KG facts across generic and domain-specific contexts. LLM4EA (Chen S. et al., 2024) aligns KGs using LLM-generated annotations, employing active learning to reduce annotation space and a label refiner to correct noisy labels.

## 6 Challenges in enhancing LLMs with KGs

Although significant progress has been made in augmenting LLM with KGs, key challenges persist. Figure 10 outlines these challenges and indicates potential pathways for further exploration.

### 6.1 Limitations in knowledge acquisition

- 1. Insufficient knowledge coverage and sparsity:** Although large-scale KGs have achieved broad coverage of general knowledge,

they often exhibit limited representation in specialized domains such as medicine and law. In these fields, many entities and relations are either missing or weakly connected. This coverage gap and structural sparsity limit the usefulness of KGs in tasks that require nuanced domain-specific reasoning. Consequently, KG-enhanced LLMs lack access to comprehensive structured support when dealing with emerging diseases, rare events, or complex procedures. Although domain-specific KGs partially address this issue, their integration with LLMs remains challenging due to heterogeneity and scale limitations (Pan et al., 2024).

- 2. High cost and scalability issues in KG construction:** Constructing and maintaining high-quality KGs typically involves significant human effort, including data cleaning, entity alignment, relation labeling, and expert validation. These processes are particularly labor-intensive in domains that require expert knowledge. Although automated or semi-automated KG construction methods, distant supervision, or neural triple extraction have made progress, they often introduce noisy or redundant triples and suffer from low precision in complex contexts. These issues not only degrade the reliability of the KG itself but also reduce the effectiveness of downstream KG-enhanced LLMs, which may propagate errors during inference (Yang et al., 2024a).
- 3. Insufficient multimodal knowledge integration:** Most existing KGs are predominantly constructed from textual data and encode information using structured triples. However, real-world knowledge often exists in multimodal formats such as images, audio, and videos, especially in domains like healthcare, autonomous driving, and robotics. The lack of integrated multimodal knowledge hinders KG-enhanced LLMs in performing tasks that require cross-modal understanding. Although early attempts at constructing multimodal knowledge graphs have shown promise, they are still in their infancy and face challenges in modality alignment, semantic consistency, and large-scale deployment (Chen et al., 2023).

## 6.2 Limitations in knowledge understanding

- Misalignment with natural language semantics:** The structured format of KGs often fails to capture the richness and flexibility of natural language. KG-enhanced LLMs frequently struggle to ground unstructured language into these rigid graph structures. This semantic gap leads to poor retrieval of relevant knowledge and ineffective reasoning over the KG. Although recent methods such as joint graph-text embeddings, prompt-based schema alignment, and co-training frameworks have been proposed to bridge this gap, they often require extensive tuning and are task-specific, lacking robust generalization (Peng et al., 2024).
- Knowledge conflict and redundancy:** KGs derived from multiple sources often contain conflicting or redundant facts. For instance, in the biomedical domain, different datasets may offer contradictory treatments for the same disease or disagree on causality between symptoms and conditions. This inconsistency poses a significant challenge for LLMs enhanced with such knowledge, as it is difficult to determine which facts to trust or prioritize. Although techniques such as triple confidence scoring, contradiction detection, and trust-aware graph filtering have been proposed, current methods remain heuristic-based and fail to generalize across domains and tasks (Wang et al., 2023).
- Difficulty in temporal and dynamic knowledge modeling:** Knowledge in the real world is not static and it evolves over time as new information becomes available. Traditional KGs are static snapshots and lack mechanisms to represent temporal dependencies or model dynamic updates. As a result, KG-enhanced LLMs struggle to reason over sequences of events, causal relationships, or time-sensitive information. Although temporal KGs attempt to incorporate time into graph structures, they are rarely combined with large language models due to scalability concerns and complex modeling requirements (Wang et al., 2023b).

## 6.3 Limitations in knowledge application

- Limited task adaptability:** Most KG-enhanced LLM architectures are designed with a specific task in mind. When applied to different tasks, especially those requiring distinct reasoning paths or domain-specific logic, they often underperform. This is because the integration mechanism is typically static and not tuned to adapt across tasks. While some research has explored multi-task graph encoders and task-specific adapters, there is a lack of a unified and generalizable framework for flexible knowledge integration across diverse LLM tasks (Ibrahim et al., 2024).
- Low real-time inference efficiency:** KG-enhanced LLMs often incur high computational overhead due to the need for graph traversal, entity linking, and dynamic retrieval during inference. These processes introduce latency that hinders the deployment of such systems in real-time applications such as dialogue systems, autonomous agents, and online recommendation. While optimization methods such as graph pruning, caching, and approximate retrieval have been introduced, they either

compromise accuracy or do not scale well with large graphs and multi-user environments (Guo et al., 2022).

## 6.4 Limitations in knowledge explainability

- Opaque reasoning chains and decision logic:** Although KGs are inherently interpretable due to their structured nature, the integration with LLMs often obscures the reasoning path. The fusion of symbolic logic with deep neural networks creates hybrid models where decisions emerge from entangled attention weights and vector operations, making them difficult to trace. Existing explainability techniques have been applied, but they often offer only shallow insight and lack user-centered interpretability (Yixin et al., 2024).
- Unclear knowledge provenance:** In KG-enhanced LLM systems, it is often unclear which knowledge source or KG triple contributes to a particular prediction or generated output. This undermines trust and hinders use in high-stakes domains such as healthcare, law, and finance, where verifiability and source traceability are crucial. Despite recent efforts to tag outputs with provenance metadata or graph node identifiers, these features are rarely integrated into the model architecture in a scalable or user-accessible manner (Pan et al., 2023).

## 7 Collaborative LLMs and KGs(LKC)

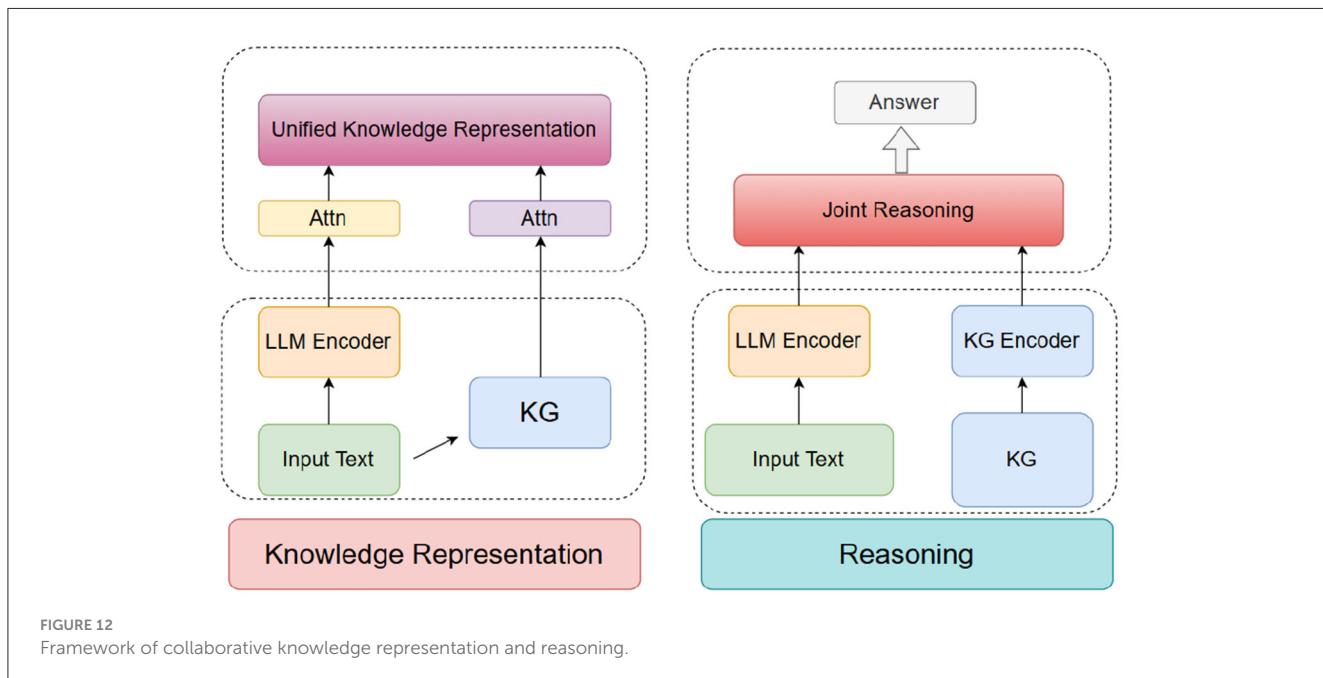
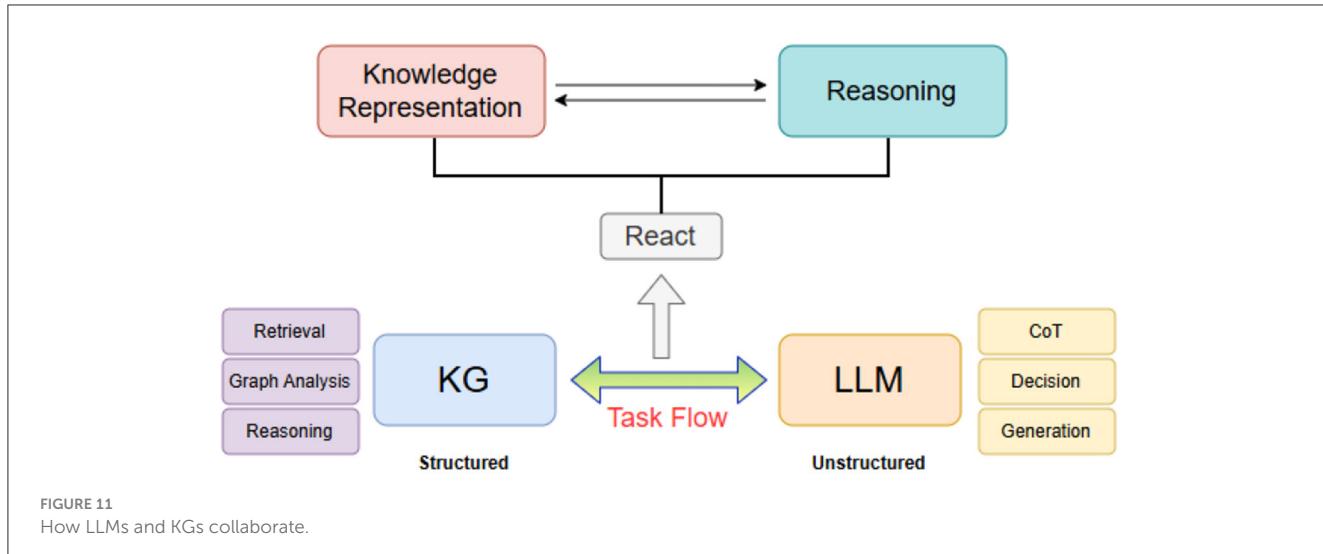
LLMs and KGs have individually demonstrated strengths in various domains. While LLMs excel in reasoning and inference, KGs provide robust frameworks for knowledge representation due to their structured nature. A collaborative approach between LLMs and KGs aims to combine the advantages of both, providing a unified model that can perform well in both knowledge representation and reasoning. Figures 11, 12 illustrate the collaborative mechanisms between LLMs and KGs, with the first figure showing how they interact and the second presenting a framework for collaborative knowledge representation and reasoning.

In this section, we examine the state-of-the-art collaborative models in knowledge representation and reasoning, focusing on studies from 2019–2025 that demonstrate significant advances in bidirectional LLM-KG collaboration.

Our selection prioritizes approaches that establish novel mechanisms for knowledge exchange between neural and symbolic systems, requiring measurable performance improvements over standalone systems on standardized benchmarks. We further emphasize methods offering practical implementations that support real-world deployment. Representative methods meeting these criteria and their key technical contributions are systematically compared in Figure 13, which organizes the selected works by their publication year and innovation type.

### 7.1 Collaborative knowledge representation

Both text corpora and KGs hold valuable information, but they each have limitations. Text corpora may lack structure and factual



consistency, making it challenging to perform precise knowledge extraction and reasoning. KGs, while structured and factual, often require natural language capabilities for more flexible interaction and knowledge understanding. Collaborative approaches between LLMs and KGs aim to combine the strengths of both to form more robust knowledge representations. Such collaborative representations are increasingly demanded in interactive settings like conversational decision support, where users expect both accurate facts and transparent reasoning traces (Amershi et al., 2019).

### 7.1.1 Integrating KG into LLM

This method enhances LLMs by incorporating knowledge from KGs directly into the model, allowing the LLM to benefit from

the structured information of KGs during language understanding tasks. ERNIE (Zhang et al., 2019) integrates KG entities and their relationships into the LLM pre-training process, where entities in the text are masked, and the model is trained to predict them by leveraging the corresponding structured information from KGs. Unlike KG-enhanced pretraining, the following adopt dynamic integration mechanisms. K-BERT (Liu et al., 2020) is a knowledge-based model, in which triples are injected into the sentences as domain knowledge. Also, to overcome knowledge noise(KN), K-BERT introduces softposition and visible matrix to limit the impact of knowledge. BERT-MK (He et al., 2019) uses a dual-encoder system, embedding both entities and their neighboring context from KGs. While these approaches improve factual consistency and entity disambiguation, they face limitations like potential latency and conflicts.

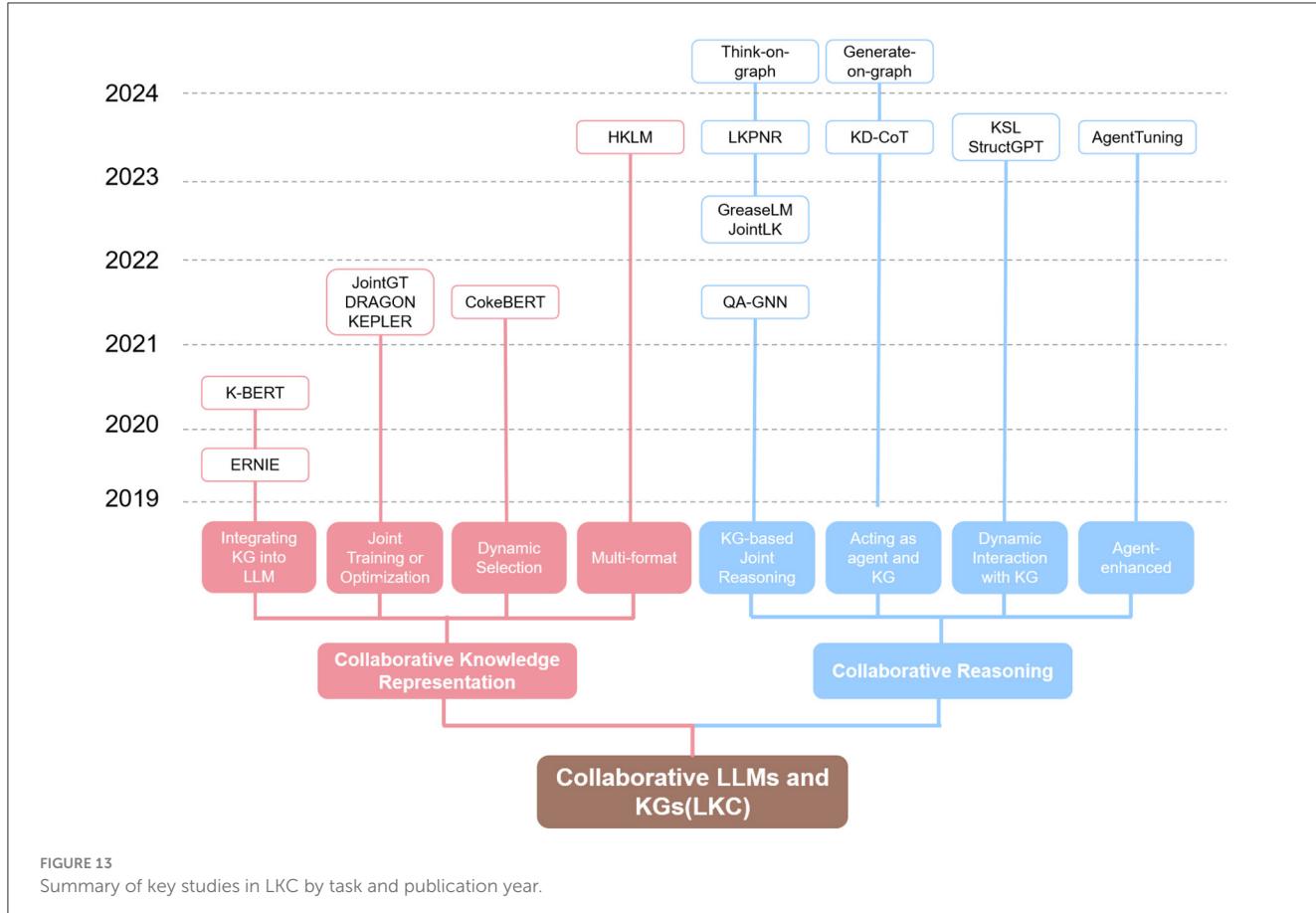


FIGURE 13  
Summary of key studies in LKC by task and publication year.

### 7.1.2 Joint training or optimization

Joint training or optimization approaches train LLMs and KGs together to align them into a unified representation space, where both language and structured knowledge can mutually reinforce each other. JointGT (Ke et al., 2021) proposes a graph-text joint representation learning framework, aiming to align the representations of graph-based and text-based data. By optimizing across tasks like graph-text alignment, node-text matching, and graph-based language modeling, JointGT (Ke et al., 2021) achieves deeper fusion of knowledge and language capabilities. KEPLER (Wang X. et al., 2021) unifies knowledge embedding with language modeling by encoding textual entity descriptions through an LLM, while simultaneously optimizing both the knowledge embedding and language modeling objectives. To compare, JointGT adopts a multi-task training scheme to bridge structural and textual semantics while KEPLER relies on textualized knowledge, jointly optimizing a masked language modeling and knowledge embedding objective. JointGT offers fine-grained control over graph-language alignment, while KEPLER provides a scalable, text-centric solution.

### 7.1.3 Other methods

We list two other strategies here. CokeBERT (Su et al., 2021) dynamically selects and integrates knowledge that are the most relevant KG sub-graphs based on the textual context via a learned relevance scorer, addressing the issue of redundant or irrelevant

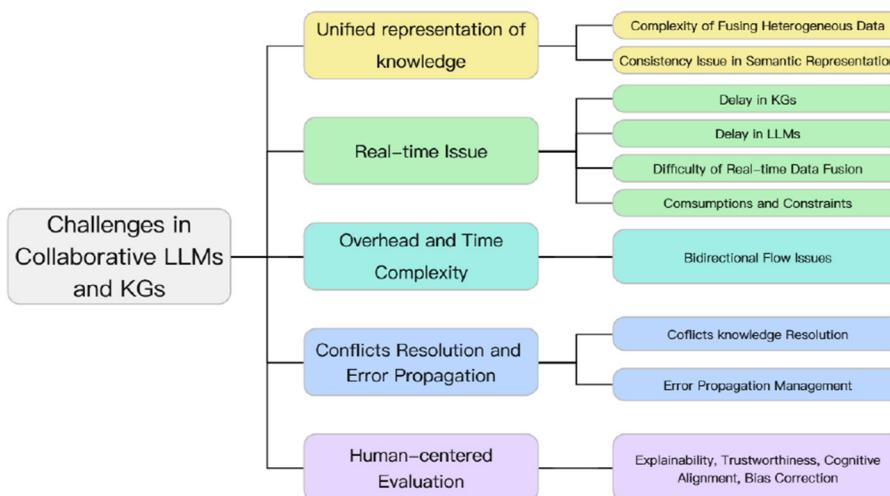
knowledge from KGs. HKLM (Zhu et al., 2023) introduces a multi-format knowledge representation approach, where the model handles unstructured, semi-structured, and structured text simultaneously. This multi-format strategy enhances the model's flexibility in dealing with diverse forms of knowledge representation. These alternative strategies shift the focus from injection to adaptation and format generalization, offering new pathways toward scalable, user-aligned knowledge reasoning. However, they also expose gaps in controllability and transparency, especially when deployed in interactive settings.

## 7.2 Collaborative reasoning

Collaborative reasoning aims to design collaborative models that can effectively conduct reasoning using both LLMs and KGs. These models leverage the structured, factual nature of KGs along with the deep contextual understanding of LLMs to achieve more robust reasoning capabilities.

### 7.2.1 KG-based joint reasoning

KG-based joint reasoning centers around leveraging the structured relational logic of knowledge graphs explicitly, the typical paradigms include GNN-enhanced models and cross-attention mechanisms. For example, QA-GNN (Yasunaga et al.,



**FIGURE 14**  
Challenges in collaborative LLMs and KGs.

2021) utilizes GNNs to reason over KGs while incorporating LLM-based semantic reasoning. The key technology is relevance scoring, where the model estimates the importance of KG nodes concerning a given question, and then applies GNN reasoning to integrate those nodes into the LLM's answer generation. GreaseLM (Zhang X. et al., 2022) employs a layer-wise modality interaction mechanism that tightly integrates a language model (LM) with a GNN, enabling bidirectional reasoning between textual and structured knowledge. JointLK (Sun et al., 2021a) uses a dense bidirectional attention module that connects question tokens with KG nodes, enabling simultaneous interaction between the two. KG nodes attend to question tokens and vice versa, enabling joint reasoning across both LLM-generated representations and KG structures. LKPNR (Runfeng et al., 2023) combines multi-hop reasoning across KGs with LLM context understanding. Think-on-Graph (Sun et al., 2023) treats the LLM as an agent that iteratively executes beam search on a KG, discovering and evaluating reasoning paths. This agent-based framing reflects a move toward interpretable, step-wise reasoning akin to human problem-solving.

### 7.2.2 Acting as both agent and KG

This paradigm breaks the traditional separation between reasoning controller and external knowledge source. The LLM acts as both roles, using its pre-trained knowledge to generate new facts while querying KGs for additional information. Typically, Generate-on-Graph (Xu et al., 2024) treats the LLM in such a paradigm. The LLM explores an incomplete KG and dynamically generates new factual triples conditioned on local graph context. These generated triples are incorporated into the reasoning path, allowing the model to “grow the graph” as it infers—mimicking a constructive reasoning agent. This approach improves robustness in sparse-KG settings. KD-CoT (Wang K. et al., 2023) integrates Chain-of-Thought (CoT) reasoning with knowledge-directed verification. The LLM produces a reasoning trace step-by-step, and

after each step, relevant KG facts are retrieved to validate or revise the intermediate conclusions.

### 7.2.3 Dynamic interaction with KG

It mainly focuses on allowing LLMs to dynamically interact with KGs in real-time, retrieving and updating knowledge during reasoning. KSL (Feng et al., 2023) empowers LLMs to search for essential knowledge from external KGs, transforming retrieval into a multi-hop decision-making process. Constructing APIs for structured data is another method. StructGPT (Jiang J. et al., 2023) creates APIs for structured data access, allowing LLMs to directly interact with structured databases during reasoning.

### 7.2.4 Agent-enhanced

Enhancing LLMs with agent-based capabilities is becoming a hot trend. AgentTuning (Zeng et al., 2023) focuses on enhancing LLMs' agent-like capabilities. By fine-tuning LLMs with structured demonstrations and interaction trajectories, it allows the model to perform more sophisticated reasoning tasks. AgentTuning enables LLMs to interact with knowledge graphs not just as memory sources, but as active environments. As demonstrated in KG retrieval tasks, models trained via AgentTuning can identify the task-relevant knowledge structure, plan multi-step actions, and dynamically query KG APIs, which showcase fine-grained collaboration.

## 8 Challenges in collaborative LLMs and KGs

Significant challenges remain in the collaborative integration of LLMs and KGs, despite promising progress. Figure 14 identifies these challenges and suggests potential solutions for further development.

## 8.1 Unified representation of knowledge

- Complexity of fusing heterogeneous data:** The knowledge sources and structures of KG and LLM have significant differences. KG's knowledge typically comes from structured data, expressed explicitly in the form of entities, relationships, and attributes, relying on manually designed patterns and rules. The knowledge of LLM mainly comes from large-scale text corpora, which capture implicit semantic relationships through unsupervised learning and present them as high-dimensional continuous vector spaces. Hence KGs and LLMs are difficult to align in terms of knowledge granularity, form, and semantics. For example, KG deviates from continuous space and is difficult to embed into the vectorized representation of LLM. The knowledge of LLM is difficult to map to the discrete structure of KG. One of the most critical subproblems under the case is to ensure entity linking pipeline (Shen et al., 2021). This process is non-trivial due to lexical ambiguity, long-tail entities, and incomplete context, especially in open-domain or multi-turn interactive settings. Failures in alignment can reduce explainability. This uncertainty negatively impacts user trust.
- Consistency issue in semantic representation:** The relationships in KG are discrete and explicitly defined, while the semantic relationships in LLM are implicit and distributed. KG may have fuzzy or incomplete knowledge (as an entity may have multiple inconsistent attributes). The knowledge captured by LLM is context sensitive and may be ambiguous due to differences in training corpus and model architecture. For example, KG may record “apples are a type of fruit,” while LLM may infer “apples may also refer to technology companies,” which increases the difficulty of unified representation due to semantic differences. In multi-hop reasoning, where a system must decide whether to rely on linked KG facts or on LLM-internal inference chains. Contradictions or divergence in knowledge may lead to unstable behavior in reasoning paths or QA answers (Zhang X. et al., 2022). If a user receives two subtly different answers depending on which component was consulted, the perceived coherence of the system breaks down. This is particularly critical in sensitive applications like healthcare and finance.

## 8.2 Real-time issue

- Delay in KGs:** KG usually exists in the form of structured data which is static, and its updates and extensions rely on manual design and rule-driven processes, with a long update cycle. KG's knowledge updates are often completed offline in batches, which results in new knowledge not being included in the model in a timely manner, especially in rapidly changing fields such as finance, news, and epidemics, where static KG cannot meet the needs of real-time decision-making. Also it faces scale limitation. As data size and complexity increase, real-time updating of KG may require significant computing and storage resources, further limiting its dynamic capabilities.
- Delay in LLMs:** The real-time performance of LLM also has significant shortcomings. One example is offline training. Most

LLMs are frozen after completing pre training and cannot dynamically learn new knowledge at runtime. (Gao P. et al., 2023) The other is that reasoning relies on historical knowledge. LLM's reasoning is based on the corpus knowledge captured by the model during training, lacking sensitivity to real-time dynamic information.

- Difficulty of real-time data fusion:** The knowledge sources and fusion mechanisms of KG and LLM further exacerbate the challenge of insufficient real-time performance.

**Asynchronous update:** The update mechanisms of KG and LLM are difficult to coordinate. For example, real-time data streams (such as sensor data, social media data) can be generated instantly, but how to synchronize updates and maintain consistency in KG and LLM is a complex task.

**Real time inference bottleneck:** Injecting real-time data dynamically into the fusion system of KG and LLM often requires complex preprocessing, relationship extraction, and context modeling operations, which significantly increases inference time.

- Consumptions and constraints: The update cost of KG is high:** Real-time updating of entities and relationships in KG may require recalculating embeddings and connections, which can introduce significant computational burden in large-scale graphs (Liu J. et al., 2024).

**The inference cost of LLMs is high:** Although generative language models support input dynamic context, the computational cost of generating long texts or complex answers in real-time scenarios is still high, making it difficult to achieve true real-time response.

From the perspective of a user in fast-moving domains, perceiving answers as out-of-date or unsafe would rapidly erode trust in decision-support systems. On the other hand, pursuing real-time performance leads to latency spikes, which reduces conversational fluidity. As a result, users may abandon interactions.

## 8.3 Overhead and time complexity

- Bidirectional flow issues:** A primary challenges in collaborative KGs and LLMs is the overhead and time complexity of managing bidirectional information flow between KGs and LLMs, typically in dynamic interaction. The process of dynamically retrieving knowledge from KGs to inform the LLM's reasoning while simultaneously enriching the KG with new insights or relations generated by the LLM is highly complex. This bidirectional interaction increases the computational overhead and complexity, especially when LLMs need to frequently query large KGs during reasoning.

## 8.4 Conflicts resolution and error propagation

- Resolving conflicting knowledge:** When there is a conflict between the knowledge provided by KG and LLM, a conflict resolution mechanism needs to be established. This may involve knowledge priority rules or confidence calculations.

Such mechanisms often rely on hybrid scoring strategies. However, these scores may not always be directly comparable across modalities or sources. Version control is a typical instance. In dynamic interaction, KG and LLM may be updated simultaneously, requiring an effective version control mechanism to track knowledge changes and ensure consistent results in bidirectional interaction.

## 2. Managing error propagation:

Bidirectional interaction may lead to circular dependencies: If KG updates the information injected into LLM, and LLM then generates knowledge based on this and updates it back to KG, without proper verification and restriction mechanisms, a feedback loop can emerge, which may lead to error propagation. Assume that a knowledge error introduced by either the KG or LLM is not properly filtered, it can be repeatedly propagated, leading to knowledge drift and factual inaccuracies. Such error propagation becomes particularly problematic when generated knowledge is later retrieved as if it were grounded truth, influencing further generations (Saparov and He, 2022). This highlights the need for causal filtering or knowledge provenance tracing, potentially using reinforcement learning to suppress self-reinforcing loops.

## 8.5 Human-centered evaluation

Evaluating collaborative KG and LLM systems is crucial for ensuring their impact on user experience. Effective evaluation not only validates the technical performance (i.e., accuracy or efficiency) of these systems but also ensures they meet user expectations in dynamic, human-facing scenarios, capturing their usability and effectiveness in real scenarios. Unlike traditional tasks, these systems often operate in interactive environments where explainability, trustworthiness, cognitive alignment, and traceability are of great significance. For example, users may require transparency on whether a generated fact was retrieved from the KG or hallucinated by the LLM, or expect the system to adapt its reasoning based on evolving dialogue context. These expectations necessitate evaluation protocols that go beyond static benchmarks, incorporating user-centric metrics such as task success rate, interaction satisfaction, and latency under real-time constraints. However, such human-centered evaluation remains underdeveloped, with limited standardized frameworks for measuring collaborative reasoning quality in real-world, interactive settings. (Kaur et al., 2022)

One typical challenge that illustrates the need is bias propagation. When a biased or incorrect piece of information is introduced by either the KG or the LLM and is subsequently reinforced through iterative reasoning, the system may amplify misleading content without awareness (Bender et al., 2021). This not only compromises factual correctness but also undermines user trust, especially in domains such as healthcare, education, or law. Imagine a knowledge graph encodes historical associations such as “CEO—typically male.” After integrating historical data based patterns into LLM, LLM may output content such as “He is a natural leader and would excel as a CEO.” It may lead to the cyclic spread of gender role bias, and

even exacerbate the structural solidification of “occupational gender stereotypes” within the model, which makes users to doubt the justice of the system. Therefore, evaluation protocols must incorporate fairness and bias-tracking dimensions to assess the long-term human-facing implications of collaborative reasoning systems.

## 9 Overarching discussion across integration paradigms

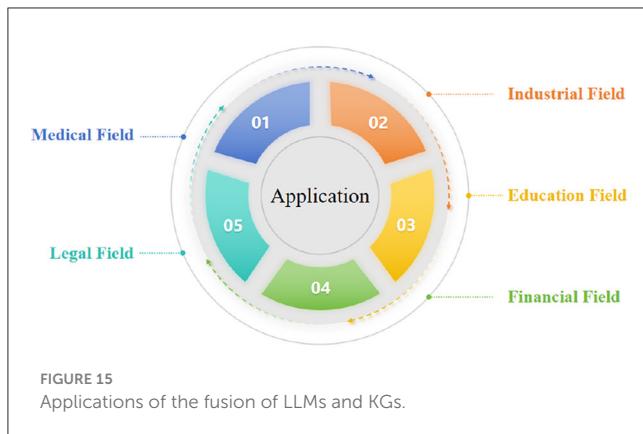
Building upon the detailed analysis of challenges in LLM-enhanced KG (Section 4), KG-enhanced LLM (Section 6), and KG-LLM synergy (Section 8), we identify several significant challenges that persist across all paradigms. The persistent representation gap between neural and symbolic knowledge systems manifests in distinct yet equally problematic ways: creating information fusion barriers in KG construction, causing semantic misalignment in LLM enhancement, and posing integration difficulties in collaborative systems. A second universal challenge involves dynamic knowledge maintenance, which encompasses both the timeliness of KG updates and the limitations of temporal reasoning in LLMs, compounded by real-time processing constraints. Furthermore, we observe an inherent tension between system performance and interpretability that consistently produces explainability-trust dilemmas. These manifest most visibly in opaque reasoning processes, ambiguous knowledge provenance, and growing demands for human-centered evaluation frameworks.

These common issues suggest that future progress will require comprehensive solutions capable of addressing shared architectural constraints while accommodating each paradigm’s specific requirements. By identifying these fundamental challenges, we establish a foundation for developing integrated research directions that could advance all three approaches to KG-LLM integration simultaneously.

## 10 Future directions

### 10.1 Knowledge reflection and dynamic update

Knowledge Reflection and Dynamic Update are key directions in dynamic knowledge graph research, aiming to ensure timeliness, accuracy, and adaptability of knowledge. Knowledge reflection identifies and corrects outdated, conflicting, or incomplete information, continuously refining existing knowledge. Dynamic updates focus on extracting and integrating new knowledge from multi-source data in real time, promoting the continuous evolution of KGs. Future research can leverage the contextual learning capabilities of LLMs to establish a feedback loop of reflection and updating, optimizing the reasoning and updating processes. Existing studies, such as Mou et al. (2024), demonstrate that reflection mechanisms enhance the dynamism and accuracy of knowledge graph construction, offering new insights for the development of adaptive KGs.



## 10.2 Integration of multimodal knowledge graphs and language models

The integration of multimodal KGs and language models (LMs) is a significant frontier in the field of artificial intelligence, aiming to build intelligent systems capable of understanding and reasoning across various modalities, including text, images, audio, and sensor data. Future research will focus on achieving unified representation learning, enabling language models to fully leverage structured and diverse data within multimodal KGs. Additionally, constructing dynamic multimodal KGs will be a key direction, requiring systems to continuously extract, update, and integrate new knowledge from different data streams.

## 10.3 Temporal reasoning

Temporal Reasoning is a significant challenge in AI reasoning, involving the understanding and prediction of temporal logic, causal relationships, and dynamic knowledge. In recent years, with the development of LLMs, new approaches have emerged. Current research primarily addresses the gap between temporal knowledge graphs (TKGs) and LLMs through retrieval-augmented generation frameworks [e.g., GenTKG (Liao et al., 2024)] and reduces computational costs by integrating few-shot learning and instruction tuning. Additionally, models like TG-LLM (Xiong et al., 2024) and chain-of-thought (CoT) reasoning enhance LLMs' ability to comprehend complex temporal logic. Furthermore, generative temporal question-answering frameworks (GenTKGQA) (Gao et al., 2024) achieve efficient reasoning by combining subgraph retrieval with virtual knowledge integration. Future research will focus on optimizing temporal data representation, improving cross-domain generalization, and deeply modeling temporal logic and causal relationships to advance the intelligence and efficiency of temporal reasoning in AI.

# 11 Applications

As shown in Figure 15, the integration of knowledge graphs (KGs) and large language models (LLMs) has been successfully

applied in five key fields: (1) medical, (2) industrial, (3) education, (4) financial, and (5) legal.

## 11.1 Medical field

In the medical domain, the integration of KGs and LLMs has shown immense potential for improving various healthcare applications. One prominent application is the use of KG-enhanced LLMs for medical question answering (QA) (Yang et al., 2024c; Cabello et al., 2024). By combining the structured medical knowledge contained in KGs with LLMs, systems can provide more accurate and contextually relevant answers to complex medical queries. For instance, MEG (Cabello et al., 2024) and LLM-KGMQA (Wang F. et al., 2024) separately integrate graph embeddings from a pre-trained KG encoder into the LLM, and leverage the reasoning capabilities of LLMs to enhance knowledge graph-based QA by refining query interpretations. In addition, KG-enhanced LLMs improve conversational agents by providing them with structured medical knowledge, allowing more informed responses during patient interactions (Varshney et al., 2023).

In biomedical area, projects like CancerKG (Gubanov et al., 2024) leverage large-scale KGs that aggregate cancer-related data from multiple sources. Furthermore, the combination of LLM and KG like DALK (Dynamic Co-Augmentation of LLM and KG) (Li D. et al., 2024) assists researchers by answering complex queries related to the disease, thus accelerating the discovery process.

## 11.2 Industrial field

In the industrial domain, the integration of KGs and LLMs has advanced intelligent systems for tasks such as quality testing and maintenance (Zhou et al., 2024; Su et al., 2024), fault diagnosis (Peifeng et al., 2024; Meng et al., 2022), and process optimization. For example, BERT-BiLSTM-CRF (Meng et al., 2022) integrates BERT, BiLSTM, and CRF modules to identify power equipment entities from Chinese technical documents and extract semantic relationships between entities. Su et al. (2024) combined LLM-based chain-of-thought (CoT) reasoning with a KG to generate highly feasible and coherent test scenarios, supporting exploratory testing and addressing issues such as inconsistent error report quality and infeasible test scenarios.

## 11.3 Education field

In the field of education, KGs help organize and visualize complex learning content, enabling students to better understand and master knowledge. Combined with the natural language capabilities of LLMs, intelligent systems can provide precise learning guidance and personalized recommendations. Jhajj et al. (2024) used GPT-4 to assist in constructing educational knowledge graphs (EduKG), integrating learning objectives and curriculum structures to validate the graphs. Abu-Rasheed et al. (2024) proposed using KGs as factual background prompts for LLMs, designing text templates filled by LLMs to provide accurate and easily understandable learning suggestions.

## 11.4 Financial field

In the financial field, the combination of KGs and LLMs provides robust technological support for financial risk control, fraud detection, and intelligent investment advisory services. By constructing financial KGs, systems can link entities such as enterprises, individuals, and transactions to identify potential risk factors. Additionally, LLMs help to extract information from vast financial reports, news, and transaction records, providing insights for risk assessment and decision-making, as exemplified by FinDKG (Li, 2023). Furthermore, LLM-enhanced KG Q&A systems can deliver financial consulting, helping individuals and enterprises make informed investment decisions.

## 11.5 Legal field

In the legal field, the integration of KGs and LLMs promotes applications such as legal intelligent Q&A (Shi et al., 2024), case prediction (Liu, 2024; Gao S. et al., 2023), and legal document generation. By constructing legal KGs, systems can organize statutes, cases, and precedents, providing structured legal knowledge support for judges, lawyers, and general users. LLMs, with their powerful language generation and reasoning capabilities, utilize these KGs to offer legal consultation, case prediction, and automated legal text generation services.

## 12 Conclusion

This study systematically analyzes three approaches for integrating KGs and LLMs: KEL (KG-enhanced LLMs), LEK (LLM-enhanced KGs), and LKC (collaborative LLMs and KGs). Through a comprehensive review of existing research, we find that such integration can effectively combine the respective strengths of structured knowledge and language models, demonstrating practical value in specific tasks such as question answering systems and decision support. However, due to inherent differences in their knowledge representation and processing methodologies, the actual integration process still faces several key challenges like efficiency issues in real-time knowledge updating and representational consistency in cross-modal learning.

## References

- Abu-Rasheed, H., Weber, C., and Fathi, M. (2024). Knowledge graphs as context sources for llm-based explanations of learning recommendations. *arXiv preprint arXiv:2403.03008*.
- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., et al. (2019). “Guidelines for human-ai interaction,” in *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems*, 1–13. doi: 10.1145/3290605.3300233
- Anelli, V. W., Di Noia, T., Di Sciascio, E., Ferrara, A., and Mancino, A. C. M. (2021). “Sparse feature factorization for recommender systems with knowledge graphs,” in *Proceedings of the 15th ACM Conference on Recommender Systems*, 154–165. doi: 10.1145/3460231.3474243
- Ayoola, T., Tyagi, S., Fisher, J., Christodoulopoulos, C., and Pierleoni, A. (2022). Refined: an efficient zero-shot-capable approach to end-to-end entity linking. *arXiv preprint arXiv:2207.04108*.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). “On the dangers of stochastic parrots: can language models be too big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. doi: 10.1145/3442188.3445922
- Biswas, R., Sack, H., and Alam, M. (2024). Madlink: Attentive multihop and entity descriptions for link prediction in knowledge graphs. *Semant. Web* 15, 83–106. doi: 10.3233/SW-222960
- Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). “Translating embeddings for modeling multi-relational data,” in *Advances in Neural Information Processing Systems*, 26.
- Boylan, J., Mangla, S., Thorn, D., Ghalandari, D. G., Ghaffari, P., and Hokamp, C. (2024). Kgvalidator: a framework for automatic validation of knowledge graph construction. *arXiv preprint arXiv:2404.15923*.

By systematically examining these technical challenges, this study provides directional references for future research.

## Author contributions

LC: Writing – original draft. CY: Writing – original draft. YK: Writing – original draft. YF: Writing – original draft. HZ: Writing – original draft. YZ: Writing – review & editing.

## Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was funded by the National Natural Science Foundation of China (NSFC) under Grant [No.62177007].

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. To provide suggestions for improving the clarity and coherence of the text and to assist in the revision process by suggesting alternative phrasing or wording.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n-gram models of natural language. *Comput. Linguist.* 18, 467–480.
- Cabello, L., Martin-Turrero, C., Akujubi, U., Søgaard, A., and Bobed, C. (2024). Meg: medical knowledge-augmented large language models for question answering. *arXiv preprint arXiv:2411.03883*.
- Caciularu, A., Cohan, A., Beltagy, I., Peters, M. E., Cattan, A., and Dagan, I. (2021). Cdlm: Cross-document language modeling. *arXiv preprint arXiv:2101.00406*.
- Cao, X., and Liu, Y. (2023). Relmk: reasoning with pre-trained language models and knowledge graphs for complex question answering. *Appl. Intell.* 53, 12032–12046. doi: 10.1007/s10489-022-04123-w
- Chen, R., Qin, C., Jiang, W., and Choi, D. (2024). “Is a large language model a good annotator for event extraction?” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 17772–17780. doi: 10.1609/aaai.v38i16.29730
- Chen, S., Zhang, Q., Dong, J., Hua, W., Li, Q., and Huang, X. (2024). Entity alignment with noisy annotations from large language models. *arXiv preprint arXiv:2405.16806*.
- Chen, W., Su, Y., Yan, X., and Wang, W. Y. (2020). Kgpt: Knowledge-grounded pre-training for data-to-text generation. *arXiv preprint arXiv:2010.02307*.
- Chen, X., Lu, T., and Wang, Z. (2024). Llm-align: utilizing large language models for entity alignment in knowledge graphs. *arXiv preprint arXiv:2412.04690*.
- Chen, X., Zhang, J., Wang, X., Zhang, N., Wu, T., Wang, Y., et al. (2023). Continual multimodal knowledge graph construction. *arXiv preprint arXiv:2305.08698*.
- Choi, B., Jang, D., and Ko, Y. (2021). Mem-kgc: masked entity model for knowledge graph completion with pre-trained language model. *IEEE Access* 9, 132025–132032. doi: 10.1109/ACCESS.2021.3113329
- Choi, B., and Ko, Y. (2023). Knowledge graph extension with a pre-trained language model via unified learning method. *Knowl.-Based Syst.* 262:110245. doi: 10.1016/j.knosys.2022.110245
- Colas, A., Alvandipour, M., and Wang, D. Z. (2022). Gap: A graph-aware language model framework for knowledge graph-to-text generation. *arXiv preprint arXiv:2204.06674*.
- Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., and Wei, F. (2021). Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.
- Ding, Y., Zeng, Q., and Weninger, T. (2024). Chatel: entity linking with chatbots. *arXiv preprint arXiv:2402.14858*.
- Ding, Z., Cai, H., Wu, J., Ma, Y., Liao, R., Xiong, B., et al. (2024). “ZRLLM: zero-shot relational learning on temporal knowledge graphs with large language models,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1877–1895. doi: 10.18653/v1/2024.nacl-long.104
- Du, H., Li, C., Zhang, D., and Zhao, D. (2024). “Bi-directional multi-granularity generation framework for knowledge graph-to-text with large language model,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 147–152. doi: 10.18653/v1/2024.acl-short.14
- Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., et al. (2024). From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.
- Fan, W., Geerts, F., and Jia, X. (2007). *Improving Data Quality: Consistency and Accuracy*. New York: ACM.
- Feng, C., Zhang, X., and Fei, Z. (2023). Knowledge solver: Teaching LLMs to search for domain knowledge from knowledge graphs. *arXiv preprint arXiv:2309.03118*.
- Galárraga, L. A., Teflioudi, C., Hose, K., and Suchanek, F. (2013). “Amie: association rule mining on incomplete evidence in ontological knowledge bases,” in *Proceedings of the 22nd International Conference on World Wide Web*, 413–422. doi: 10.1145/2488388.2488425
- Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., et al. (2023). Llama-adapter v2: parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.
- Gao, S., Sa, R., Li, Y., Ge, F., Yu, H., Wang, S., et al. (2023). “How legal knowledge graph can help predict charges for legal text,” in *International Conference on Neural Information Processing* (Springer), 408–420. doi: 10.1007/978-981-99-8076-5\_30
- Gao, Y., Qiao, L., Kan, Z., Wen, Z., He, Y., and Li, D. (2024). Two-stage generative question answering on temporal knowledge graph using large language models. *arXiv preprint arXiv:2402.16568.2402.16568*. doi: 10.18653/v1/2024.findings-acl.401
- Gottschalk, S., and Demidov, E. (2018). “Eventkg: a multilingual event-centric temporal knowledge graph,” in *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings* 15 (Springer), 272–287. doi: 10.1007/978-3-319-93417-4\_18
- Gubanov, M., Pyayt, A., and Karolak, A. (2024). “Cancerkg: org-a web-scale, interactive, verifiable knowledge graph-llm hybrid for assisting with optimal cancer treatment and care,” in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 4497–4505. doi: 10.1145/3627673.3680094
- Guo, L., Bo, Z., Chen, Z., Zhang, Y., Chen, J., Yarong, L., et al. (2025). “MKGL: mastery of a three-word language,” in *Advances in Neural Information Processing Systems*, 140509–140534.
- Guo, Q., Cao, S., and Yi, Z. (2022). A medical question answering system using large language models and knowledge graphs. *Int. J. Intell. Syst.* 37, 8548–8564. doi: 10.1002/int.22955
- Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. (2020). “Retrieval augmented language model pre-training,” in *International Conference on Machine Learning (PMLR)*, 3929–3938.
- Hao, S., Tan, B., Tang, K., Ni, B., Shao, X., Zhang, H., et al. (2022). Bertnet: Harvesting knowledge graphs with arbitrary relations from pretrained language models. *arXiv preprint arXiv:2206.14268*.
- He, B., Zhou, D., Xiao, J., Liu, Q., Yuan, N. J., Xu, T., et al. (2019). Integrating graph contextualized knowledge into pre-trained language models. *arXiv preprint arXiv:1912.00147*.
- Heim, D., Meyer, L.-P., Schröder, M., Frey, J., and Dengel, A. (2025). How do scaling laws apply to knowledge graph engineering tasks? The impact of model size on large language model performance. *arXiv preprint arXiv:2505.16276*.
- Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansky, D., Cohen, V., et al. (2022). True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*.
- Huang, N., Deshpande, Y. R., Liu, Y., Alberts, H., Cho, K., Vania, C., et al. (2022). Endowing language models with multimodal knowledge graph representations. *arXiv preprint arXiv:2206.13163*.
- Ibrahim, N., Aboulela, S., Ibrahim, A., and Kashef, R. (2024). A survey on augmenting knowledge graphs (kgs) with large language models (LLMs): models, evaluation metrics, benchmarks, and challenges. *Disc. Artif. Intell.* 4:76. doi: 10.1007/s44163-024-00175-8
- Jhajj, G., Zhang, X., Gustafson, J. R., Lin, F., and Lin, M. P.-C. (2024). “Educational knowledge graph creation and augmentation via LLMS,” in *International Conference on Intelligent Tutoring Systems* (Springer), 292–304. doi: 10.1007/978-3-031-63031-6\_25
- Jiang, J., Zhou, K., Dong, Z., Ye, K., Zhao, W. X., and Wen, J.-R. (2023). Structgpt: a general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645*.
- Jiang, J., Zhou, K., Zhao, W. X., Song, Y., Zhu, C., Zhu, H., et al. (2024). Kg-agent: an efficient autonomous agent framework for complex reasoning over knowledge graph. *arXiv preprint arXiv:2402.11163*.
- Jiang, P., Agarwal, S., Jin, B., Wang, X., Sun, J., and Han, J. (2023). Text-augmented open knowledge graph completion via pre-trained language models. *arXiv preprint arXiv:2305.15597*.
- Jiang, P., Cao, L., Xiao, C. D., Bhatia, P., Sun, J., and Han, J. (2024). “Kg-fit: knowledge graph fine-tuning upon open-world knowledge,” in *Advances in Neural Information Processing Systems*, 136220–136258.
- Jiang, Z., Xu, F. F., Araki, J., and Neubig, G. (2020). How can we know what language models know? *Trans. Assoc. Comput. Linguist.* 8, 423–438. doi: 10.1162/tacl\_a\_00324
- Kang, M., Baek, J., and Hwang, S. J. (2022). Kala: knowledge-augmented language model adaptation. *arXiv preprint arXiv:2204.10555*.
- Kaur, D., Uslu, S., Rittichier, K. J., and Durresi, A. (2022). Trustworthy artificial intelligence: a review. *ACM Comput. Surv.* 55, 1–38. doi: 10.1145/3491209
- Ke, P., Ji, H., Ran, Y., Cui, X., Wang, L., Song, L., et al. (2021). Jointgt: Graph-text joint representation learning for text generation from knowledge graphs. *arXiv preprint arXiv:2106.10502*.
- Kim, B., Hong, T., Ko, Y., and Seo, J. (2020). “Multi-task learning for knowledge graph completion with pre-trained language models,” in *Proceedings of the 28th International Conference on Computational Linguistics*, 1737–1743. doi: 10.18653/v1/2020.coling-main.153
- Lee, J., Wang, Y., Li, J., and Zhang, M. (2024). Multimodal reasoning with multimodal knowledge graph. *arXiv preprint arXiv:2406.02030*.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). “Retrieval-augmented generation for knowledge-intensive nlp tasks,” in *Advances in Neural Information Processing Systems*, 9459–9474.
- Li, D., Yang, S., Tan, Z., Baik, J. Y., Yun, S., Lee, J., et al. (2024). Dalk: dynamic co-augmentation of LLMs and kg to answer alzheimer’s disease questions with scientific literature. *arXiv preprint arXiv:2405.04819*.
- Li, D., Zhu, B., Yang, S., Xu, K., Yi, M., He, Y., et al. (2023). Multi-task pre-training language model for semantic network completion. *ACM Trans. Asian Low-Resour. Lang. Inf. Proc.* 22, 1–20. doi: 10.1145/3627704
- Li, H., Appleby, G., and Suh, A. (2024). A preliminary roadmap for LLMs as assistants in exploring, analyzing, and visualizing knowledge graphs. *arXiv preprint arXiv:2404.01425*.
- Li, J., Tang, T., Zhao, W. X., Wei, Z., Yuan, N. J., and Wen, J.-R. (2021). Few-shot knowledge graph-to-text generation with pretrained language models. *arXiv preprint arXiv:2106.01623*.

- Li, X. V. (2023). *Findkg: dynamic knowledge graph with large language models for global finance*. Available at SSRN 4608445. doi: 10.2139/ssrn.4608445
- Liao, Q. V., and Vaughan, J. W. (2023). Ai transparency in the age of LLMs: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941, 01910*.
- Liao, R., Jia, X., Li, Y., Ma, Y., and Tresp, V. (2023). Gentkg: Generative forecasting on temporal knowledge graph with large language models. *arXiv preprint arXiv:2310.07793*.
- Liao, R., Jia, X., Li, Y., Ma, Y., and Tresp, V. (2024). "Gentkg: generative forecasting on temporal knowledge graph with large language models," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 4303–4317. doi: 10.18653/v1/2024.findings-naacl.268
- Lin, B. Y., Chen, X., Chen, J., and Ren, X. (2019). Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*.
- Lin, C.-Y. (2004). "Rouge: a package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 74–81.
- Liu, F. (2024). "Design of legal judgment prediction on knowledge graph and deep learning," in *2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA)* (IEEE), 1192–1195. doi: 10.1109/ICIPCA61593.2024.10709293
- Liu, H., Wang, S., Zhu, Y., Dong, Y., and Li, J. (2024). Knowledge graph-enhanced large language models via path selection. *arXiv preprint arXiv:2406.13862*.
- Liu, J., Ke, W., Wang, P., Shang, Z., Gao, J., Li, G., et al. (2024). "Towards continual knowledge graph embedding via incremental distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 8759–8768. doi: 10.1609/aaai.v38i8.28722
- Liu, Q., He, Y., Xu, T., Lian, D., Liu, C., Zheng, Z., et al. (2024). "Unimel: a unified framework for multimodal entity linking with large language models," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 1909–1919. doi: 10.1145/3627673.3679793
- Liu, S., Semnani, S. J., Triedman, H., Xu, J., Zhao, I. D., and Lam, M. S. (2024). Spinach: Sparql-based information navigation for challenging real-world questions. *arXiv preprint arXiv:2407.11417*.
- Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., et al. (2020). "K-bert: enabling language representation with knowledge graph," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2901–2908. doi: 10.1609/aaai.v34i03.5681
- Liu, Y., Tian, X., Sun, Z., and Hu, W. (2024). "Finetuning generative large language models with discrimination instructions for knowledge graph completion," in *International Semantic Web Conference* (Springer), 199–217. doi: 10.1007/978-3-031-77844-5\_11
- Lukovnikov, D., Fischer, A., and Lehmann, J. (2019). "Pretrained transformers for simple question answering over knowledge graphs," in *The Semantic Web-ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I* 18 (Springer), 470–486. doi: 10.1007/978-3-030-30793-6\_27
- Luo, H., Tang, Z., Peng, S., Guo, Y., Zhang, W., Ma, C., et al. (2023). Chatkbqa: a generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models. *arXiv preprint arXiv:2310.08975*.
- Luo, H., Ju, J., Xiong, B., Li, Y.-F., Haffari, G., and Pan, S. (2023a). Chatrule: mining logical rules with large language models for knowledge graph reasoning. *arXiv preprint arXiv:2309.01538*.
- Luo, H., Li, Y.-F., Haffari, G., and Pan, S. (2023b). Reasoning on graphs: faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*.
- Luo, L., Vu, T.-T., Phung, D., and Haffari, G. (2023c). Systematic assessment of factual knowledge in large language models. *arXiv preprint arXiv:2310.11638*.
- Luo, Z., Xie, Q., and Ananiadou, S. (2024). Factual consistency evaluation of summarization in the era of large language models. *Expert Syst. Appl.* 254:124456. doi: 10.1016/j.eswa.2024.124456
- Ma, M. D., Wang, X., Kung, P.-N., Brantingham, P. J., Peng, N., and Wang, W. (2024). "Star: Boosting low-resource information extraction by structure-to-text data generation with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 18751–18759. doi: 10.1609/aaai.v38i17.29839
- Ma, S., Xu, C., Jiang, X., Li, M., Qu, H., and Guo, J. (2024). Think-on-graph 2.0: deep and interpretable large language model reasoning with knowledge graph-guided retrieval. *arXiv e-prints*, arXiv-2407.
- Ma, Y., Wang, A., and Okazaki, N. (2023). Dreeam: Guiding attention with evidence for improving document-level relation extraction. *arXiv preprint arXiv:2302.08675*.
- Meng, F., Yang, S., Wang, J., Xia, L., and Liu, H. (2022). Creating knowledge graph of electric power equipment faults based on bert-bilstm-crf model. *J. Electr. Eng. Technol.* 17, 2507–2516. doi: 10.1007/s42835-022-01032-3
- Meng, Z., Liu, F., Shareghi, E., Su, Y., Collins, C., and Collier, N. (2021). Rewire-then-probe: a contrastive recipe for probing biomedical knowledge of pre-trained language models. *arXiv preprint arXiv:2110.08173*.
- Min, Q., Guo, Q., Hu, X., Huang, S., Zhang, Z., and Zhang, Y. (2024). Synergetic event understanding: a collaborative approach to cross-document event coreference resolution with large language models. *arXiv preprint arXiv:2406.02148*.
- Mirkhani, K., Chaggares, C., Masterson, C., Jastrzebski, M., Dusatko, T., Sinclair, A., et al. (2004). Optimal design of emat transmitters. *NDT e Int.* 37, 181–193. doi: 10.1016/j.ndteint.2003.09.005
- Moiseev, F., Dong, Z., Alfonseca, E., and Jaggi, M. (2022). Skill: Structured knowledge infusion for large language models. *arXiv preprint arXiv:2205.08184*.
- Mou, Y., Liu, L., Sowe, S., Collarana, D., and Decker, S. (2024). "Leveraging LLMs few-shot learning to improve instruction-driven knowledge graph construction," in *Proceedings of the VLDB Endowment*. ISSN 2150, 8097.
- Nath, A., Manafi, S., Chelle, A., and Krishnaswamy, N. (2024). Okay, let's do this! Modeling event coreference with generated rationales and knowledge distillation. *arXiv preprint arXiv:2404.03196*.
- Pan, J. Z., Razniewski, S., Kalo, J.-C., Singhania, S., Chen, J., Dietze, S., et al. (2023). Large language models and knowledge graphs: opportunities and challenges. *arXiv preprint arXiv:2308.06374*.
- Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., and Wu, X. (2024). Unifying large language models and knowledge graphs: a roadmap. *IEEE Trans. Knowl. Data Eng.* 36, 3580–3599. doi: 10.1109/TKDE.2024.3352100
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318. doi: 10.3115/1073083.1073135
- Peifeng, L., Qian, L., Zhao, X., and Tao, B. (2024). Joint knowledge graph and large language model for fault diagnosis and its application in aviation assembly. *IEEE Trans. Ind. Inform.* 20, 8160–8169. doi: 10.1109/TII.2024.3366977
- Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., et al. (2024). Graph retrieval-augmented generation: a survey. *arXiv preprint arXiv:2408.08921*.
- Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., et al. (2019). Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.
- Pradeep, R., Lee, D., Mousavi, A., Pound, J., Sang, Y., Lin, J., et al. (2024). Convkgarn: Spinning configurable and scalable conversational knowledge graph qa datasets with large language models. *arXiv preprint arXiv:2408.05948*.
- Rabiner, L., and Juang, B. (1986). An introduction to hidden markov models. *IEEE Assp Magaz.* 3, 4–16. doi: 10.1109/MASSP.1986.1165342
- Razouk, H., Liu, X. L., and Kern, R. (2023). Improving fmea comprehensibility via common-sense knowledge graph completion techniques. *IEEE Access* 11, 127974–127986. doi: 10.1109/ACCESS.2023.3331585
- Ren, X., Tang, J., Yin, D., Chawla, N., and Huang, C. (2024). "A survey of large language models for graphs," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6616–6626. doi: 10.1145/3637528.3671460
- Runfeng, X., Xiangyang, C., Zhou, Y., Xin, W., Zhanwei, X., Kai, Z., et al. (2023). LKPNR: LLM and kg for personalized news recommendation framework. *arXiv preprint arXiv:2308.12028*.
- Sanmartin, D. (2024). KG-RAG: bridging the gap between knowledge and creativity. *arXiv preprint arXiv:2405.12035*.
- Sansford, H., Richardson, N., Maretic, H. P., and Saada, J. N. (2024). Grapheval: a knowledge-graph based llm hallucination evaluation framework. *arXiv preprint arXiv:2407.10793*.
- Saparov, A., and He, H. (2022). Language models are greedy reasoners: a systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.
- Saxena, A., Kochsiek, A., and Gemulla, R. (2022). Sequence-to-sequence knowledge graph completion and question answering. *arXiv preprint arXiv:2203.10321*.
- Shen, J., Wang, C., Gong, L., and Song, D. (2022). Joint language semantic and structure embedding for knowledge graph completion. *arXiv preprint arXiv:2209.08721*.
- Shen, W., Li, Y., Liu, Y., Han, J., Wang, J., and Yuan, X. (2021). Entity linking meets deep learning: Techniques and solutions. *IEEE Trans. Knowl. Data Eng.* 35, 2556–2578. doi: 10.1109/TKDE.2021.3117715
- Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D* 404:132306. doi: 10.1016/j.physd.2019.132306
- Shi, J., Guo, Q., Liao, Y., Wang, Y., Chen, S., and Liang, S. (2024). "LEGAL-LLM: knowledge graph enhanced large language models for law consulting," in *International Conference on Intelligent Computing* (Springer), 175–186. doi: 10.1007/978-981-97-5672-8\_15
- Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. (2020). Autoprompt: eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.
- Shu, D., Chen, T., Jin, M., Zhang, C., Du, M., and Zhang, Y. (2024). Knowledge graph large language model (KG-LLM) for link prediction. *arXiv preprint arXiv:2403.07311*.
- Su, D., Xu, Y., Winata, G. I., Xu, P., Kim, H., Liu, Z., et al. (2019). "Generalizing question answering system with pre-trained language model fine-tuning," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 203–211.

- Su, Y., Han, X., Zhang, Z., Lin, Y., Li, P., Liu, Z., et al. (2021). Cokebert: contextual knowledge selection and embedding towards enhanced pre-trained language models. *AI Open* 2, 127–134. doi: 10.1016/j.aiopen.2021.06.004
- Su, Y., Liao, D., Xing, Z., Huang, Q., Xie, M., Lu, Q., et al. (2024). “Enhancing exploratory testing by large language model and knowledge graph,” in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 1–12. doi: 10.1145/3597503.3639157
- Su, Z., Wang, D., Miao, C., and Cui, L. (2023). “Multi-aspect explainable inductive relation prediction by sentence transformer,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 6533–6540. doi: 10.1609/aaai.v37i5.25803
- Sun, J., Xu, C., Tang, L., Wang, S., Lin, C., Gong, Y., et al. (2023). Think-on-graph: deep and responsible reasoning of large language model on knowledge graph. *arXiv preprint arXiv:2307.07697*.
- Sun, L., Zhang, P., Gao, F., An, Y., Li, Z., and Zhao, Y. (2025). SF-GPT: a training-free method to enhance capabilities for knowledge graph construction in LLMs. *Neurocomputing* 613:128726. doi: 10.1016/j.neucom.2024.128726
- Sun, T., Shao, Y., Qiu, X., Guo, Q., Hu, Y., Huang, X., et al. (2020). Colake: contextualized language and knowledge embedding. *arXiv preprint arXiv:2010.00309*.
- Sun, Y., Shi, Q., Qi, L., and Zhang, Y. (2021a). Jointlk: joint reasoning with language models and knowledge graphs for commonsense question answering. *arXiv preprint arXiv:2112.02732*.
- Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., et al. (2021b). Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.
- Sung, M., Lee, J., Yi, S., Jeon, M., Kim, S., and Kang, J. (2021). Can language models be biomedical knowledge bases? *arXiv preprint arXiv:2109.07154*.
- Tao, Z., Chen, X., Jin, Z., Bai, X., Zhao, H., and Lou, Y. (2024). Evit: Event-oriented instruction tuning for event reasoning. *arXiv preprint arXiv:2404.11978*.
- Tian, Y., Song, H., Wang, Z., Wang, H., Hu, Z., Wang, F., et al. (2024). “Graph neural prompting with large language models,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 19080–19088. doi: 10.1609/aaai.v38i17.29875
- Varshney, D., Zafar, A., Behera, N. K., and Ebkal, A. (2023). Knowledge graph assisted end-to-end medical dialog generation. *Artif. Intell. Med.* 139:102535. doi: 10.1016/j.artmed.2023.102535
- Vaswani, A. (2017). “Attention is all you need,” in *Advances in Neural Information Processing Systems*.
- Waldock, W. J., Zhang, J., Guni, A., Nabeel, A., Darzi, A., and Ashrafiyan, H. (2024). The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: systematic review and meta-analysis. *J. Med. Internet Res.* 26:e56532. doi: 10.2196/56532
- Wang, B., Shen, T., Long, G., Zhou, T., Wang, Y., and Chang, Y. (2021). “Structure-augmented text representation learning for efficient knowledge graph completion,” in *Proceedings of the Web Conference 2021*, 1737–1748. doi: 10.1145/3442381.3450043
- Wang, F., Shi, D., Aguilar, J., Cui, X., Jiang, J., Shen, L., et al. (2024). Llm-kgmqa: Large language model-augmented multi-hop question-answering system based on knowledge graph in medical field. *Knowl. Inf. Syst.* 2025:s1115. doi: 10.1007/s10115-025-02399-1
- Wang, J., Huang, W., Shi, Q., Wang, H., Qiu, M., Li, X., et al. (2022). Knowledge prompting in pre-trained language model for natural language understanding. *arXiv preprint arXiv:2210.08536*.
- Wang, J., Sun, Q., Li, X., and Gao, M. (2023a). Boosting language models reasoning with chain-of-knowledge prompting. *arXiv preprint arXiv:2306.06427*.
- Wang, J., Wang, B., Qiu, M., Pan, S., Xiong, B., Liu, H., et al. (2023b). A survey on temporal knowledge graph completion: Taxonomy, progress, and prospects. *arXiv preprint arXiv:2308.02457*.
- Wang, K., Duan, F., Wang, S., Li, P., Xian, Y., Yin, C., et al. (2023). Knowledge-driven cot: Exploring faithful reasoning in LLMs for knowledge-intensive question answering. *arXiv preprint arXiv:2308.13259*.
- Wang, K., Xu, Y., Wu, Z., and Luo, S. (2024). Llm as prompter: low-resource inductive reasoning on arbitrary knowledge graphs. *arXiv preprint arXiv:2402.11804*.
- Wang, L., Zhao, W., Wei, Z., and Liu, J. (2022). SIMKGC: simple contrastive knowledge graph completion with pre-trained language models. *arXiv preprint arXiv:2203.02167*.
- Wang, P., Xie, X., Wang, X., and Zhang, N. (2023). “Reasoning through memorization: Nearest neighbor knowledge graph embeddings,” in *CCF International Conference on Natural Language Processing and Chinese Computing* (Springer), 111–122. doi: 10.1007/978-3-031-44693-1\_9
- Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., et al. (2023). GPT-ner: named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.
- Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., et al. (2021). Kepler: a unified model for knowledge embedding and pre-trained language representation. *Trans. Assoc. Comput. Linguist.* 9, 176–194. doi: 10.1162/tacl\_a\_00360
- Wang, X., He, Q., Liang, J., and Xiao, Y. (2022). Language models as knowledge embeddings. *arXiv preprint arXiv:2206.12617*.
- Wang, Y., Feng, S., Wang, H., Shi, W., Balachandran, V., He, T., et al. (2023). Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.
- Wang, Y., Hu, M., Huang, Z., Li, D., Yang, D., and Lu, X. (2024). Kc-genre: a knowledge-constrained generative re-ranking method based on large language models for knowledge graph completion. *arXiv preprint arXiv:2403.17532*.
- Wen, Y., Wang, Z., and Sun, J. (2023). Mindmap: knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*.
- Xie, X., Zhang, N., Li, Z., Deng, S., Chen, H., Xiong, F., et al. (2022). “From discrimination to generation: Knowledge graph completion with generative transformer,” in *Companion Proceedings of the Web Conference 2022*, 162–165. doi: 10.1145/3487553.3524238
- Xin, A., Qi, Y., Yao, Z., Zhu, F., Zeng, K., Bin, X., et al. (2024). LLMAEL: Large language models are good context augmenters for entity linking. *arXiv:2407.04020*.
- Xiong, S., Payani, A., Kompella, R., and Fekri, F. (2024). Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853*.
- Xiong, W., Du, J., Wang, W. Y., and Stoyanov, V. (2019). Pretrained encyclopedic weakly supervised knowledge-pretrained language model. *arXiv preprint arXiv:1912.09637*.
- Xu, W., Liu, B., Peng, M., Jia, X., and Peng, M. (2023). Pre-trained language model with prompts for temporal knowledge graph completion. *arXiv preprint arXiv:2305.07912*.
- Xu, Y., He, S., Chen, J., Wang, Z., Song, Y., Tong, H., et al. (2024). Generate-on-graph: treat llm as both agent and kg in incomplete knowledge graph question answering. *arXiv preprint arXiv:2404.14741*.
- Yang, L., Chen, H., Li, Z., Ding, X., and Wu, X. (2024a). Give us the facts: enhancing large language models with knowledge graphs for fact-aware language modeling. *IEEE Trans. Knowl. Data Eng.* 36, 3091–3110. doi: 10.1109/TKDE.2024.3360454
- Yang, L., Chen, H., Wang, X., Yang, J., Wang, F.-Y., and Liu, H. (2024b). Two heads are better than one: Integrating knowledge from knowledge graphs and large language models for entity alignment. *arXiv preprint arXiv:2401.16960*.
- Yang, R., Liu, H., Marrese-Taylor, E., Zeng, Q., Ke, Y. H., Li, W., et al. (2024c). Kg-rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. *arXiv preprint arXiv:2403.05881*.
- Yang, R., Yang, B., Ouyang, S., She, T., Feng, A., Jiang, Y., et al. (2024d). Graphusion: Leveraging large language models for scientific knowledge graph fusion and construction in nlp education. *arXiv preprint arXiv:2407.10794*.
- Yao, L., Mao, C., and Luo, Y. (2019). Kg-bert: bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.
- Yasunaga, M., Ren, H., Bosselut, A., Liang, P., and Leskovec, J. (2021). Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*. doi: 10.18653/v1/2021.nacl-main.45
- Ye, H., Zhang, N., Deng, S., Chen, X., Chen, H., Xiong, F., et al. (2022). “Ontology-enhanced prompt-tuning for few-shot learning,” in *Proceedings of the ACM Web Conference 2022*, 778–787. doi: 10.1145/3485447.3511921
- Yixin, X., Zongbao, Y., Yuchen, L., Jinlong, H., and Shoubin, D. (2024). Interpretable biomedical reasoning via deep fusion of knowledge graph and pre-trained language models. *Beijing Da Xue Xue Bao* 60, 62–70. doi: 10.13209/j.0479-8023.2023.073
- You, J., and Tagkopoulos, I. (2022). Kglm: Integrating knowledge graph structure in language models for link prediction. *arXiv preprint arXiv:2211.02744*.
- Yu, D., Zhu, C., Yang, Y., and Zeng, M. (2022). Jaket: joint pre-training of knowledge graph and language understanding,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 11630–11638. doi: 10.1609/aaai.v36i10.21417
- Zeng, A., Liu, M., Lu, R., Wang, B., Liu, X., Dong, Y., et al. (2023). Agenttuning: enabling generalized agent abilities for LLMs. *arXiv preprint arXiv:2310.12823*.
- Zha, H., Chen, Z., and Yan, X. (2022). “Inductive relation prediction by bert,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 5923–5931. doi: 10.1609/aaai.v36i5.20537
- Zhang, D., Yuan, Z., Liu, Y., Zhuang, F., Chen, H., and Xiong, H. (2020). E-bert: a phrase and product knowledge enhanced language model for e-commerce. *arXiv preprint arXiv:2009.02835*.
- Zhang, M., Dai, R., Dong, M., and He, T. (2022). “Drlk: dynamic hierarchical reasoning with language model and knowledge graph for question answering,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5123–5133. doi: 10.18653/v1/2022.emnlp-main.342
- Zhang, M., Yang, G., Liu, Y., Shi, J., and Bai, X. (2024). Knowledge graph accuracy evaluation: an llm-enhanced embedding approach. *Int. J. Data Sci. Anal.* 2024, 1–15. doi: 10.1007/s41060-024-00661-3
- Zhang, R., Su, Y., Trisedy, B. D., Zhao, X., Yang, M., Cheng, H., et al. (2023). Autoalign: fully automatic and effective knowledge graph alignment enabled by large language models. *IEEE Trans. Knowl. Data Eng.* 36, 2357–2371. doi: 10.1109/TKDE.2023.3325484

- Zhang, T., Wang, C., Hu, N., Qiu, M., Tang, C., He, X., et al. (2022). "DKPLM: decomposable knowledge-enhanced pre-trained language model for natural language understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 11703–11711. doi: 10.1609/aaai.v36i10.21425
- Zhang, X., Bosselut, A., Yasunaga, M., Ren, H., Liang, P., Manning, C. D., et al. (2022). Greaselm: graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860*.
- Zhang, Y., Chen, Z., Guo, L., Xu, Y., Zhang, W., and Chen, H. (2023). Making large language models perform better in knowledge graph completion. *arXiv preprint arXiv:2310.06671*.
- Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2019). Ernie: enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.
- Zhang, Z., Lee, S., Wu, J., Zhang, D., Li, S., Cambria, E., et al. (2024a). "Cross-domain ner with generated task-oriented knowledge: an empirical study from information density perspective," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1595–1609. doi: 10.18653/v1/2024.emnlp-main.95
- Zhang, Z., Liu, X., Zhang, Y., Su, Q., Sun, X., and He, B. (2020). "Pretrain-kge: learning knowledge representation from pretrained language models," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 259–266. doi: 10.18653/v1/2020.findings-emnlp.25
- Zhang, Z., Wang, Y., Wang, C., Chen, J., and Zheng, Z. (2024b). Llm hallucinations in practical code generation: Phenomena, mechanism, and mitigation. *arXiv preprint arXiv:2409.20550*.
- Zhao, R., Zhao, F., Wang, L., Wang, X., and Xu, G. (2024). "Kg-cot: chain-of-thought prompting of large language models over knowledge graphs for knowledge-aware question answering," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)* (International Joint Conferences on Artificial Intelligence), 6642–6650. doi: 10.24963/ijcai.2024/734
- Zheng, J., Liang, M., Yu, Y., Li, Y., and Xue, Z. (2024). "Knowledge graph enhanced multimodal transformer for image-text retrieval," in *2024 IEEE 40th International Conference on Data Engineering (ICDE)* (IEEE), 70–82. doi: 10.1109/ICDE60146.2024.00013
- Zheng, L., Chen, B., Fei, H., Li, F., Wu, S., Liao, L., et al. (2024). "Self-adaptive fine-grained multi-modal data augmentation for semi-supervised muti-modal coreference resolution," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 8576–8585. doi: 10.1145/3664647.3680966
- Zhou, B., Li, X., Liu, T., Xu, K., Liu, W., and Bao, J. (2024). Causalkgpt: industrial structure causal knowledge-enhanced large language model for cause analysis of quality problems in aerospace product manufacturing. *Adv. Eng. Inform.* 59:102333. doi: 10.1016/j.aei.2023.102333
- Zhu, H., Peng, H., Lyu, Z., Hou, L., Li, J., and Xiao, J. (2023). Pre-training language model incorporating domain-specific heterogeneous knowledge into a unified representation. *Expert Syst. Appl.* 215:119369. doi: 10.1016/j.eswa.2022.119369