Graph-Assisted Large Language Models: A Perspective on Mitigating Intrinsic Limitations

Anonymous ACL submission

Abstract

Large language models (LLMs) have made progress in knowledge-intensive tasks, reasoning and planning, and collaborative problem solving, yet they exhibit intrinsic limitations such as knowledge cutoff, single-threaded reasoning that hinders finer-grained branch and aggregation, and rigid collaboration mechanisms that struggle to coordinate specialized capabilities. Graphs, with their ability to represent relational knowledge and complex dependencies, offer a natural means to address these limitations: they provide structured, high-density knowledge for augmenting or correcting LLMs' generation; enable revisitable inference by organizing intermediate steps as graphs; and support dynamic coordination among experts or agents in collaborative settings. Motivated by these developments, we present the first systematic survey of graph-assisted LLMs from the perspective of how graph structures mitigate LLMs' limitations. We introduce a taxonomy spanning *Graph-Assisted Knowledge Augmentation, Graph-Assisted Reasoning and Planning, and Graph-Assisted LLM Collaboration*, and analyze representative methods, summarize common design patterns, and outline open challenges and future directions for advancing LLMs with graph-based enhancements.

1 Introduction

Large language models (LLMs) exhibit strong generalization and adaptability across diverse applications, serving as compact parametric knowledge bases in knowledge-intensive tasks (Wang et al., 2024a; Dernbach et al., 2024), producing multi-step reasoning chains for reliable problem solving (Yao et al., 2023; Wang et al., 2025a), and collaborating via multi-agent systems or mixture-of-experts to handle dynamic task demands (Shinn et al., 2024; Dong et al., 2024; Wang et al., 2025c).

Despite their strong capabilities, LLMs also exhibit inherent limitations in knowledge, reasoning, and collaboration. **(i) Knowledge limitations.** Because LLMs learn from static text corpora, their parametric knowledge suffers from *knowledge cut-off* (Agarwal et al., 2023; Wu et al., 2024a), factual errors arising from noisy training data (Huang et al., 2025a), and *unintentional memorization* of sensitive or copyrighted content (Yang et al., 2025; Qiu et al., 2024a). **(ii) Structured reasoning limitations.** Token-by-token generation forces reasoning into a linear "single-thread" process in which early mistakes propagate (Ren et al., 2024; Li et al., 2023b; Luo et al., 2024a), preventing backtracking or branching and making LLMs unreliable for tasks requiring network-structured inference. **(iii) Collaboration limitations.** Although multi-LLM and mixture-of-expert-based approaches aim to combine complementary capabilities, LLMs lack mechanisms for effective coordination, and mixture-of-expert (MoE) or multi-agent system (MAS) degrade when routing selects suboptimal experts or agents (Zhang et al., 2024a; Zhao et al., 2024).

Graphs, as a canonical structured data form, are well-suited for storing knowledge, guiding LLMs toward structured outputs, and orchestrating experts/agents, thereby addressing key limitations in knowledge, reasoning, and collaboration. **(i) Knowledge enhancement.** Knowledge graphs can mitigate issues such as knowledge cutoff and factual errors (Jiang et al., 2025a; Chen et al., 2025c) and help identify or remove unintentionally memorized content (Yang et al., 2025) by providing accurate and up-to-date knowledge. In addition, their structured triples, high information density, and rich resources (Bodenreider, 2004; GeoNames, 2004) facilitate graph-based knowledge injection, editing, and unlearning of LLMs. **(ii) Reasoning and planning enhancement.** While LLMs can perform multi-step reasoning, their reasoning traces typically remain linear, limiting exploration of alternative paths. By introducing an explicit topology of reasoning into the reasoning process, graph structures enable parallel branching and revisitable inference (Yao et al., 2023; Besta et al., 2024), improving the robustness of complex planning. **(iii) Collaboration enhancement.** The interaction and coordination of experts (Bai et al., 2024) or agents (Zhang et al., 2025h) naturally forms a collaboration graph. Explicitly modeling these interactions allows for dynamic graph optimization and adaptive module selection, facilitating effective orchestration in both MoE and MAS settings.

A wide range of graph-enhanced LLM methods have been proposed for knowledge augmentation (Chen et al., 2025a; Zhang et al., 2024d), reasoning and planning (Yao et al., 2023; Besta et al., 2024), and LLM collaboration (Zhang et al., 2024a; Zhao et al., 2024) (more works see Tab. 1, 2, 3), leveraging graph structures to address core LLM limitations. Yet, the field lacks a systematic survey examining these approaches from the perspective of how graphs help mitigate such limitations. To fill this gap, we categorize graph-assisted LLMs into three aspects: **Graph-Assisted Knowledge Augmentation, Structural Reasoning and Planning, and LLM Collaboration**, and develop a taxonomy spanning parametric and non-parametric knowledge augmentation, knowledge validation/-correction, graph-assisted reasoning/planning, and collaboration (Sec. 2 and Fig.12). We then summarize key mechanisms that graphs help LLMs and outline promising research directions.

Differences from Existing Surveys. While various surveys examine graph applications in the LLM era, the majority (Liu et al., 2023; Li et al., 2023b; Jin et al., 2024; Ren et al., 2024; Huang et al., 2024a; Wang et al., 2025i; Wei et al., 2025a) focus on *LLMs for Graphs*, with limited discussion on the reciprocal benefits of graphs for LLMs. Surveys that do address *Graphs for LLMs* are often restricted to single dimensions: Pan et al. (2024); Yang et al. (2024a); Ma et al. (2025a) focus on KG-enhanced LLMs; Zhang et al. (2025g); Procko and Ochoa (2024); Peng et al. (2024) on Graph RAG; and Bei et al. (2025); Liu et al. (2025) on graph-based agents. We fill this gap by providing a holistic survey from the perspective of **LLM limitations**, synthesizing how graphs address deficits in knowledge, reasoning, and collaboration.

2 Taxonomy

This section summarizes graph-assisted LLMs into five paradigms (Fig. 12). To address the three core limitations, we divide the knowledge enhancement into three functional categories based on their operational stage, while maintaining reasoning and collaboration as distinct structural enhancements.

- **Parametric Knowledge Augmentation.** Knowledge graphs support parametric knowledge within LLMs, including: ❶ *knowledge injection*, where KG information is incorporated through fine-tuning; ❷ *knowledge editing*, which leverages KGs to modify specific internal facts; ❸ *knowledge unlearning*, where KGs guide the removal of undesired or sensitive knowledge.
- **Non-Parametric Knowledge Augmentation.** Graphs support non-parametric augmentation in two settings: ❶ *external RAG*, using graphs as indices or stores over external knowledge resources; ❷ *internal memory*, using graphs to index or store system- and user-derived memory.
- **Knowledge Validation and Correction.** Knowledge graphs support post-inference knowledge enhancement: ❶ *post-inference validation*, which checks the correctness of generated content; ❷ *post-inference correction*, which revises erroneous outputs using KG evidence.
- **Graph-Assisted Reasoning and Planning.** Reasoning and planning are formalized as graph-structured processes to improve reliability: ❶ *graph-structured reasoning* organizes intermediate reasoning steps into graph form to enable multi-branch reasoning; ❷ *graph-structured planning* models tasks, tools, and environments as graphs to guide complex action sequences.
- **Graph-Assisted Collaboration.** Collaboration among LLMs is represented as a graph to facilitate optimization and adaptation to dynamic task demands: ❶ *graph-assisted MoE* optimizes expert graphs for dynamic expert selection; ❷ *graph-assisted MAS* organizes agent interactions for effective multi-agent collaboration.

3 Parametric Knowledge Augmentation

LLMs suffer from knowledge limitations, including outdated information, factual errors, and unintended memorization. To address these issues, recent work explores parametric knowledge operations, including *injection*, *modification*, and *unlearning*. While many operate on text, KGs provide greater control and interpretability. This section reviews KG-based approaches for injecting, editing, and unlearning LLMs' parametric knowledge.

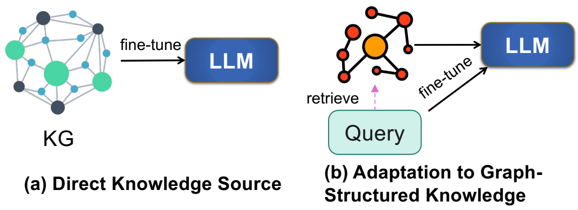

Figure 1: Full-parameter injecting KGs into LLMs.

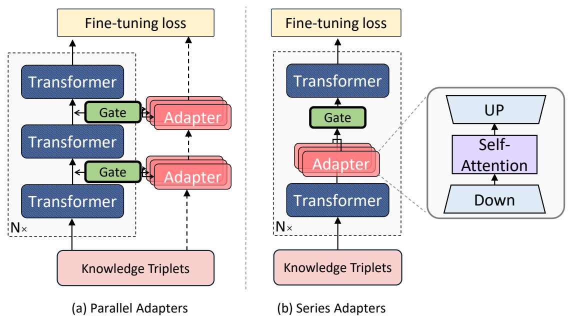

Figure 2: Adapter-based injecting KGs into LLMs.

3.1 Knowledge Injection

By the update scope, KG-based parametric knowledge injection includes (1) full-parameter fine-tuning, which updates all parameters, and (2) adapter-based injection, which infuses KG knowledge via adapters while freezing the base model.

(1) Full-Parameter Fine-Tuning Techniques. KGs play multifaceted roles in full-parameter knowledge injection, serving both as direct knowledge sources and as structure mentors that help LLMs' adapt to graph-structured information (Fig. 1). *(i) Direct Knowledge Source:* Methods such as KITLM (Agarwal et al., 2023) convert KG triples into text for next-token training, while GraphVis (Deng et al., 2024) encodes subgraphs as visual inputs for multimodal tuning. PMC-LLaMA (Wu et al., 2024a) and SKILL (Moiseev et al., 2022) use entity-centric KG explanations for continual learning. SSQR (Lin et al., 2025) aligns KG entities with LLM embeddings, enabling efficient vocabulary expansion. *(ii) Adaptation to Graph-Structured Information:* GLaM (Dernbach et al., 2024) and ELPF (Jiang et al., 2024b) fine-tune models on textualized subgraphs and reasoning QA, while GALLa (Zhang et al., 2025j) integrates code graphs via Graph2Code and GraphQA to better align graph and language representations.

(2) Adapter-Based Injection Methods. To reduce computational cost, recent work uses adapters to inject knowledge from KGs without updating base parameters, primarily by converting KG triples into natural language for targeted adapter training (Fig. 2). The adapters mainly treat the KGs as knowledge sources. *(i) Parallel Adapters:* These operate alongside Transformer layers, merging adapter and model outputs. For example, InfuserKI (Wang et al., 2024a) injects missing KG knowledge via gated parallel adapters; MixDA (Diao et al., 2023) employs a mixture-of-adapters with adaptive gating; K-Adapter (Wang et al., 2021) enables modular domain-specific knowledge injection; LightPROF (Ao et al., 2025) encodes KG prompts using a graph token; STR-CMP (Li et al., 2025c) integrates GNN-derived structural embeddings for combinatorial optimization. *(ii) Series Adapters:* Inserted within Transformer stacks, they jointly propagate KG and LLM embeddings. KG-Adapter (Tian et al., 2024) injects node and relation features; multilingual adapters (Gurgurov et al., 2024; Gurgurov et al.), KnowLA (Luo et al., 2024c), and domain-specific methods (Park et al., 2023; Meng et al., 2021) extend this strategy to various settings.

3.2 Knowledge Modification

This subsection focuses on KG-based knowledge modification (editing), which includes (1) enhancing cascading effects via knowledge subgraphs and (2) constructing KG-derived editing datasets.

(1) Enhancing Cascading Effects via Knowledge Subgraphs. Isolated fact edits often fail to generalize to related knowledge, whereas KGs encode dependencies that enable coherent cascading updates. GLAME (Zhang et al., 2024d) propagates edits over n-hop KG subgraphs via causal tracing, and HYPE (Atri et al., 2025) further promotes consistent propagation using hyperbolic subgraphs.

(2) Constructing KG-derived Editing Datasets. Existing knowledge editing benchmarks largely target commonsense reasoning and underrepresent domain-specific settings. KGs provide a principled basis for realistic benchmarks by defining edits over factual triples. MedCF (Xu et al., 2024) derives medical QA-based editing data from DRKG with counterfactual tail replacements, while VLKEB (Huang et al., 2024b) extends this paradigm to multimodal MMKGs. KGs also enable portability evaluation through one-hop QA tests that assess generalization to related facts.

3.3 Knowledge Unlearning

Privacy and copyright risks in LLMs motivate regulated unlearning, yet most methods ignore dependencies between forgotten and retained knowledge. Graphs explicitly encode such correlations by enabling (1) graph-based unlearning evaluation for more reliable assessment, and (2) graph-based forget-set construction for systemic forgetting.

(1) Graph-Based Unlearning Evaluation. Existing unlearning benchmarks test forgetting of target facts while preserving unrelated ones (Shi et al., 2024; Maini et al., 2024), but they ignore dependencies between facts. KGs offer a structured basis for more reliable evaluation. FaithUn (Yang et al., 2025) assesses robustness, multi-hop forgetting, and retention via KG-derived QA sets. PISTOL (Qiu et al., 2024a) builds relational graphs for multi-scenario evaluation. HANKER (Jiang et al., 2025b) removes overlapping triples and generates diverse queries to ensure faithful auditing.

(2) Graph-Based Forget-set Construction. Graph structures can also enhance forget-set construction. KGUnL (Mai et al., 2024) builds forget sets by converting harmful outputs into knowledge graphs and replacing unsafe triples with ethical ones, enabling interpretable unlearning that removes malicious content while reinforcing safe knowledge. Concept Unlearning (Yamashita et al., 2025) targets scenarios where only a concept-level forgetting request is given, constructing a concept-centric knowledge graph via LLM-generated triplets and converting it into textual forget samples for iterative unlearning until no new concept-related triplets remain.

4 Non-Parametric Knowledge Augmentation

Non-parametric methods store knowledge outside the parameters and retrieve it at inference time. Based on the source, they include *external* and *internal* knowledge augmentation. External knowledge, typically retrieved from resources such as Wikipedia or Wikidata, underpins *RAG*, where graphs index text corpora or store structured relational knowledge. Internal knowledge comprises long-context inputs, user preferences, and interaction experience, maintained as *memory*. In both cases, graphs act as indexing or memory structures. Accordingly, we organize this section into *Graph-Assisted RAG* and *Graph-Assisted LLM Memory*.

4.1 Graph-Assisted RAG

Graph-assisted RAG includes two roles of graphs: *graphs for knowledge indexing* that organize textual knowledge into retrievable graphs, and *graphs as knowledge stores* that supply explicit context.

(1) Graph for Knowledge Indexing. Traditional

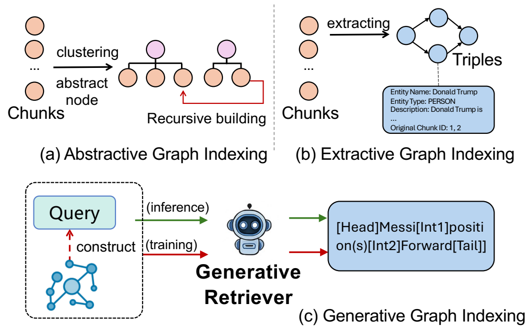

Figure 3: Graph-based indexing methods.

RAG indexes isolated text chunks, missing inter-document relations. Graph-based indexing structures documents into graphs, reducing redundancy and improving retrieval. Fig. 3 shows that existing methods follow three paradigms: (i) abstractive, (ii) extractive, and (iii) generative graph indexing.

(i) Abstractive Graph Indexing: Independent chunk indexing scatters related information and degrades global summarization; abstractive graph indexing clusters text into hierarchies, with lower-level nodes capturing details and higher-level nodes encoding abstractions (Fig. 3(a)). RAPTOR (Sarthi et al., 2024) constructs retrieval trees for multi-level retrieval, while PECAN (Wang et al., 2025h) and GraphRAG (Edge et al., 2024) extend this to hierarchical graphs for more flexible exploration.

(ii) Extractive Graph Indexing: Chunk-based retrieval ignores structural relations, often producing fragmented answers. Extractive graph indexing addresses this by extracting entities, relations, and events into graphs that preserve logical connections and links to source text (Fig. 3(b)). Methods such as GEAR (Shen et al., 2025), HippoRAG (Jimenez Gutierrez et al., 2024), LightRAG (Guo et al., 2024), MiniRAG (Fan et al., 2025), GFM-RAG (Luo et al., 2025c), and KAG (Liang et al., 2025) enhance multi-hop retrieval. Extensions include GoR (Zhang et al., 2024b), and HyperGraphRAG (Luo et al., 2025a).

(iii) Generative Graph Indexing: KG retrieval often relies on breadth-first expansion, yielding irrelevant triples, or agent-based search, which is slow. Generative graph indexing addresses these limitations by fine-tuning LLMs to map queries directly to relevant relation sets or induced subgraphs (Fig. 3(c)). GSR (Huang et al., 2024c) generates query-specific relation chains, DialogGSR (Park et al., 2024) extends this to dialogue, and DP (Ma et al., 2025b) injects the structural and constraint priors from KGs for faithful subgraph generation.

(2) Graph as Knowledge Store. Beyond graph

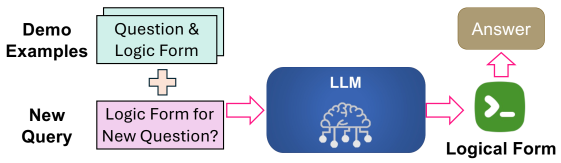

Figure 4: Semantic parsing-based methods.

indexing, RAG uses KGs as external structured knowledge for QA and reasoning. KG querying is challenged by the language–schema gap, schema heterogeneity, and incompleteness. Recent methods leverage LLMs' generative querying and internal knowledge to address these issues. Thus, KG-sourced RAG approaches are grouped into (i) semantic parsing–based, (ii) information retrieval-based, and (iii) agent-based multi-turn methods.

(i) Semantic Parsing–based Methods: Semantic parsing enables LLMs to access KGs by translating natural-language queries into executable forms (e.g., SPARQL), bridging unstructured inputs and graph schemas. Early approaches rely on KG-specific parsers with heavy supervision (Yih et al., 2015; He et al., 2021). Recent methods leverage LLM in-context learning for low-annotation query generation (Fig. 4), including KB-BINDER (Li et al., 2023a), and KB-Coder (Nie et al., 2024).

(ii) Information Retrieval–based Methods: IR-based KG methods retrieve relevant subgraphs or reasoning paths without explicit logical forms, alleviating limitations of semantic parsing under KG incompleteness (Fig. 5). They typically involve (i) query representation and entity matching, (ii) multi-hop subgraph or path retrieval, (iii) relevance scoring and reranking, and (iv) KG-grounded answer generation. Representative approaches include MedRAG (Zhao et al., 2025c), K-RagRec (Wang et al., 2025e), GNN-RAG (Mavromatis and Karypis, 2024), HyKGE (Jiang et al., 2025c), and constraint-based generation methods (Ma et al., 2025b; Luo et al., 2025b, 2024b).

(iii) Agent–Based Methods: Agent-based methods enable multi-turn LLM–KG interaction, overcoming rigid schemas in semantic parsing and one-shot retrieval limits under KG incompleteness (Fig. 6). Agents iteratively plan, query, verify, and update reasoning via atomic KG actions until termination. Representative systems include StructGPT (Jiang et al., 2023), Think-on-Graph (Sun et al., 2024), Interactive-KBQA (Xiong et al., 2024), CoK (Li et al., 2024), and KG-Agent (Jiang et al., 2025a).

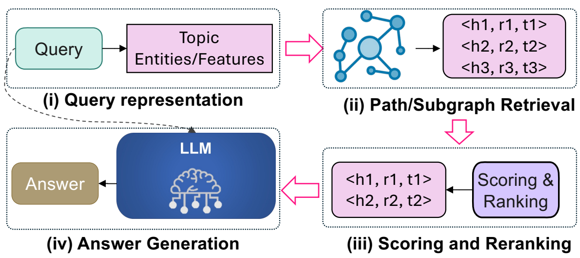

Figure 5: Information retrieval-based methods.

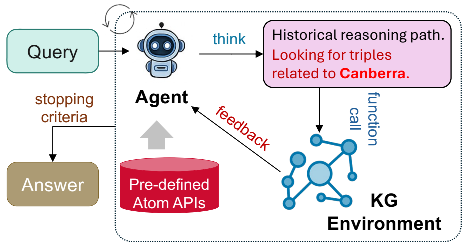

Figure 6: Agent-based methods.

4.2 Graph-Assisted Memory

Graphs organize internal memory either as structured indices for efficient retrieval or as explicit stores for dense, interrelated information. Accordingly, we distinguish *graph for memory indexing* and *graph as memory store*, as shown in Fig. 7.

(1) Graph for Memory Indexing. Graph-based memory indexing is categorized into *planar* and *hierarchical* graphs. Planar graphs directly connect raw interaction records and enrich representations via neighborhood aggregation but are less explored (e.g., A-mem (Xu et al., 2025)). Hierarchical graphs dominate, organizing memory into multiple abstraction levels for coarse-to-fine retrieval. Representative methods include CAM (Li et al., 2025a), Zep (Rasmussen et al., 2025), SG-Mem (Wu et al., 2025), G-Memory (Zhang et al., 2025b), and AriGraph (Anokhin et al., 2024).

(2) Graph as Memory Store. Graph-based memory stores encode memory as knowledge graphs, discarding raw input–output records in favor of structured, editable representations. By extract-

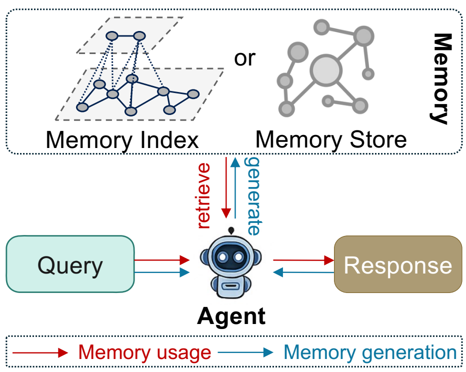

Figure 7: Memory graph for indexing or as stores.

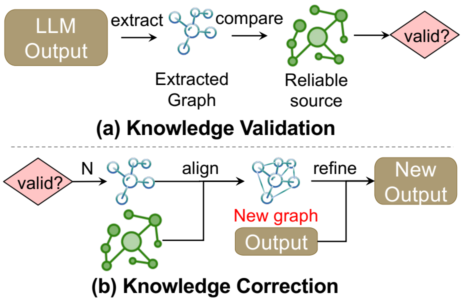

Figure 8: Knowledge validation and correction.

ing explicit relations from past interactions, they enable controllable, up-to-date retrieval and reasoning. For example, D-SMART (Lei et al., 2025) builds dialogue-specific KGs to maintain factual consistency, while EMG-RAG (Wang et al., 2024c) constructs user-centric, multi-level KGs from daily interactions to support personalized retrieval.

5 Knowledge Validation and Correction

Despite knowledge augmentation, LLMs still exhibit stochastic errors that are unacceptable in high-stakes settings post-inference; therefore, verified KGs enable efficient fact-checking with structured triples and correcting errors. Accordingly, this section focuses on *graph-assisted knowledge validation* and *graph-assisted knowledge correction*.

(1) Graph-Assisted Knowledge Validation. They verify factual consistency by comparing graph-structured representations of LLM outputs with trusted sources (Fig. 8(a)). Graphs play two roles. *(i) Graphs as External Knowledge:* Outputs are converted into triples (e.g., via OpenIE) and matched against verified KGs; CoKGLM (Hasegawa and Ichise, 2024) detects inconsistent relations, while domain-specific methods (Delmas et al., 2025) resolve terminology mismatches. *(ii) Graphs for Internal Validation:* Outputs are transformed into graphs and checked via graph reasoning, as Context-aware Hallucination Detection (Fang et al., 2025), GraphCheck (Chen et al., 2025b), and GENUINE (Wang et al., 2025g).

(2) Graph-Assisted Knowledge Correction. Once erroneous knowledge is detected, this process aligns LLM outputs with verified KG facts. The system constructs a corrected graph from retrieved triples and retrofits the response for consistency (Fig. 8(b)). Representative methods include KGR (Guan et al., 2024) and GraphCheck (Chen et al., 2025c), which revise answers using Wikidata, and CoG (Zhao et al., 2025b), which fixes entity, relation, and path errors to ensure faithfulness.

Discussion on Graph-Assisted LLM Knowledge

Graph-Assisted LLM knowledge augmentation acts across three stages: fine-tuning, inference, and post-inference.

- In the fine-tuning stage, parametric knowledge editing methods, ranging from full-parameter to adapter-based approaches, e.g., inject new or domain-specific KG knowledge into model parameters but still suffer from slow update cycles. Knowledge unlearning is a new area in data rights that requires deeper investigation.
- During the inference stage, external knowledge graphs or internal/textual knowledge organized as graphs provide more real-time information, yet LLMs may still hallucinate due to stochastic generation or incorrect beliefs.
- Post-inference stage employs external KGs for validation and correction, identifying inconsistencies and supplying verified facts to further refine model outputs.

6 Graph-Assisted Reasoning/Planning

LLMs often struggle with tasks requiring long-range dependencies in reasoning and planning due to implicit relationships in reasoning traces and planning steps. Graph-based methods address this by making dependencies explicit, enabling structured reasoning and planning. This section reviews *graph-structured reasoning* for logical thinking and *graph-assisted planning* for goal-oriented action sequences, summarized in Table 2.

6.1 Graph-Structured Reasoning

LLMs can perform multi-step reasoning when prompted, but their reasoning traces remain linear, limiting exploration of alternative paths. Graph structures address these limitations by introducing explicit graph-form organization into either the reasoning process or the input context. Current work falls into two categories: (1) graph-structured input and (2) graph-structured reasoning.

(1) Graph-Structured Input. LLMs possess strong contextual understanding but struggle with long text, where key entities and relations are dispersed. Graph-based input structuring addresses this by converting raw text into explicit relational graphs that highlight semantic, temporal, or causal dependencies, producing a compact and less noisy prompt (Fig. 9(a)). Representative methods include

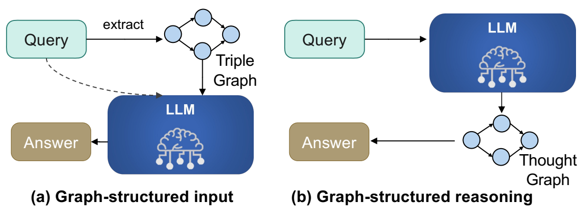

Figure 9: Graph-structured reasoning.

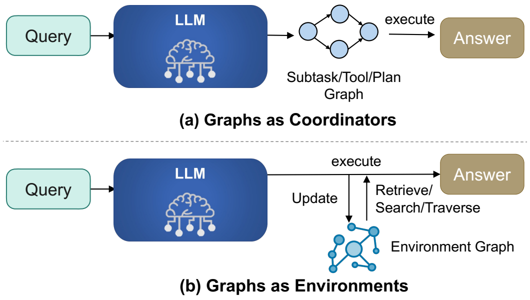

Figure 10: Graph-assisted planning.

Structure-Guided Prompting (Cheng et al., 2024), which builds concept maps to aid zero-shot reasoning; RwG (Han et al., 2025), which iteratively constructs and verifies reasoning graphs; and Talk-like-a-Graph (Fatemi et al., 2023), which assesses LLM reasoning over graph-encoded text.

(2) Graph-Structured Reasoning. LLMs can perform multi-step reasoning using Chain-of-Thought (CoT) prompting (Wei et al., 2022; Wang et al., 2023), but linear reasoning limits their ability to capture complex dependencies, explore alternatives, or recover from early errors. Graph-structured reasoning addresses this by organizing thoughts into relational structures where nodes represent states and edges encode dependencies (Fig. 9(b)), enabling multi-path exploration. Recent methods extend CoT into structured frameworks: ToT (Yao et al., 2023) performs tree search over partial solutions; Tree Prompting (Singh et al., 2023) learns static decision trees; GoT (Besta et al., 2024) generalizes to arbitrary directed graphs of logical dependencies; and SaGoT (Bai et al., 2025) infers thought dependencies via attention scores.

6.2 Graph-Assisted Planning

LLMs often struggle to plan long action sequences, manage inter-tool dependencies, and adapt to dynamic contexts. Graph structures address these challenges by providing explicit representations of task dependencies, tool relations, and action flows. We organize works into two approaches: (1) graphs as coordinators and (2) graphs as environments.

(1) Graphs as Coordinators. For multi-stage composite tasks, graphs unify decomposition, tool selection, and sequencing, addressing dependency modeling and hallucination (Fig. 10(a)). *(i) Sub-Task Pool as Graphs:* Agents model subtask dependencies by constructing task graphs for long-horizon planning (Shen et al., 2023; Lin et al., 2024) or using pre-defined pools via GNN-based retrieval for reliability (Wu et al., 2024b). *(ii) Tool Management as Graphs:* Tool graphs structure the tool space to model functional dependencies. Approaches like (Liu et al., 2024b,a; Ding et al., 2025) use traversal to find optimal paths in large libraries, while others (Wang et al., 2025j) employ graph-based sampling to generate tuning data for enhanced capability. *(iii) Plan as Graphs:* Execution is modeled as static workflows or dynamic, context-aware graphs (Zhang et al., 2024c; Wu et al., 2024c). Synthetic plan graphs enhance parallel planning via fine-tuning (Zhang et al., 2025i), a core paradigm in agent benchmarks (Qiao et al., 2024).

(2) Graphs as Environments. Beyond internal task coordination, agents frequently need to perceive and interact with external environments that are inherently structured and dynamic. Graph representations serve as a critical bridge, enabling agents to model intricate dependencies within codebases, memory systems, or physical spaces (Fig. 10(b)). For instance, LocAgent (Chen et al., 2025d) models codebases as dependency graphs for bug localization; Mem0g (Chhikara et al., 2025) organizes memory into dynamic KGs for context-aware retrieval; and Huang et al. (2025b) constructs spatio-semantic graphs for safety-aware planning.

Discussion on Graph-Assisted Reasoning and Planning

Graph-assisted reasoning and planning offer a principled solution to LLMs' structural limitations in complex tasks. By embedding explicit relational structures into input contexts or intermediate thoughts, graphs enable LLMs to externalize long-range dependencies, recover from early errors, explore multiple reasoning paths, and integrate dispersed information. In planning, graph-based representations of tasks, tools, or environments support multi-step coordination and adaptive decision-making. Remaining challenges include automatically constructing reliable graph representations in real-world settings and developing scalable evaluations for structured reasoning and planning.

7 Graph-Assisted LLM Collaboration

Single LLMs struggle to meet diverse task requirements, motivating collaborative mechanisms. Two dominant paradigms emerge: Mixture-of-Experts (MoE), which enables implicit division of labor within a single model via expert routing, and Multi-Agent Systems (MAS), where heterogeneous LLM agents interact. Both induce structured interactions that are naturally captured by graphs, which we analyze for organizing and optimizing collaboration; therefore, we focus on *(1) graph-assisted MoE* and *(2) graph-assisted MAS*, summarized in Table 3.

(1) Graph-Assisted Mixture-of-Experts. MoE architectures rely on routing mechanisms that select

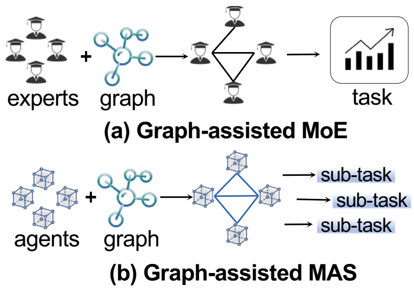

Figure 11: Graph-assisted LLM collaboration.

a subset of experts for each input. Traditional routing treats experts as independent units, often leading to redundant learning and weak specialization. Graph-assisted routing addresses this by modeling experts as nodes in a graph to enable relational reasoning and coordinated activation via message passing (Fig. 11(a)). This richer structure, however, introduces challenges such as load imbalance and communication overhead in large models. Recent methods explore solutions: GraphLoRA (Bai et al., 2024) and GraphRouter (Zhao et al., 2024) employ GNN-guided routing for balanced selection, while GraphMOE (Tang et al., 2025) enhances inter-expert knowledge transfer. Additionally, Lancet (Jiang et al., 2024a) optimizes system-level communication in graph-structured pipelines.

(2) Graph-Assisted Multi-Agent System. LLMs coordinate specialized agents by decomposing tasks and assigning them to appropriate units. These methods model the systems as graphs where nodes represent agents and edges encode communication or dependency constraints, enabling coordinated execution via optimization (Fig. 11(b)). *(i) For Performance:* Collaboration is formulated as graph optimization, including DAG-based frameworks (Dong et al., 2024; Park et al., 2025), topology learning (Zhang et al., 2024a; Li et al., 2025b), KG-driven (Yu and McQuade, 2025), search-based (Zhang et al., 2024c; Wei et al., 2025b), evolutionary (Zhang et al., 2025a; Hu et al., 2024), RL-based (Shinn et al., 2024; Wang et al., 2025b; Zhou et al., 2025), and hypergraphs (Zhao et al., 2025a; Zhang et al., 2025f,c). *(ii) For Efficiency:* Communication efficiency is improved via AgentPrune (Zhang et al., 2025d), AgentDropout (Wang et al., 2025l), and budget-aware AgentBalance (Cai et al., 2025). *(iii) For Robustness:* Robustness is enhanced by filtering failure-prone paths (Gao et al., 2025), cooperative evolution (Wei et al., 2025b), and GNN-based detection with topological intervention (Wang et al., 2025f).

Discussion on Graph-Assisted LLM Collaboration

Graph-assisted collaboration provides explicit structures for coordinating experts or agents, improving routing and task decomposition beyond traditional MoE or MAS systems. By modeling collaboration as a graph, systems can adapt communication patterns and specialize more effectively, efficiently, and robustly. Remaining challenges include scalability bottlenecks in large systems and the lack of standard benchmarks for graph-structured cooperation.

8 Applications

Graph-assisted LLMs address intrinsic limitations via structural modeling. In **biomedical and healthcare**, methods bridge domain gaps through precise knowledge injection (Wu et al., 2024a) and retrieval (Zhao et al., 2025c). In **scientific reasoning**, graph structures enable non-linear inference for math and biology (Bai et al., 2025; Sengupta et al., 2025). For **software engineering**, graphs model code dependencies to enhance bug localization (Chen et al., 2025d) and understanding (Zhang et al., 2025j). Finally, graphs enhance **trustworthiness** using external KGs to detect (Chen et al., 2025c) and correct (Zhao et al., 2025b) hallucinations. Detailed reviews are in Appendix A.

9 Challenges and Future Directions

While graph-assisted LLMs address key deficits, challenges remain. In knowledge maintenance, slow parametric updates necessitate lightweight editing. For reasoning, a major bottleneck is automated construction of reliable graphs from unstructured text, alongside a need for process-oriented evaluations beyond final accuracy. Finally, collaborative systems face scalability bottlenecks due to communication overhead, requiring topology-aware compression and standardized benchmarks. A detailed discussion is provided in Appendix B.

10 Conclusion

We present a systematic survey of graph-assisted LLMs focusing on mitigating intrinsic limitations. We propose a taxonomy spanning five paradigms: Parametric and Non-Parametric Knowledge Augmentation, Knowledge Validation and Correction, Reasoning and Planning, and Collaboration. These methods address three core deficits: alleviating knowledge cutoffs and hallucinations, overcoming linear reasoning constraints, and optimizing collaboration. By demonstrating how graphs complement LLMs, our work serves as a roadmap for developing trustworthy and intelligent systems.

Limitations

Although our work introduces a systematic framework for analyzing Graph-LLM integration, several limitations remain. First, given the rapid evolution of this field, new techniques emerge daily. While we strive for comprehensive coverage, certain recent preprints or niche domain applications may not be fully captured, and further community insights are encouraged to complement this study. Second, overlaps exist among the categorized limitations. For instance, the objective of enhancing domain knowledge via KGs naturally intersects with mitigating hallucination, as factual grounding serves both ends. While we categorize methods based on their primary motivation, many graph-based solutions effectively address multiple deficits simultaneously. We expect that addressing these complexities will motivate future comprehensive surveys and deeper investigations into the interpretability and scalability of Graph-LLM systems.

References

Ankush Agarwal, Sakharam Gawade, Amar Prakash Azad, and Pushpak Bhattacharyya. 2023. Kitlm: domain-specific knowledge integration into language models for question answering. *arXiv preprint arXiv:2308.03638*.

Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, Andrey Kravchenko, Mikhail Burtsev, and Evgeny Burnaev. 2024. Arigraph: Learning knowledge graph world models with episodic memory for llm agents. *arXiv preprint arXiv:2407.04363*.

Tu Ao, Yanhua Yu, Yuling Wang, Yang Deng, Zirui Guo, Liang Pang, Pinghui Wang, Tat-Seng Chua, Xiao Zhang, and Zhen Cai. 2025. Lightprof: A lightweight reasoning framework for large language model on knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23424–23432.

Yash Kumar Atri, Ahmed Alaa, and Thomas Hartvigsen. 2025. Lifelong model editing with graph-based external memory. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13336–13352, Vienna, Austria. Association for Computational Linguistics.

Ruiqiao Bai, Xue Han, Shuo Lei, Junlan Feng, Yanyan Luo, and Chao Deng. 2025. Self-attention-based graph-of-thought for math problem solving. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6112–6125, Vienna, Austria. Association for Computational Linguistics.

Ting Bai, Yue Yu, Le Huang, Zenan Xu, Zhe Zhao, and Chuan Shi. 2024. Graphlora: Empowering llms fine-tuning via graph collaboration of moe. *arXiv preprint arXiv:2412.16216*.

Yuanchen Bei, Weizhi Zhang, Siwen Wang, Weizhi Chen, Sheng Zhou, Hao Chen, Yong Li, Jiajun Bu, Shirui Pan, Yizhou Yu, et al. 2025. Graphs meet ai agents: Taxonomy, progress, and future opportunities. *arXiv preprint arXiv:2506.18019*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Shuowei Cai, Yansong Ning, and Hao Liu. 2025. Agentbalance: Backbone-then-topology design for cost-effective multi-agent systems under budget constraints. *arXiv preprint arXiv:2512.11426*.

Hanzhu Chen, Xu Shen, Jie Wang, Zehao Wang, Qitan Lv, Junjie He, Rong Wu, Feng Wu, and Jieping Ye. 2025a. Knowledge graph finetuning enhances knowledge manipulation in large language models. In *The Thirteenth International Conference on Learning Representations*.

Liyi Chen, Panrong Tong, Zhongming Jin, Ying Sun, Jieping Ye, and Hui Xiong. 2024. Plan-on-graph: Self-correcting adaptive planning of large language model on knowledge graphs. *Advances in Neural Information Processing Systems*, 37:37665–37691.

Yingjian Chen, Haoran Liu, Yinhong Liu, Jinxiang Xie, Rui Yang, Han Yuan, Yanran Fu, Peng Yuan Zhou, Qingyu Chen, James Caverlee, and Irene Li. 2025b. GraphCheck: Breaking long-term text barriers with extracted knowledge graph-powered fact-checking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14976–14995, Vienna, Austria. Association for Computational Linguistics.

Yingjian Chen, Haoran Liu, Yinhong Liu, Jinxiang Xie, Rui Yang, Han Yuan, Yanran Fu, Peng Yuan Zhou, Qingyu Chen, James Caverlee, et al. 2025c. Graphcheck: Breaking long-term text barriers with extracted knowledge graph-powered fact-checking. *arXiv preprint arXiv:2502.16514*.

Zhaoling Chen, Robert Tang, Gangda Deng, Fang Wu, Jialong Wu, Zhiwei Jiang, Viktor Prasanna, Arman Cohan, and Xingyao Wang. 2025d. LocAgent: Graph-guided LLM agents for code localization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8697–8727, Vienna, Austria. Association for Computational Linguistics.

Kewei Cheng, Nesreen Ahmed, Theodore Willke, and Yizhou Sun. 2024. Structure guided prompt: Instructing large language model in multi-step reasoning by exploring graph structure of the text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9407–9430.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

Maxime Delmas, Magdalena Wysocka, Danilo Miranda Gusicuma, and André Freitas. 2025. Accelerating antibiotic discovery with large language models and knowledge graphs. *CoRR*, abs/2503.16655.

Yihe Deng, Chenchen Ye, Zijie Huang, Mingyu Derek Ma, Yiwen Kou, and Wei Wang. 2024. Graphvis: Boosting LLMs with visual knowledge graph integration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Stefan Dernbach, Khushbu Agarwal, Alejandro Zuniga, Michael Henry, and Sutanay Choudhury. 2024. Glam: Fine-tuning large language models for domain knowledge graph alignment via neighborhood partitioning and generative subgraph encoding. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 82–89.

Shizhe Diao, Tianyang Xu, Ruijia Xu, Jiawei Wang, and Tong Zhang. 2023. Mixture-of-domain-adapters: Decoupling and injecting domain knowledge to pre-trained language models' memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5113–5129.

Keyan Ding, Jing Yu, Junjie Huang, Yuchen Yang, Qiang Zhang, and Huajun Chen. 2025. Scitoolagent: a knowledge-graph-driven scientific agent for multitool integration. *Nature Computational Science*, 5(10):962–972.

Yubo Dong, Xukun Zhu, Zhengzhe Pan, Linchao Zhu, and Yi Yang. 2024. Villageragent: A graph-based multi-agent framework for coordinating complex task dependencies in minecraft. *arXiv preprint arXiv:2406.05720*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Tianyu Fan, Jingyuan Wang, Xubin Ren, and Chao Huang. 2025. Minirag: Towards extremely simple retrieval-augmented generation. *arXiv preprint arXiv:2501.06713*.

Xinyue Fang, Zhen Huang, Zhiliang Tian, Minghui Fang, Ziyi Pan, Quntian Fang, Zhihua Wen, Hengyue Pan, and Dongsheng Li. 2025. Zero-resource hallucination detection for text generation via graph-based contextual knowledge triples modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23868–23877.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2023. Talk like a graph: Encoding graphs for large language models. *arXiv preprint arXiv:2310.04560*.

Hongcheng Gao, Yue Liu, Yufei He, Longxu Dou, Chao Du, Zhijie Deng, Bryan Hooi, Min Lin, and Tianyu Pang. 2025. Flowreasoner: Reinforcing query-level meta-agents. *arXiv preprint arXiv:2504.15257*.

GeoNames. 2004. Geonames ontology. `http://www.geonames.org/ontology/`.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18126–18134.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.

Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. 2024. Adapting multilingual llms to low-resource languages with knowledge graphs via adapters. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 63–74.

Daniil Gurgurov, Ivan Vykopal, Josef van Genabith, and Simon Ostermann. Small models, big impact: Efficient corpus and graph-based adaptation of small multilingual language models for low-resource languages. In *ACL 2025 Student Research Workshop*.

Haoyu Han, Yaochen Xie, Hui Liu, Xianfeng Tang, Sreyashi Nag, William Headden, Yang Li, Chen Luo, Shuiwang Ji, Qi He, and Jiliang Tang. 2025. Reasoning with graphs: Structuring implicit knowledge to enhance LLMs reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25698–25714, Vienna, Austria. Association for Computational Linguistics.

Rie Hasegawa and Ryutaro Ichise. 2024. Cokglm: Detecting hallucinations generated by large language models via knowledge graph verification. In *Knowledge Graphs and Semantic Web - 6th International Conference, KGSWC 2024, Paris, France, December 11-13, 2024, Proceedings*, volume 15459 of *Lecture Notes in Computer Science*, pages 212–224. Springer.

Gaole He, Yunshi Lan, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Improving multi-hop knowledge base question answering by learning intermediate supervision signals. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 553–561.

Mengliang He, Aimin Zhou, and Xiaoming Shi. 2024. Enhancing textbook question answering with knowledge graph-augmented large language models. In *The 16th Asian Conference on Machine Learning (Conference Track)*.

Yue Hu, Yuzhu Cai, Yaxin Du, Xinyu Zhu, Xiangrui Liu, Zijie Yu, Yuchen Hou, Shuo Tang, and Siheng Chen. 2024. Self-evolving multi-agent collaboration networks for software development. *arXiv preprint arXiv:2410.16946*.

Chao Huang, Xubin Ren, Jiabin Tang, Dawei Yin, and Nitesh Chawla. 2024a. Large language models for graphs: Progresses and directions. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1284–1287.

Han Huang, Haitian Zhong, Tao Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2024b. Vlkeb: A large vision-language model knowledge editing benchmark.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Wanjing Huang, Tongjie Pan, and Yalan Ye. 2025b. Graphormer-guided task planning: Beyond static rules with llm safety perception. *arXiv preprint arXiv:2503.06866*.

Wenyu Huang, Guancheng Zhou, Hongru Wang, Pavlos Vougiouklis, Mirella Lapata, and Jeff Pan. 2024c. Less is more: Making smaller language models competent subgraph retrievers for multi-hop kgqa. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15787–15803.

Chenyu Jiang, Ye Tian, Zhen Jia, Shuai Zheng, Chuan Wu, and Yida Wang. 2024a. Lancet: Accelerating mixture-of-experts training via whole graph computation-communication overlapping. *Proceedings of Machine Learning and Systems*, 6:74–86.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251, Singapore. Association for Computational Linguistics.

Jinhao Jiang, Kun Zhou, Xin Zhao, Yang Song, Chen Zhu, Hengshu Zhu, and Ji-Rong Wen. 2025a. KG-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9505–9523, Vienna, Austria. Association for Computational Linguistics.

Weipeng Jiang, Juan Zhai, Shiqing Ma, Ziyan Lei, Xiaofei Xie, Yige Wang, and Chao Shen. 2025b. Holistic audit dataset generation for llm unlearning via knowledge graph traversal and redundancy removal. *arXiv preprint arXiv:2502.18810*.

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, et al. 2025c. Hykge: A hypothesis knowledge graph enhanced rag framework for accurate and reliable medical llms responses. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11856.

Zhouyu Jiang, Ling Zhong, Mengshu Sun, Jun Xu, Rui Sun, Hui Cai, Shuhan Luo, and Zhiqiang Zhang. 2024b. Efficient knowledge infusion via kg-llm alignment. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2986–2999.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37:59532–59569.

Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. 2024. Large language models on graphs: A comprehensive survey. *IEEE Transactions on Knowledge and Data Engineering*.

Xiang Lei, Qin Li, and Min Zhang. 2025. D-smart: Enhancing llm dialogue consistency via dynamic structured memory and reasoning tree. *arXiv preprint arXiv:2510.13363*.

Rui Li, Zeyu Zhang, Xiaohe Bo, Zihang Tian, Xu Chen, Quanyu Dai, Zhenhua Dong, and Ruiming Tang. 2025a. CAM: A constructivist view of agentic memory for LLM-based reading comprehension. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Shiyuan Li, Yixin Liu, Qingsong Wen, Chengqi Zhang, and Shirui Pan. 2025b. Assemble your crew: Automatic multi-agent communication topology design via autoregressive graph generation. *arXiv preprint arXiv:2507.18224*.

Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. 2023a. Few-shot in-context learning on knowledge base question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6966–6980, Toronto, Canada. Association for Computational Linguistics.

Xijun Li, Jiexiang Yang, Jinghao Wang, Bo Peng, Jianguo Yao, and Haibing Guan. 2025c. Strcmp: Integrating graph structural priors with language models for combinatorial optimization. *arXiv preprint arXiv:2506.11057*.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing.

2024. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*.

Yuhan Li, Zhixun Li, Peisong Wang, Jia Li, Xiangguo Sun, Hong Cheng, and Jeffrey Xu Yu. 2023b. A survey of graph meets large language model: Progress and future directions. *arXiv preprint arXiv:2311.12399*.

Lei Liang, Zhongpu Bo, Zhengke Gui, Zhongshu Zhu, Ling Zhong, Peilong Zhao, Mengshu Sun, Zhiqiang Zhang, Jun Zhou, Wenguang Chen, et al. 2025. Kag: Boosting llms in professional domains via knowledge augmented generation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 334–343.

Fangru Lin, Emanuele La Malfa, Valentin Hofmann, Elle Michelle Yang, Anthony G Cohn, and Janet B Pierrehumbert. 2024. Graph-enhanced large language models in asynchronous plan reasoning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 30108–30134.

Qika Lin, Tianzhe Zhao, Kai He, Zhen Peng, Fangzhi Xu, Ling Huang, Jingying Ma, and Mengling Feng. 2025. Self-supervised quantized representation for seamlessly integrating knowledge graphs with large language models. *arXiv preprint arXiv:2501.18119*.

Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S Yu, et al. 2023. Towards graph foundation models: A survey and beyond. *arXiv preprint arXiv:2310.11829*.

Xukun Liu, Zhiyuan Peng, Xiaoyuan Yi, Xing Xie, Lirong Xiang, Yuchen Liu, and Dongkuan Xu. 2024a. Toolnet: Connecting large language models with massive tools via tool graph. *arXiv preprint arXiv:2403.00839*.

Yixin Liu, Guibin Zhang, Kun Wang, Shiyuan Li, and Shirui Pan. 2025. Graph-augmented large language model agents: Current progress and future prospects. *arXiv preprint arXiv:2507.21407*.

Zhaoyang Liu, Zeqiang Lai, Zhangwei Gao, Erfei Cui, Ziheng Li, Xizhou Zhu, Lewei Lu, Qifeng Chen, Yu Qiao, Jifeng Dai, et al. 2024b. Controlllm: Augment language models with tools by searching on graphs. In *European Conference on Computer Vision*, pages 89–105. Springer.

Haitong Luo, Xuying Meng, Suhang Wang, Tianxiang Zhao, Fali Wang, Hanyun Cao, and Yujun Zhang. 2024a. Enhance graph alignment for large language models. *arXiv preprint arXiv:2410.11370*.

Haoran Luo, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, et al. 2025a. Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation. *arXiv preprint arXiv:2503.21322*.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2024b. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth International Conference on Learning Representations*.

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Yuan-Fang Li, Chen Gong, and Shirui Pan. 2025b. Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models. In *Forty-second International Conference on Machine Learning*.

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Dinh Phung, Chen Gong, and Shirui Pan. 2025c. Gfm-rag: graph foundation model for retrieval augmented generation. *arXiv preprint arXiv:2502.01113*.

Xindi Luo, Zequn Sun, Jing Zhao, Zhe Zhao, and Wei Hu. 2024c. Knowla: Enhancing parameter-efficient finetuning with knowledgeable adaptation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7146–7159.

Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. 2025a. Large language models meet knowledge graphs for question answering: Synthesis and opportunities. *arXiv preprint arXiv:2505.20099*.

Jie Ma, Ning Qu, Zhitao Gao, Rui Xing, Jun Liu, Hongbin Pei, Jiang Xie, Linyun Song, Pinghui Wang, Jing Tao, et al. 2025b. Deliberation on priors: Trustworthy reasoning of large language models on knowledge graphs. *arXiv preprint arXiv:2505.15210*.

Peihua Mai, Hao Jiang, Ran Yan, Youjia Yang, Zhe Huang, and Yan Pang. 2024. Knowledge graph unlearning to defend language model against jailbreak attack. In *The Second Tiny Papers Track at ICLR 2024*.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Costas Mavromatis and George Karypis. 2024. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*.

Zaiqiao Meng, Fangyu Liu, Thomas Clark, Ehsan Shareghi, and Nigel Collier. 2021. Mixture-of-partitions: Infusing large biomedical knowledge graphs into bert. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4672–4681.

Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. 2022. Skill: Structured knowledge infusion for large language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1581–1588.

Zhijie Nie, Richong Zhang, Zhongyuan Wang, and Xudong Liu. 2024. Code-style in-context learning for knowledge-based question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18833–18841.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Chiwan Park, Wonjun Jang, Daeryong Kim, Aelim Ahn, Kichang Yang, Woosung Hwang, Jihyeon Roh, Hyerin Park, Hyosun Wang, Min Seok Kim, and Jihoon Kang. 2025. A practical approach for building production-grade conversational agents with workflow graphs. *CoRR*, abs/2505.23006.

Hyeryun Park, Jiye Son, Jeongwon Min, and Jinwook Choi. 2023. Selective umls knowledge infusion for biomedical question answering. *Scientific Reports*, 13(1):14214.

Jinyoung Park, Minseok Joo, Joo-Kyung Kim, and Hyunwoo Kim. 2024. Generative subgraph retrieval for knowledge graph–grounded dialog generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21167–21182.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*.

Tyler Thomas Procko and Omar Ochoa. 2024. Graph retrieval-augmented generation for large language models: A survey. In *2024 Conference on AI, Science, Engineering, and Technology (AIxSET)*, pages 166–169. IEEE.

Shuofei Qiao, Runnan Fang, Zhisong Qiu, Xiaobin Wang, Ningyu Zhang, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024. Benchmarking agentic workflow generation. *arXiv preprint arXiv:2410.07869*.

Xinchi Qiu, William F Shen, Yihong Chen, Nicola Cancedda, Pontus Stenetorp, and Nicholas D Lane. 2024a. Pistol: Dataset compilation pipeline for structural unlearning of llms. *arXiv preprint arXiv:2406.16810*, pages 2021–2025.

Zhangchi Qiu, Linhao Luo, Shirui Pan, and Alan Wee-Chung Liew. 2024b. Unveiling user preferences: A knowledge graph and llm-driven approach for conversational recommendation. *arXiv preprint arXiv:2411.14459*.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*.

Xubin Ren, Jiabin Tang, Dawei Yin, Nitesh Chawla, and Chao Huang. 2024. A survey of large language models for graphs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6616–6626.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations (ICLR)*.

Saptarshi Sengupta, Shuhua Yang, Paul Kwong Yu, Fali Wang, and Suhang Wang. 2025. Biomol-mqa: A multi-modal question answering dataset for llm reasoning over bio-molecular interactions. *arXiv preprint arXiv:2506.05766*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180.

Zhili Shen, Chenxin Diao, Pavlos Vougiouklis, Pascual Merita, Shriram Piramanayagam, Enting Chen, Damien Graux, Andre Melo, Ruofei Lai, Zeren Jiang, Zhongyang Li, Ye Qi, Yang Ren, Dandan Tu, and Jeff Z. Pan. 2025. GeAR: Graph-enhanced agent for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12049–12072, Vienna, Austria. Association for Computational Linguistics.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2024. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*.

N. Shinn et al. 2024. Gptswarm: Llm agents as (optimizable) graphs. *arXiv preprint arXiv:2403.07656*.

Chandan Singh, John Morris, Alexander M Rush, Jianfeng Gao, and Yuntian Deng. 2023. Tree prompting: Efficient task adaptation without fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6253–6267.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations*.

Chen Tang, Bo Lv, Zifan Zheng, Bohao Yang, Kun Zhao, Ning Liao, Xiaoxing Wang, Feiyu Xiong, Zhiyu Li, Nayu Liu, et al. 2025. Graphmoe: Amplifying cognitive depth of mixture-of-experts network via introducing self-rethinking mechanism. *arXiv preprint arXiv:2501.07890*.

Shiyu Tian, Yangyang Luo, Tianze Xu, Caixia Yuan, Huixing Jiang, Chen Wei, and Xiaojie Wang. 2024. Kg-adapter: Enabling knowledge graph integration in large language models through parameter-efficient fine-tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3813–3828.

Fali Wang, Runxue Bao, Suhang Wang, Wenchao Yu, Yanchi Liu, Wei Cheng, and Haifeng Chen. 2024a. Infuserki: Enhancing large language models with knowledge graphs via infuser-guided knowledge integration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3675–3688.

Fali Wang, Jihai Chen, Shuhua Yang, Ali Al-Lawati, Linli Tang, Hui Liu, and Suhang Wang. 2025a. A survey on collaborating small and large language models for performance, cost-effectiveness, cloud-edge privacy, and trustworthiness. *arXiv preprint arXiv:2510.13890*.

Fali Wang, Jihai Chen, Shuhua Yang, Runxue Bao, Tianxiang Zhao, Zhiwei Zhang, Xianfeng Tang, Hui Liu, Qi He, and Suhang Wang. 2025b. Generalizing test-time compute-optimal scaling as an optimizable graph. *arXiv preprint arXiv:2511.00086*.

Fali Wang, Hui Liu, Zhenwei DAI, Jingying Zeng, Zhiwei Zhang, Zongyu Wu, Chen Luo, Zhen Li, Xianfeng Tang, Qi He, and Suhang Wang. 2025c. AgentTTS: Large language model agent for test-time compute-optimal scaling strategy in complex tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Nengbo Wang, Xiaotian Han, Jagdip Singh, Jing Ma, and Vipin Chaudhary. 2025d. Causalrag: Integrating causal graphs into retrieval-augmented generation. *arXiv preprint arXiv:2503.19878*.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418.

Shijie Wang, Wenqi Fan, Yue Feng, Lin Shanru, Xinyu Ma, Shuaiqiang Wang, and Dawei Yin. 2025e. Knowledge graph retrieval-augmented generation for LLM-based recommendation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27152–27168, Vienna, Austria. Association for Computational Linguistics.

Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. 2025f. G-safeguard: A topology-guided security lens and treatment on llm-based multi-agent systems. *arXiv preprint arXiv:2502.11127*.

Tuo Wang, Adithya Kulkarni, Tyler Cody, Peter A Beling, Yujun Yan, and Dawei Zhou. 2025g. Genuine: Graph enhanced multi-level uncertainty estimation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20522–20541.

Xinyu Wang, Yanzheng Xiang, Lin Gui, and Yulan He. 2025h. Pecan: Llm-guided dynamic progress control with attention-guided hierarchical weighted graph for long-document qa. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13317–13335.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024b. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19206–19214.

Zehong Wang, Zheyuan Liu, Tianyi Ma, Jiazheng Li, Zheyuan Zhang, Xingbo Fu, Yiyang Li, Zhengqing Yuan, Wei Song, Yijun Ma, et al. 2025i. Graph foundation models: A comprehensive survey. *arXiv preprint arXiv:2505.15116*.

Zezhong Wang, Xingshan Zeng, Weiwen Liu, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2025j. Toolflow: Boosting llm tool-calling through natural and coherent dialogue synthesis. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4246–4263.

Zhefan Wang, Huanjun Kong, Jie Ying, Wanli Ouyang, and Nanqing Dong. 2025k. Rograg: A robustly optimized graphrag framework. *arXiv preprint arXiv:2503.06474*.

Zheng Wang, Zhongyang Li, Zeren Jiang, Dandan Tu, and Wei Shi. 2024c. Crafting personalized agents through retrieval-augmented generation on editable memory graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4891–4906.

Zhexuan Wang, Yutong Wang, Xuebo Liu, Liang Ding, Miao Zhang, Jie Liu, and Min Zhang. 2025l. Agentdropout: Dynamic agent elimination for token-efficient and high-performance llm-based multi-agent collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24013–24035. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shaopeng Wei, Jun Wang, Yu Zhao, Xingyan Chen, Xiaochun Hu, Qing Li, Fuzhen Zhuang, Fuji Ren, and Gang Kou. 2025a. Graph learning and its advancements on large language models: A holistic survey. *Neurocomputing*, page 132396.

Yangbo Wei, Zhen Huang, Huang Li, Wei W Xing, Ting-Jung Lin, and Lei He. 2025b. Vflow: Discovering optimal agentic workflows for verilog generation. *arXiv preprint arXiv:2504.03723*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024a. Pmc-llama: toward building open-source language models for medicine. *J. Am. Medical Informatics Assoc.*

Xixi Wu, Yifei Shen, Caihua Shan, Kaitao Song, Siwei Wang, Bohang Zhang, Jiarui Feng, Hong Cheng, Wei Chen, Yun Xiong, et al. 2024b. Can graph learning improve planning in llm-based agents? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yaxiong Wu, Yongyue Zhang, Sheng Liang, and Yong Liu. 2025. Sgmem: Sentence graph memory for long-term conversational agents. *arXiv preprint arXiv:2509.21212*.

Yue Wu, Yewen Fan, So Yeon Min, Shrimai Prabhumoye, Stephen McAleer, Yonatan Bisk, Ruslan Salakhutdinov, Yuanzhi Li, and Tom Mitchell. 2024c. Agentkit: structured llm reasoning with dynamic graphs. *arXiv preprint arXiv:2404.11483*.

Guanming Xiong, Junwei Bao, and Wen Zhao. 2024. Interactive-KBQA: Multi-turn interactions for knowledge base question answering with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10561–10582, Bangkok, Thailand. Association for Computational Linguistics.

Derong Xu, Ziheng Zhang, Zhihong Zhu, Zhenxi Lin, Qidong Liu, Xian Wu, Tong Xu, Wanyu Wang, Yuyang Ye, Xiangyu Zhao, et al. 2024. Editing factual knowledge and explanatory ability of medical large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2660–2670.

Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See Kiong Ng, and Jiashi Feng. 2023. Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. *arXiv preprint arXiv:2311.08562*.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

Tomoya Yamashita, Yuuki Yamanaka, Masanori Yamada, Takayuki Miura, Toshiki Shibahara, and Tomoharu Iwata. 2025. Concept unlearning in large language models via self-constructed knowledge triplets. *arXiv preprint arXiv:2509.15621*.

Jian Yang, Xinyu Hu, Gang Xiao, and Yulong Shen. 2024a. A survey of knowledge enhanced pre-trained language models. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Nakyeong Yang, Minsung Kim, Seunghyun Yoon, Joongbo Shin, and Kyomin Jung. 2025. Faithun: Toward faithful forgetting in language models by investigating the interconnectedness of knowledge. *arXiv preprint arXiv:2502.19207*.

Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yuhe Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, and Irene Li. 2024b. KG-rank: Enhancing large language models for medical QA with knowledge graphs and ranking techniques. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 155–166, Bangkok, Thailand. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1321–1331, Beijing, China. Association for Computational Linguistics.

Hong Qing Yu and Frank McQuade. 2025. Rag-kg-il: A multi-agent hybrid framework for reducing hallucinations and enhancing llm reasoning through rag and incremental knowledge graph learning integration. *arXiv preprint arXiv:2503.13514*.

Guibin Zhang, Kaijie Chen, Guancheng Wan, Heng Chang, Hong Cheng, Kun Wang, Shuyue Hu, and Lei Bai. 2025a. Evoflow: Evolving diverse agentic workflows on the fly. *arXiv preprint arXiv:2502.07373*.

Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. 2025b. G-memory: Tracing hierarchical memory for multi-agent systems. *arXiv preprint arXiv:2506.07398*.

Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. 2025c. Multi-agent architecture search via agentic supernet. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. 2025d. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. In *The Thirteenth International Conference on Learning Representations,* ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net.

Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. 2024a. G-designer: Architecting multi-agent communication topologies via graph neural networks. *arXiv preprint arXiv:2410.11782*.

Haozhen Zhang, Tao Feng, and Jiaxuan You. 2024b. Graph of records: Boosting retrieval augmented generation for long-context summarization with graphs. *arXiv preprint arXiv:2410.11001*.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. 2024c. Aflow: Automating agentic workflow generation. *arXiv preprint arXiv:2410.10762*.

Liangliang Zhang, Zhuorui Jiang, Hongliang Chi, Haoyang Chen, Mohammed ElKoumy, Fali Wang, Qiong Wu, Zhengyi Zhou, Shirui Pan, Suhang Wang, and Yao Ma. 2025e. Diagnosing and addressing pitfalls in KG-RAG datasets: Toward more reliable benchmarking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Mengqi Zhang, Xiaotian Ye, Qiang Liu, Pengjie Ren, Shu Wu, and Zhumin Chen. 2024d. Knowledge graph enhanced large language model editing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22647–22662.

Qianqian Zhang, Jiajia Liao, Heting Ying, Yibo Ma, Haozhan Shen, Jingcheng Li, Peng Liu, Lu Zhang, Chunxin Fang, Kyusong Lee, et al. 2025f. Unifying language agent algorithms with graph-based orchestration engine for reproducible agent research. *arXiv preprint arXiv:2505.24354*.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, et al. 2025g. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and Qingyun Wu. 2025h. Which agent causes task failures and when? on automated failure attribution of LLM multi-agent systems. In *Forty-second International Conference on Machine Learning*.

Shiqi Zhang, Xinbei Ma, Zouying Cao, Zhuosheng Zhang, and Hai Zhao. 2025i. Plan-over-graph: Towards parallelable llm agent schedule. *arXiv preprint arXiv:2502.14563*.

Ziyin Zhang, Hang Yu, Sage Lee, Peng Di, Jianguo Li, and Rui Wang. 2025j. GALLa: Graph aligned large language models for improved source code understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13784–13802, Vienna, Austria. Association for Computational Linguistics.

Chenxu Zhao et al. 2024. Graphrouter: A graph-based router for llm selections. *arXiv preprint arXiv:2403.03019*.

Rui Zhao et al. 2025a. Hypergraph coordination networks with dynamic grouping for multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*.

Ruilin Zhao, Feng Zhao, and Hong Zhang. 2025b. Correcting on graph: Faithful semantic parsing over knowledge graphs with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5364–5376.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025c. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pages 4442–4457.

Heng Zhou, Hejia Geng, Xiangyuan Xue, Li Kang, Yiran Qin, Zhiyong Wang, Zhenfei Yin, and Lei Bai. 2025. Reso: A reward-driven self-organizing llm-based multi-agent system for reasoning tasks. *arXiv preprint arXiv:2503.02390*.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8912–8924.

A Applications

Graph-assisted Large Language Models have demonstrated significant potential across domains by explicitly modeling structural dependencies to mitigate intrinsic limitations, such as knowledge cutoffs, reasoning gaps, and hallucinations. We summarize representative works in Figure 12 and detail their applications below.

Biomedical and Healthcare. As a quintessential knowledge-intensive domain, healthcare demands high precision and the integration of specialized, up-to-date information. Graph-assisted methods effectively bridge the domain gap in this context by enhancing knowledge injection and retrieval. For instance, PMC-LLaMA (Wu et al., 2024a) and MedCF (Xu et al., 2024) demonstrate how graph-based injection and editing ensure precise medical adaptation. Similarly, in retrieval tasks, systems like MedRAG (Zhao et al., 2025c) and

HyKGE (Jiang et al., 2025c) use medical graphs to ground generation, improving the accuracy of diagnostic and discovery tasks (Delmas et al., 2025).

Software Engineering. Code data inherently possesses a graph structure (e.g., Abstract Syntax Trees, dependency graphs), which is crucial for understanding complex logic. LocAgent (Chen et al., 2025d) utilizes these structures to model dependencies between code files, enabling autonomous agents to accurately locate bugs. KB-Coder (Nie et al., 2024) enhances knowledge-based question answering via code-style in-context learning, while GALLa (Zhang et al., 2025j) improves source code understanding by aligning structural code representations with textual models.

Scientific Reasoning. Scientific domains demand rigorous logic and the exploration of non-linear solution spaces. In these contexts, graph structures enable models to explicitly plan and navigate complex reasoning paths. Works such as SaGoT (Bai et al., 2025) in mathematics and Biomol-mqa (Sengupta et al., 2025) in biology exemplify this capability, demonstrating how graph-based inputs facilitate structural inference over mathematical proofs or bio-molecular interactions.

Trustworthiness and Factuality. Combating hallucinations is a critical application area for deploying trustworthy AI services. Graph-assisted systems provide a verifiable reference for validation, serving as a safeguard for reliable information dissemination. Tools like GraphCheck (Chen et al., 2025c) and CoKGLM (Hasegawa and Ichise, 2024) utilize external KGs to cross-reference and verify generated content. Furthermore, frameworks such as KGR (Guan et al., 2024) and CoG (Zhao et al., 2025b) extend this utility by actively correcting erroneous outputs based on trusted graph evidence.

B Challenges and Future Directions

While graph-assisted LLMs have addressed key deficits, several critical challenges remain. Based on the discussions above, we highlight primary directions for future research.

Efficient Knowledge Maintenance and Safety. Current parametric knowledge augmentation faces a significant bottleneck in update efficiency: injection methods often suffer from slow update cycles, making them ill-suited for rapidly evolving information. Future research must explore lightweight, modular editing techniques that allow for rapid knowledge integration without the cost of full parameter updates. Furthermore, Knowledge unlearning represents a nascent but critical area for data rights and safety. Deeper investigation is required to develop rigorous frameworks that can selectively erase sensitive or outdated information from graph-enhanced models while preserving the integrity of retained knowledge.

Automated Graph Construction and Process Evaluation. Graph-assisted reasoning and planning offer a principled solution to the structural limitations of LLMs by enabling them to externalize long-range dependencies and recover from early errors. However, a major bottleneck is automatically constructing reliable graph representations from unstructured contexts without relying on expensive human annotation or pre-defined schemas. Furthermore, current benchmarks largely focus on final answer accuracy. The field urgently needs scalable evaluations that assess the quality of the structured reasoning process itself, measuring the logical validity of graph traversal paths and the coherence of intermediate planning steps.

Scalability and Standardization in Collaboration. For graph-assisted collaboration, scalability bottlenecks pose a serious challenge. As the number of experts (in MoE) or agents (in MAS) grows, the communication overhead inherent in complex graph topologies can degrade system efficiency. Research is needed into topology-aware compression and dynamic pruning strategies that enable large-scale coordination without exhausting resources. Finally, there is a distinct lack of standard benchmarks for graph-structured cooperation. Establishing unified evaluation platforms that rigorously test coordination efficiency, robustness, and role specialization is essential for measuring progress in collaborative intelligent systems.

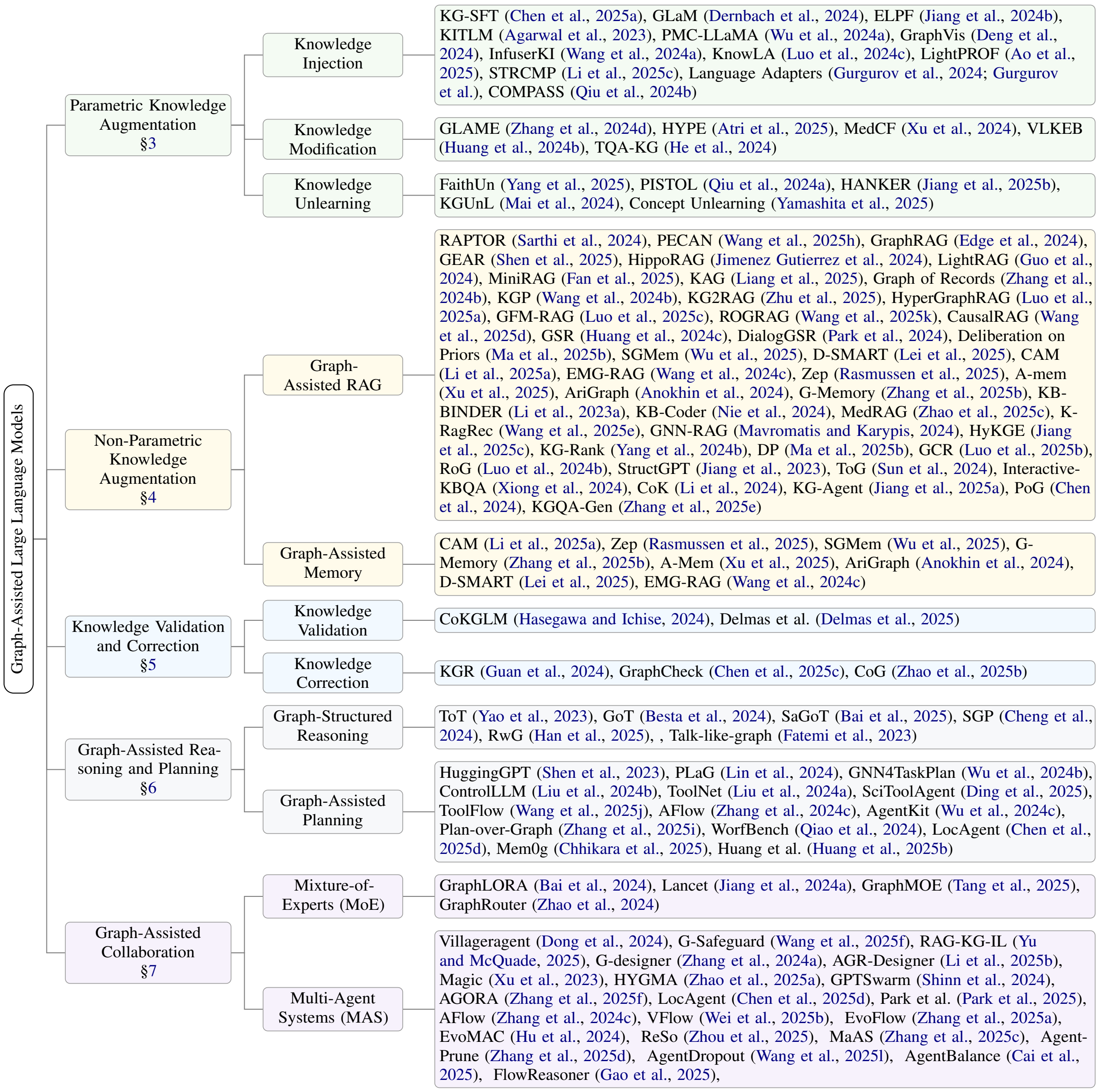

Figure 12: A taxonomy of graph-assisted large language models.

Table 1: Overview of Graph-Augmented LLM Knowledge across domains, datasets, and paradigms.

Paper	Domain	Knowledge Graph	Dataset(s)	Model	Paradigm
Knowledge Injection					
KG-SFT (2025a)	Biomedical	UMLS	MedQA, IgakuQA, RuMedDaNet, etc.	LLaMA-2-7B	Fine-tuning
GLaM (2024)	Biomedical	UMLS	UMLS tasks	LLaMA-7B	Fine-tuning
ELPF (2024b)	Biomedical	–	CMedQA, BioASQ	ChatGLM2-6B	Fine-tuning
GraphVis (2024)	General	ConceptNet	CSQA, OBQA	Llava-Mistral-7B	Fine-tuning
KITLM (2023)	Aviation	AviationKG	AeroQA	T5	Fine-tuning
PMC-LLaMA (2024a)	Biomedical	UMLS	PubMedQA, MedMCQA, etc.	LLaMA-7B	Fine-tuning
InfuserKI (2024a)	Biomedical, Movie	UMLS, MetaQA	–	LLaMA-7B	Adapters
KnowLA (2024c)	General	ConceptNet, Wikidata	CWQ, SIQA, WebQSP, etc.	LLaMA-2	Adapters
LightPROF (2025)	General	Freebase	WebQSP, CWQ	LLaMA-7B	Adapters
STRCMP (2025c)	Math	–	MILP, SAT	Code LLaMA-7B	Adapters
Language Adapters (2024)	General	ConceptNet	WikiANN, DGurgurov	mBERT	Adapters
COMPASS (2024b)	Recommendation	Self-construct	ReDial, INSPIRED	LLaMA-3.1-8B	Adapters
Knowledge Modification					
GLAME (2024d)	General	Wikipedia	COUNTERFACT/+, MQUAKE	GPT-2 XL, GPT-J	Knowledge association
HYPE (2025)	General	Wikipedia	COUNTERFACT/+, MQUAKE	GPT-2 XL, GPT-J	Knowledge association
MedCF (2024)	Biomedical	DRKG	MedCF, MedFE	ChatDoctor-7B, Meditron-7B	Dataset construction
VLKEB (2024b)	Multimodal	MMKG	VLKEB	BLIP2, MiniGPT-4, Qwen-VL, LLaVA	Dataset construction
TQA-KG (2024)	Education	CK12-QA	CK12-QA	GPT-3.5	Dataset construction
Knowledge Unlearning					
FaithUn (2025)	General	Wikidata	FaithUn	Gemma2, LLaMA-3.2-3B	Unlearning Evaluation
PISTOL (2024a)	Business	Self-construct	PISTOL	LLaMA-2-7B, Gemma-7B, Mistral-7B	Unlearning Evaluation
HANKER (2025b)	General	Self-construct	HANKER	LLaMA-2-7B	Unlearning Evaluation
KGUnL (2024)	Safety	Self-construct	AdvBench	LLaMA-2-7B	Forget-set Construction
CU (2025)	General	Self-construct	Wiki-Fact, TOFU	Mistral-7B, Llama3.1-8B, etc.	Forget-set Construction
Graph-Assisted RAG					
RAPTOR (2024)	General	–	QuALITY	GPT-4 retriever+generator	Abstractive indexing
PECAN (2025h)	General	–	NarrativeQA, Qasper, HotpotQA, etc.	Llama-3.x guided LLM	Abstractive indexing
GraphRAG (2024)	General	–	Private corpora	GPT-4 family	Abstractive indexing
GEAR (2025)	General	–	MuSiQue, etc.	BM25 + LLM agent	Extractive indexing
HippoRAG (2024)	General	–	HotpotQA, 2WikiMultihopQA, etc.	LLM + PPR traversal	Extractive indexing
LightRAG (2024)	General	–	Public corpora	LLM various	Extractive indexing
MiniRAG (2025)	On-device	–	QA benchmarks	Small SLM-based	Extractive indexing
KAG (2025)	Professional	–	HotpotQA, 2WikiMultihopQA, etc.	Framework various LLMs	Extractive indexing
Graph of Records (2024b)	General	–	WCEP, QMSum, etc.	BERT + GNN	Extractive indexing
KGP (2024b)	General	–	IIRC, HotpotQA, etc.	LLM traversal agent	Extractive indexing
KG2RAG (2025)	General	–	HotpotQA, etc.	LLM retriever + generator	Extractive indexing
HyperGraphRAG (2025a)	General	–	Multi-domain datasets	LLM various	Extractive indexing
GFM-RAG (2025c)	General	–	Multi-hop/domain sets	GraphFM	Extractive indexing
ROGRAG (2025k)	Domain-specific	–	SeedBench, etc.	Qwen2.5-7B eval	Extractive indexing
CausalRAG (2025d)	General	–	EventStoryLine, DramaQA, etc.	LLM various	Extractive indexing
GSR (2024c)	General	–	Subgraph retrieval sets	220M SLM retriever	Generative indexing
DialogGSR (2024)	General	–	OpenDialKG, KOMODIS	LLM with graph decoding	Generative indexing
DP (2025b)	General	–	ComplexWebQuestions, etc.	LLM w/ priors	Generative indexing
KB-BINDER (2023a)	General, Movie	Freebase, WikiMovies	WebQSP, GrailQA, etc.	Codex	SP-based
KB-Coder (2024)	General	Freebase	WebQSP, GrailQA, GraphQA	GPT-3.5	SP-based
MedRAG (2025c)	Biomedical	Self-construct	DDXPlux, CPDD	Mixtral-8×7B, GPT-4o, etc.	IR-based
K-RagRec (2025e)	Recommendation	Freebase	MovieLens-1M/20M, Amazon Book	LLaMA-2/3 variants	IR-based
GNN-RAG (2024)	General	Freebase	WebQSP, CWQ	LLaMA-2-Chat-7B	IR-based
HyKGE (2025c)	Biomedical	CpubMed-KG, etc.	MMCU-Medical, CMB-Exam, etc.	GPT-3.5, Baichuan	IR-based
KG-Rank (2024b)	Biomedical	UMLS	LiveQA, ExpertQA-Bio/Med, etc.	GPT-4, Baize-healthcare	IR-based
DP (2025b)	General	Freebase	WebQSP, CWQ, MetaQA	LLaMA-3.1-8B, Qwen-3-8B	IR-based
GCR (2025b)	General, Biomedical	Freebase, MedicalKG	WebQSP, CWQ, etc.	Llama-3-8B	IR-based
RoG (2024b)	General	Freebase	WebQSP, CWQ	LLaMA-2-7B-Chat	IR-based
StructGPT (2023)	General	Freebase	WebQSP, MetaQA	GPT-3/3.5	Agent-based
ToG (2024)	General	Freebase, Wikidata	CWQ, GrailQA, QALD10-en, etc.	GPT-3.5/4, LLaMA-70B	Agent-based
Interactive-KBQA (2024)	General	Freebase, Wikidata	WebQSP, CWQ, KQA-Pro	GPT-4, Mistral, LLaMA	Agent-based
CoK (2024)	General	Wikidata	FEVER, HotpotQA, FeTaQA	LLaMA-2-7B, ChatGPT	Agent-based
KG-Agent (2025a)	General	Freebase, Wikidata	WebQSP, CWQ, etc.	LLaMA-2-7B	Agent-based
PoG (2024)	General	Freebase	CWQ, WebQSP, GrailQA	GPT-3.5/4	Agent-based
Graph-Assisted Memory					
CAM (2025a)	General	–	NovelQA, QMSum	GPT-4 retriever+generator	Graph for indexing
Zep (2025)	General	–	DMR, LongMemEval	GPT-4o/4o-mini	Graph for indexing
SGMem (2025)	General	–	LongMemEval, LoCoMo	Qwen2.5-32B	Graph for indexing
A-mem (2025)	General	–	LoCoMo, DialSim	GPT-4o/4o-mini, etc.	Graph for indexing
AriGraph (2024)	General	–	TextWorld, MuSiQue, etc.	GPT-3.5/4	Graph for indexing
G-Memory (2025b)	Embodied, Game	-	ALFWorld, PDDL, etc.	Qwen-2.5-7B/14B,GPT-4o	Graph for indexing
D-SMART (2025)	General	–	MT-Bench-101	GPT-4o, Qwen-8B	Graph as store
EMG-RAG (2024c)	General	–	Self-construct	GPT-4,ChatGLM-6B,etc.	Graph as store
Knowledge Validation and Correction					
CoKGLM (2024)	General	OpenDialKG	OpenDialKG	BART	Knowledge Validation
Delmas et al. (2025)	Biomedical	ABROAD-KG	LOTUS, PubMed	Mixtral-8×7B	Knowledge Validation
KGR (2024)	General	Wikidata	SimpleQuestions, Mintaka, etc.	GPT-3.5, Vicuna	Knowledge Correction
GraphCheck (2025c)	General	Wikidata	FEVER, SciFact	GPT-4, Llama-2, ChatGLM	Knowledge Correction
CoG (2025b)	General	Self-construct	LC-QuAD, QALD-9	T5, BART	Knowledge Correction

Table 2: Overview of Graph-Assisted Reasoning and Planning across domains, datasets, and paradigms.

Paper	Domain	Graph	Dataset(s)	Model	Paradigm
Graph-Assisted Reasoning					
SGP (2024)	General	Concept Graph	CLUTRR, BBH, etc.	GPT-3.5, GPT-4, etc.	Graph-Structured Input
RwG (2025)	General	Concept Graph	AIW, LogiQA	GPT-4o, Claude-3.5, etc.	Graph-Structured Input
Talk-like-graph (2023)	Graph Problems	General Graph	GraphQA	PaLM-62B	Graph-Structured Input
ToT (2023)	General	Reasoning Tree	Game of 24, Creative Writing, etc.	GPT-4	Graph-Structured Reasoning
GoT (2024)	General	Reasoning Graph	Sorting, Keyword Counting, etc.	GPT-3.5, LLaMA-2, etc.	Graph-Structured Reasoning
SaGoT (2025)	Math	Reasoning Graph	GSM8K, MathBenchA	Qwen2-1.5B, LLaMA3-8B, etc.	Graph-Structured Reasoning
Graph-Assisted Planning					
HuggingGPT (2023)	General	Task Graph	HuggingGPT	Alpaca-7B, Vicuna-7B, etc.	Graphs as Coordinators
PLaG (2024)	General	Task Graph	AsyncHow	GPT-3.5, GPT-4, etc.	Graphs as Coordinators
GNN4TaskPlan (2024b)	General	Task Graph	TaskBench, RestBench	Baichuan2-13B, CodeLlama, etc.	Graphs as Coordinators
ControlLLM (2024b)	Multimodal	Tool Graph	Self-construct	GPT-3.5, LLaMA-7B	Graphs as Coordinators
ToolNet (2024a)	General	Tool Graph	SciQA, TabMWP, etc.	GPT-3.5	Graphs as Coordinators
SciToolAgent (2025)	General	Tool Graph	SciToolEval, Safeguard Database, etc.	OpenAI o1, Qwen2.5-72B, etc.	Graphs as Coordinators
ToolFlow (2025j)	General	Tool Graph	ToolBench, BFCL-v2, etc.	GPT-4, LLaMA-3.1-8B	Graphs as Coordinators
AFlow (2024c)	General	Task Graph	GSM8K, HumanEval, etc.	DeepSeekV2.5, GPT-4o, etc.	Graphs as Coordinators
AgentKit (2024c)	General	Task Graph	Crafter, WebShop	GPT-3.5, GPT-4	Graphs as Coordinators
Plan-over-Graph (2025i)	General	Task Graph	Self-construct	GPT-4o, Llama-3.1, Qwen2.5, etc.	Graphs as Coordinators
WorfBench (2024)	General	Task Graph	WorfBench	GPT-4, GPT-3.5, etc.	Graphs as Coordinators
LocAgent (2025d)	Code	Code Graph	Loc-Bench	Qwen-2.5-Coder	Graphs as Environments
Mem0g (2025)	General	Concept Graph	LOCOMO	GPT-4o-mini	Graphs as Environments
Huang et al. (2025b)	General	Task Graph	AI2-THOR	not mentioned	Graphs as Environments

Table 3: Overview of Graph-Assisted Collaboration across domains, datasets, and paradigms.

Paper	Domain	Dataset(s)	Model	Paradigm
Mixture-of-Experts (MoE)				
GraphLORA (2024)	General	ARC-Challenge, BoolQ, OpenBookQA, SocialIQA	LLaMA-3-8B, Qwen2-7B, Yi-1.5-9B	MoE
Lancet (2024a)	General	WikiText	LLaMA-3-8B	MoE
GraphMOE (2025)	General	ARC, BoolQ, OpenBookQA, SocialIQA, etc.	GPT2-S-MoE, GPT2-L-MoE	MoE
GraphRouter (2024)	General	Natural Instructions v2, TriviaQA, MMLU, HumanEval	T5, GPT-3.5	MoE
Multi-Agent Systems (MAS)				
Villageragent (2024)	General	OpenDialKG	GPT-4, Gemini Pro, GLM-4	For Performance
RAG-KG-IL (2025)	Biomedical	Private 20 questions	GPT-4o	For Performance
G-designer (2024a)	General	MMLU, GSM8K, MultiArith, SVAMP, HumanEval	GPT-3.5/4	For Performance
Magic (2023)	Customer Service	Magic benchmark	GPT-3.5/4, LLaMA-2-70B, PaLM 2, Claude 2, etc.	For Performance
HYGMA (2025a)	MARL, Robotics	MAPF, StarCraft II micromanagement	Actor-Critic architecture with Hypergraph neural modules	For Performance
GPTSwarm (2024)	General	MMLU, Mini CrossWords, HumanEval, GAIA	GPT-3.5/4	For Performance
AFlow (2024c)	General	GSM8K, HumanEval, etc.	DeepSeekV2.5, GPT-4o, etc.	For Performance
AGORA (2025f)	Math, Multimodal	GSM8K, AQuA, MATH-500, MME-RealWorld	Deepseek-R1-1.5B, Qwen2-1.5B-Instruct, etc.	For Performance
LocAgent (2025d)	Code	CodeSearchNet, CoSQA	GPT-4, CodeLLaMA, StarCoder, SantaCoder	For Performance
Park et al. (2025)	Industrial	Self-construct	Qwen2.5-32B, Gemma-3-27B	For Performance
EvoFlow (2025a)	General	GSM8K, MultiArith, HumanEval, MBPP, ALFWorld	GPT-4o-mini, LLaMA-3.1-70B, Qwen-2-72B, etc.	For Performance
EvoMAC (2024)	General	rSDE-Bench, HumanEval	GPT-4o-mini, Claude-3.5-Sonnet, and Gemini-1.5-flash	For Performance
ReSo (2025)	General	MATH, SciBench	Gemini-2.0-Flash, GPT-4o, Qwen-2.5- Max, etc.	For Performance
ARG-Designer (2025b)	General	GSM8K, MMLU, SVAMP, AQuA, HumanEval	GPT-4o	For Performance
MaAS (2025c)	General	GSM8K, MATH, HumanEval, MBPP, GAIA	GPT-4o-mini, Qwen2.5-72B, LLaMA-3.1-70B	For Performance
AgentPrune (2025d)	General	MMLU, SVAMP, AQuA, HumanEval	GPT-3.5/4	For Eficiency
AgentDropout (2025l)	General	GSM8K, MMLU, AQuA, MultiArith, etc.	Llama3-8B, Qwen2.5-72B-Instruct, etc.	For Eficiency
AgentBalance (2025)	General	MMLU, HumanEval, MATH	Qwen3-8B, DeepSeek-R1, etc.	For Eficiency
FlowReasoner (2025)	Code	BigCodeBench, HumanEval, MBPP	GPT-4o-mini, DeepSeek-R1-Distill-Qwen	For Robustness
VFlow (2025b)	Code	VerilogEval	GPT-4o, DeepSeek-V3, GPT-4o-mini	For Robustness
G-Safeguard (2025f)	Security	CSQA, MMLU, GSM8K	GPT-4o, LLaMA-3.1-70B, Claude-3.5-haiku, etc.	For Robustness