Synergizing Large Language Models and Knowledge Graphs in Science: A Survey

Zhihui Zhu[1], **Yuqi Tang**[1,2], **Qiang Zhang**[1,2]*, **Keyan Ding**[1]*
[1]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University
[2]ZJU-UIUC Institute, Zhejiang University
Zhihui.Zhu01@outlook.com, {qiang.zhang.cs, dingkeyan}@zju.edu.cn

Abstract

The integration of large language models (LLMs) and scientific knowledge graphs (SciKGs) is emerging as a powerful paradigm in AI for science. This survey examines their bidirectional synergy: LLMs accelerate SciKG construction via automated extraction, completion, and maintenance, while SciKGs make LLMs more factual and explainable and strengthen scientific reasoning and comprehension. We organize the survey around these two directions and adopt a task-centered framework that aligns technical methods with scientific objectives. Building on this framework, we (i) chart techniques that automate and sustain SciKG construction with LLMs, (ii) systematize how SciKGs ground and guide LLMs to improve factuality, explainability, and reasoning, (iii) synthesize representative applications in biomedicine, chemistry, and materials, and (iv) outline open problems and research directions around knowledge consistency and conflict handling, temporal modeling and updating, scalable retrieval and inference, and rigorous evaluation. This work's insights recast LLMs and SciKGs as complementary components of a dynamic, self-improving knowledge infrastructure for scientific discovery, providing a clear foundation for building grounded, transparent, and knowledge-driven models in high-stakes scientific domains.

1 Introduction

The rapid advancement of large language models (LLMs) has transformed the landscape of artificial intelligence in science, enabling new forms of hypothesis generation, literature integration, and experimental planning [1–5]. However, there are major barriers to the reliable application of LLMs in high-stakes scientific areas due to their inherent shortcomings, especially their propensity for hallucinations, lack of verifiability, and opacity in reasoning [6–9]. At the same time, scientific knowledge graphs (SciKGs), which encode structured, curated, and interlinked facts, offer a principled foundation for trustworthy knowledge representation[10–14]. Yet, their construction and maintenance remain labor-intensive and often lag behind the pace of scientific discovery[15, 16].

A growing body of work suggests that the solution lies not in choosing between LLMs and KGs, but in synergizing them[17–21]. This survey aims to explore the bidirectional synergy between LLMs and SciKGs, a paradigm in which each component compensates for the other's weaknesses and amplifies its strengths (Figure 1). On one hand, LLMs serve as powerful engines for automating the construction and evolution of SciKGs: through domain-adaptive pre-training on scientific corpora and end-to-end pipeline design, they extract entities and relations from unstructured text, complete missing links, and dynamically update knowledge bases[22–26]. On the other hand, SciKGs act as grounding mechanisms for LLMs, providing structured, verifiable facts that reduce hallucinations, enhance

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science.

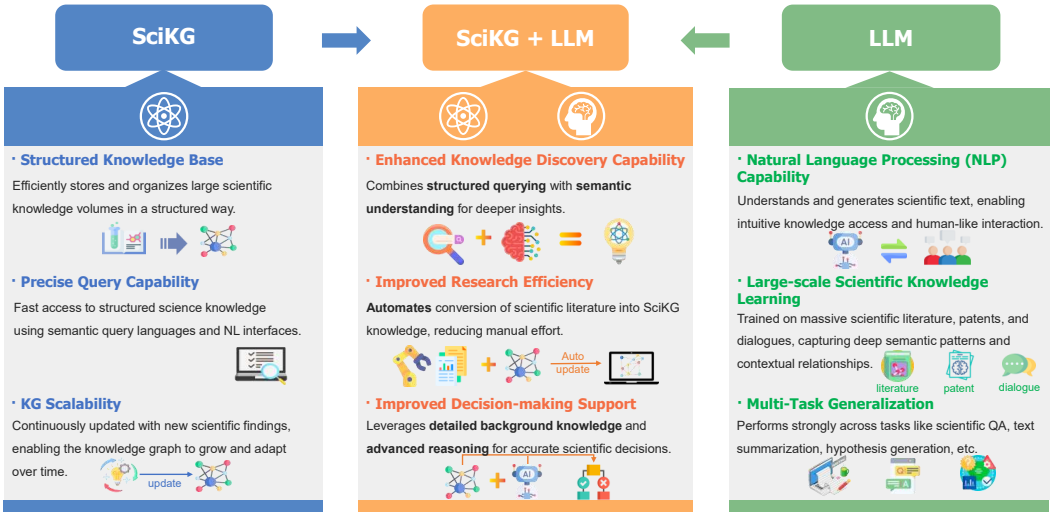

Figure 1: Illustration of the bidirectional synergy between scientific knowledge graphs (SciKGs) and large language models (LLMs). SciKGs serve as structured, scalable repositories of scientific knowledge, while LLMs bring robust natural language understanding, generation, and generalization. Their integration creates a synergistic framework that enhances knowledge discovery by merging structured querying with semantic reasoning, improves research efficiency through automated knowledge extraction from scientific literature, and strengthens decision-making support by leveraging detailed background knowledge and advanced inference.

interpretability, and enable complex, multi-step scientific reasoning through retrieval-augmented generation (RAG) and knowledge-guided prompting[27–34].

This dual role, with "SciKGs for LLMs" and "LLMs for SciKGs", represents a fundamental shift in how we approach scientific knowledge engineering. It moves beyond static, manually curated knowledge bases and ungrounded generative models toward a dynamic, self-improving ecosystem of knowledge creation and utilization. This integration is further enhanced by multimodal scientific data, including text, images, and molecular structures. In these cases, LLMs work as adaptable interfaces for knowledge extraction and reasoning, while SciKGs integrate disparate sources to facilitate cross-modal understanding[35, 36].

In this survey, we provide a comprehensive and structured review of this emerging synergy. We examine both directions of the interaction: (1) how LLMs empower the automated construction, completion, and maintenance of SciKGs; and (2) how SciKGs enhance the reliability, accuracy, and interpretability of LLMs in scientific tasks. We analyze key methodologies, including prompt engineering and retrieval mechanisms, and discuss real-world applications across biomedicine, chemistry, and materials science. We also critically assess challenges such as knowledge consistency, efficiency, and scalability, and outline future directions for building more robust and trustworthy LLM systems for scientific discovery.

2 LLMs for SciKGs

LLMs are reshaping the paradigm of SciKG construction by automating and enriching the integration of knowledge from heterogeneous sources. This section outlines the dual function of LLMs in this process (Figure 2): (a) extracting and aligning entities, relations, and facts across heterogeneous scientific data and achieving end-to-end construction; and (b) enhancing the SciKG through knowledge completion, reasoning, and automated maintenance.

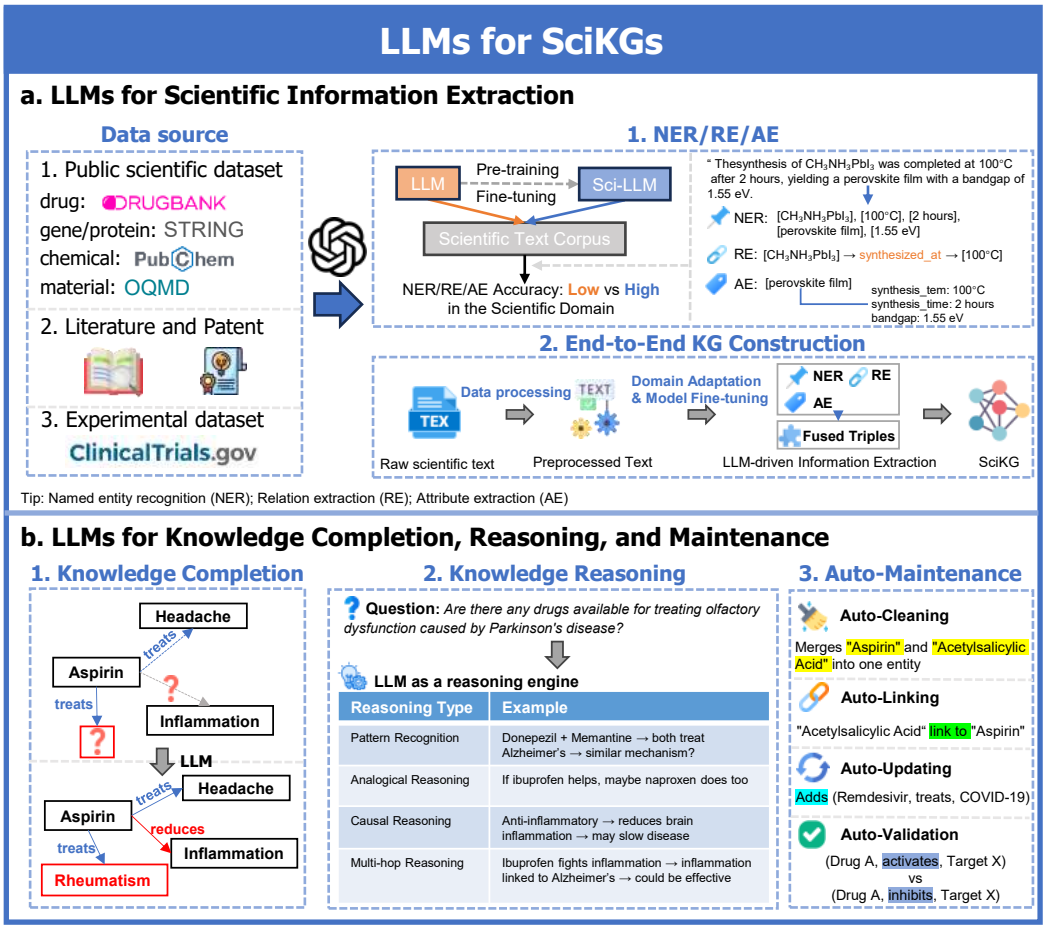

Figure 2: LLMs for SciKGs. (a) LLMs facilitate automated information extraction from scientific databases; (b) LLMs empower automated completion, reasoning, and maintenance.

2.1 LLMs enhance scientific information extraction

Named Entity Recognition (NER) and Relation Extraction (RE) are foundational steps in SciKG construction. Traditional methods rely on rule-based systems or feature engineering, which require extensive manual design and domain knowledge [37–41]. These approaches struggle with the terminological diversity, syntactic complexity, and semantic nuance of scientific texts, limiting scalability and generalization. LLMs address these challenges through deep semantic understanding and contextual reasoning[42–46]. They can automatically identify entities and relations from diverse scientific sources, including published literature, patents, public domain-specific databases, and raw experimental records, significantly improving extraction accuracy, coverage, and efficiency.

Pre-trained language models such as BERT [47] and T5 [48] capture rich linguistic representations and serve as strong backbones for scientific information extraction. Through domain-adaptive pre-training on large corpora of scientific text, such as PubMed and arXiv, these models acquire a deeper understanding of domain-specific terminology and context. When further fine-tuned on labeled datasets, they achieve high accuracy and scalability in NER and RE. For instance, BioBERT [22], a domain-adapted version of BERT, achieves state-of-the-art performance in biomedical NER and RE after fine-tuning. Similarly, MatSciBERT [49] demonstrates strong efficacy in tasks such as chemical component recognition when adapted to downstream tasks. These models show that the combination of domain-adaptive pre-training and task-specific fine-tuning enables precise and generalizable scientific knowledge extraction.

Building on these advances, modern frameworks pursue fully end-to-end SciKG construction with automated NER and RE as the core, integrated into a unified pipeline that turns raw text directly into structured graphs. Unlike modular systems (which separate entity extraction, relation identification, and graph assembly, risking error propagation), these approaches use LLMs to jointly optimize NER and RE while organizing results into graphs in one pass. For example, ReguloGPT [50] leverages GPT-4 and structured prompting to simultaneously identify entities (e.g., m6A regulators, cancer types via NER) and extract N-ary relations (e.g., regulatory interactions via RE), then builds the m6A-KG from 400 PubMed titles. Similarly, Ma et al. [24] combine prompt engineering and chain-of-thought (CoT) reasoning to anchor polymer retrosynthesis SciKG construction to NER and RE: the LLM first recognizes entities (e.g., monomers, catalysts, reaction types) from scientific text, then extracts directional relations (e.g., "monomer A reacts with monomer B under catalyst C to form polymer D"), and finally maps these entity-relation pairs directly to reaction graphs. These works reflect a paradigm shift: from "extract-then-integrate" (separate NER/RE then graph assembly) to "generate-as-structure", where LLMs act as unified compilers.

2.2 LLMs empower scientific knowledge completion, reasoning, and maintenance

In the construction of SciKGs, LLMs serve as effective technological drivers for overcoming three principal challenges of conventional KGs: the incompleteness of triples, delays in the integration of new knowledge, and the inefficiency of manual upkeep by operating as advanced knowledge completion instruments, intelligent reasoning engines for discovery, and automated maintenance systems[23, 51–53]. These features jointly enhance the completeness, accuracy, and dynamic adaptability of SciKGs, hence maintaining their cutting-edge quality and reliability.

LLMs excel at identifying missing facts in SciKGs through knowledge graph completion (KGC), including entity prediction, relation prediction, and triple classification. Two primary paradigms dominate: Predictor as Encoder (PaE) and Predictor as Generator (PaG) [17]. The PaE technique optimizes and evaluates using prediction heads on the LLM encoding representation; nevertheless, it is constrained by resource utilization and generalization capacity. Conversely, the PaG technique leverages the generative capacity of LLMs to produce absent entities or relationships without necessitating supplementary parameter training, rendering it appropriate for novel entity identification in open-world contexts. For example, DDI-GPT[54] conceptualizes DDI as a natural language generation task, integrating data such as SMILES structures, targets, and pathways to formulate prompt templates, thereby directing the model to identify interaction types (e.g., synergistic or antagonistic), thus proficiently addressing critical deficiencies in drug SciKGs. Similarly, KGAREVION [23] dynamically generates candidate triples in response to biomedical queries and refines them through verification and correction modules, enabling robust reasoning over incomplete knowledge.

LLMs not only thrive in KGC tasks but also demonstrate diverse capabilities as fundamental reasoning engines for scientific knowledge discovery, including pattern recognition, analogical reasoning, causal inference, and multi-hop reasoning. These capabilities allow LLMs to discern hidden linkages within complex, unstructured scientific literature and formulate novel insights that can be structured and integrated to enhance the completeness and depth of SciKGs. For instance, BioGPT [55] leverages large-scale pattern recognition to autonomously identify potential associations between pharmaceuticals and diseases (e.g., "Aspirin $\rightarrow$ reduced risk $\rightarrow$ heart attack"). These co-occurrence-derived hypotheses, while emerging from text, can serve as candidate triples for enriching biomedical SciKGs. In chemistry, StructChem [56] employs a multi-step prompting technique to guide LLMs in systematically extracting products from reactants, ensuring adherence to chemical principles through a self-validation process. The resulting reaction pathways can be directly mapped to chemical KG schemas, facilitating automated KG population. In materials science, LLMs in MKG creation employ analogical reasoning via semantic similarity to generate novel material-application combinations [57], showcasing the capacity for generalization from the known to the unknown. Similarly, Bai et al. [27] demonstrate causal reasoning capabilities of LLMs in analyzing ferroelectric materials, using ChatGPT to elucidate relationships between synthesis conditions and material properties. When formalized, such analogical and causal insights can be transformed into structured knowledge to expand and deepen SciKGs.

LLMs also exhibit distinct benefits in the automated maintenance and updating of SciKGs, facilitating the ongoing evolution and quality assurance of KGs. Due to the highly dynamic and varied characteristics of scientific information, conventional manual maintenance approaches encounter

considerable constraints regarding timeliness and scalability. LLMs, with their robust semantic comprehension and creation skills, offer systematic, automated assistance for SciKGs. Initially, LLMs are capable of detecting and rectifying semantic discrepancies and factual inaccuracies inside SciKGs. The SAC-KG framework [58] employs multi-dimensional validation methods to assess the numerical rationality of triples and rectify logical consistency, thereby significantly improving the quality and coherence of the SciKG. Furthermore, LLMs can semantically match newly retrieved things with pre-existing knowledge. NetMe 2.0 [59] employs a tailored entity linker, OntoTagMe, for biomedical entity recognition and normalization, alongside the semantic analysis tool SpaCy for constructing syntactic dependency trees, to precisely extract and standardize gene-disease and drug-target interactions. Moreover, LLMs facilitate real-time knowledge integration from literary sources. LightRAG [60] utilizes a dual-layer retrieval system to swiftly identify newly incorporated nodes and associations within local subgraphs, while ensuring their semantic coherence through the global graph framework, hence facilitating efficient and resilient knowledge growth. Ultimately, LLMs can evaluate the reliability of information sources and curate high-quality triples. TrustLLM [61] employs criteria like citation count and journal impact factor to rank various data sources, thereby substantially improving the authority and stability of SciKGs. In conclusion, LLMs are emerging as the fundamental technological force propelling the sustained advancement of SciKGs. Their extensive capabilities in knowledge cleansing, standardization, updating, and validation offer robust assistance for constructing high-precision, adaptive scientific knowledge infrastructures.

3 SciKGs for LLMs

This section demonstrates how SciKGs augment LLMs across two dimensions (Figure 3): (a) enhancing factuality and explainability by grounding outputs in structured knowledge, enabling fact verification and evidence-based reasoning; and (b) improving scientific reasoning and comprehension through context-enhanced QA, multi-hop inference, and structured summarization.

3.1 SciKGs strengthen factuality and explainability in LLMs

SciKGs serve as organized knowledge repositories for LLMs, significantly contributing to the improvement of their specialized comprehension, reasoning capabilities, and generalization efficacy. Integrating SciKGs into the pretraining or inference stages of LLMs significantly enhances the model's capacity to semantically analyze intricate scientific texts. Common methodologies encompass knowledge injection utilizing structured embeddings (e.g., CoLAKE [62], which employs a unified graph structure to represent language-knowledge interaction), modular adapter-driven knowledge integration (e.g., K-ADAPTER [63], which engages domain knowledge without altering backbone parameters), dynamic knowledge fusion within the retrieval-augmented generation (RAG) framework (e.g., KG-FM [27], which leverages materials science SciKGs to markedly enhance question-answering precision), and multimodal knowledge alignment for heterogeneoous data (e.g., MR-MKG [64], which processes text, images, and structured triples via dedicated encoders and aligns them with a knowledge adapter). These tactics not only augment the LLM's comprehension of long-tail entities and multi-hop relationships but also furnish knowledge backing for its deployment in high-precision scientific endeavors.

Building on this foundation, SciKG is extensively utilized to develop fact-checking mechanisms for LLM outputs to alleviate the "hallucination" issue frequently encountered in the scientific domain, thereby circumventing the potential hazards associated with factual inaccuracies in high-stakes areas such as pharmaceutical research and biological diagnostics. Researchers have methodically included SciKGs in three phases: pre-generation retrieval augmentation, dynamic validation during generation, and post-generation triple validation, establishing a closed-loop verification system. During the pre-generation phase, SciKG is employed to augment input comprehension; for instance, MedRAG [65] connects symptom and disease entity pathways via a four-tiered medical SciKG to direct the model in forming a reasoning foundation grounded in authoritative information. In the generation phase, techniques such as KG-Rank [32] integrate the UMLS to create a high-quality triple repository and implement multi-tiered ranking systems to dynamically evaluate the alignment of output with established knowledge during decoding. During the post-generation phase, CoK (Chain-of-Knowledge) [66] necessitates that LLMs explicitly cite structured triples from SciKG, so creating traceable and verifiable reasoning chains to guarantee that each conclusion is interpretable and aligns with established knowledge. These methods together establish a closed-loop verification system that

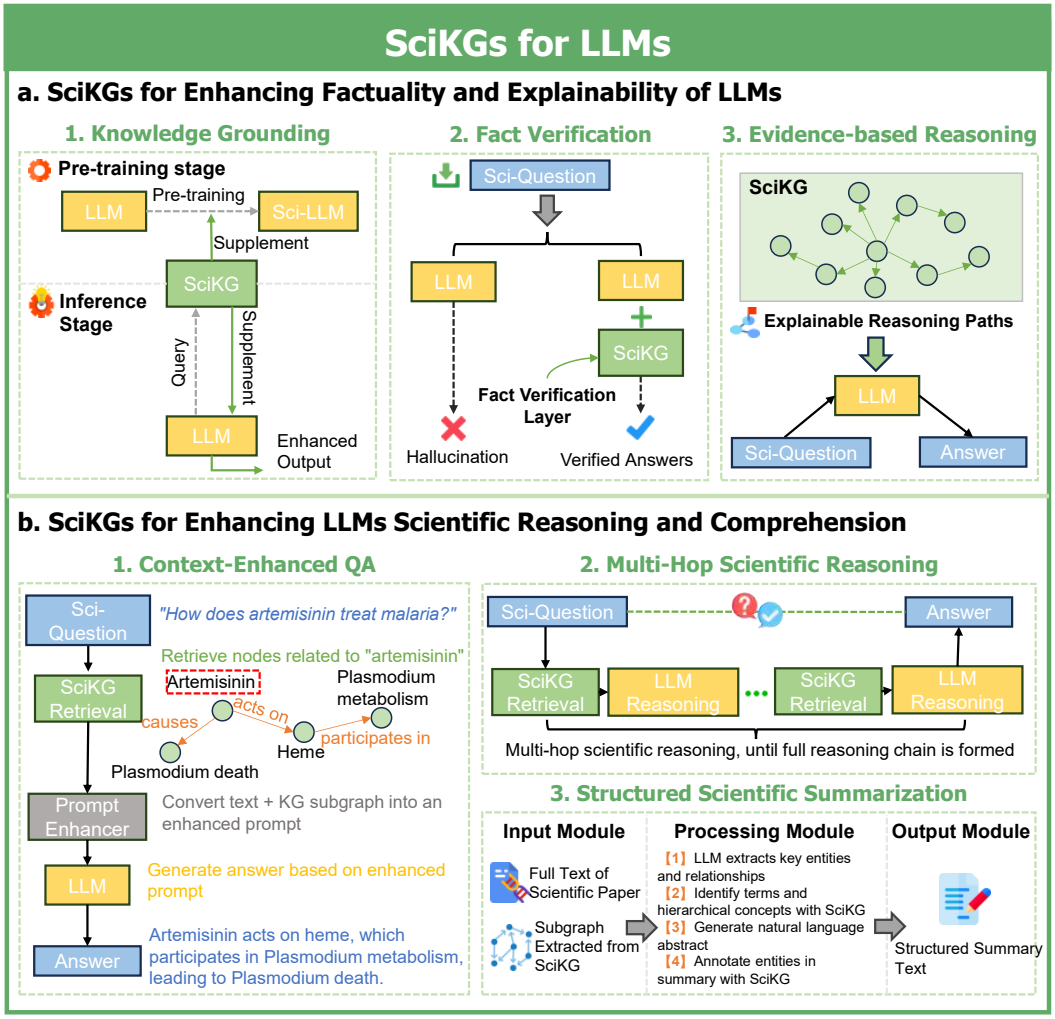

Figure 3: SciKGs for LLMs. (a) Enhancing factuality and explainability of LLMs via knowledge grounding, fact verification, and evidence-based reasoning. (b) Improving scientific reasoning and comprehension of LLM through context-enhanced QA, multi-hop scientific reasoning, and structured scientific summarization.

spans from input comprehension to output rectification, markedly improving the professionalism and dependability of model outputs. It is important to acknowledge that while SciKG offers robust knowledge support, certain inquiries may beyond the existing boundaries of SciKGs. The CogMG framework [67] resolves this issue through a collaborative enhancement method. When a query exceeds the SciKG's scope, LLMs are instructed to methodically disaggregate the requisite knowledge triples and autonomously validate them by consulting external documents. This method addresses the issue of insufficient knowledge coverage while guaranteeing the precision of newly integrated information via external validation, hence facilitating new avenues for knowledge creation in the scientific field.

Moreover, the organized depiction of SciKG offers a viable means to enhance the transparency and interpretability of the LLM reasoning process. The "black-box" nature of traditional LLMs restricts their reliability in research settings, whereas the symbolic and verifiable knowledge frameworks offered by SciKG facilitate the visualization and auditing of reasoning chains. In biomedical question-answering tasks, the KGT [33] framework correlates user questions with entity routes in SciKG, dynamically constructing the best reasoning subgraph and identifying the most pertinent knowledge paths via text embeddings to provide a comprehensive, traceable reasoning chain. MechGPT [68] attains graphical support for reasoning and improved control in physical modeling activities by

developing a material mechanics ontology graph that includes nodes, edges, and subgraph hierarchy data. These methods utilize the structured triples and graph framework of SciKG, ensuring that each reasoning step is linked to specific knowledge facts, hence greatly improving the transparency and scientific validity of model outputs.

3.2 SciKGs improve scientific reasoning and comprehension in LLMs

Scientific questions are generally domain-specific, dependent on fundamental knowledge, and frequently entail intricate multi-step reasoning[69–71]. In this context, SciKGs act as essential structured knowledge repositories that substantially enhance the question-answering (QA), reasoning, and summarizing capacities of LLMs. SciKG gives traceable factual assistance and organized advice for constructing reasoning paths, guaranteeing logical coherence, and producing coherent summaries, so enhancing the professionalism and credibility of scientific text processing.

A central challenge in scientific QA is identifying relevant knowledge paths within large-scale, heterogeneous knowledge networks. GraphRAG [65, 72, 73] introduces a paradigm shift by retrieving task-specific subgraphs from SciKGs and injecting them as structured context into the LLM prompt. This approach grounds generation in verifiable facts and supports dynamic knowledge retrieval. For instance, MedGraphRAG [72] leverages Meta-MedGraphs to construct disease-drug-mechanism pathways, significantly enhancing the interpretability of medical QA. Similarly, KGT [33] integrates clinical guidelines and therapeutic targets into cancer-related reasoning, employing subgraph pruning to filter noise and focus the model on high-relevance knowledge segments. KG-FM [27] utilizes material science SciKGs with Cypher-based querying to dynamically retrieve synthesis routes and performance metrics, improving precision in materials QA.

In addition to single-hop retrieval, scientific reasoning frequently necessitates multi-step causal inference. LLMs alone are prone to logical inconsistencies due to incomplete or implicit knowledge. SciKGs provide external symbolic support that enables explicit CoT construction. Self-BioRAG [34] employs a "reason-reflect-verify" loop, where medical KGs are used to generate candidate inference subgraphs, which are then validated against known facts to correct erroneous reasoning paths. MedReason [28] autonomously discovers latent reasoning chains between queries and answers, using SciKG to generate structured CoT prompts that guide the LLM toward clinically valid outputs. These methods transform reasoning into a transparent, auditable process, where each inference step is anchored to a specific triple or subgraph. For tasks with heterogeneous inputs (e.g., language, images, and KG subgraphs), the KAM-CoT system [36] enhances CoT reasoning without LLM fine-tuning. It uses cross-attention to align these three modalities and a dual-phase training strategy, achieving a high accuracy on ScienceQA and showing how multimodal SciKG integration boosts logical coherence in complex scientific inference.

In scientific summarization, where factual consistency is crucial, SciKGs augment reliability by systematic fact validation. Traditional LLMs may introduce knowledge drift or vague expressions. Incorporating SciKG into the summarizing process provides structured knowledge assistance, allowing the model to validate facts and arrange content according to explicit triples. FASUM [74] addresses this by extracting entity-relation triples from input texts to build a local KG, which is then integrated via a graph attention network to guide summary generation. A post-hoc Fact Checker (FC) module further validates output consistency. These methods illustrate how SciKGs restructure the LLM's reasoning in the QA process, transforming it from a generative black box into a knowledge-grounded, traceable, and verifiable agent, enabling more trustworthy and scientifically rigorous AI-assisted discovery.

4 Challenges and Future Directions

4.1 Key Challenges

Despite the promising synergy between LLMs and SciKGs, several critical challenges hinder their robust and scalable integration in scientific domains.

- **Knowledge Consistency and Conflict Resolution.** Automated knowledge extraction by LLMs may generate contradictions or redundancies in SciKGs, particularly when analyzing developing or conflicting literature, such as disputed protein functions or contradictory material qualities.

Addressing these inconsistencies necessitates both logical reasoning and the implementation of provenance-aware conflict detection and domain-specific validation procedures, which are currently insufficiently developed.

- **Dynamic Knowledge Updating and Temporal Reasoning.** Scientific knowledge is inherently temporal and evolving. Although LLMs can extract new facts, it is still difficult to integrate them into a time-aware knowledge graph structure that tracks the evolution of beliefs, preserves historical states, and supports temporal queries. Most current SciKGs lack explicit temporal modeling, limiting their utility for longitudinal analysis.
- **Scalability and Efficiency.** End-to-end pipelines combining LLM inference with large-scale KG querying (e.g., in RAG) suffer from high computational costs and latency. As SciKGs expand, indexing, retrieval, and grounding operations become bottlenecks, particularly in domains like biomedicine with millions of entities. Efficient indexing strategies and lightweight LLMs tailored for scientific reasoning are urgently needed.
- **Evaluation and Benchmarking.** There is a lack of standardized, domain-specific benchmarks for evaluating the joint performance of LLM–SciKG systems. Metrics often focus on isolated tasks (e.g., relation extraction F1-score or QA accuracy), failing to capture higher-order outcomes such as scientific insight generation, hypothesis validity, or reproducibility. Ground-truth curation for such complex outputs is labor-intensive and subjective.

4.2 Future Directions

To realize the full potential of LLM–SciKG integration, we envision several transformative directions for future research.

- **Neural-Symbolic Fusion Architectures.** Rather than treating LLMs and SciKGs as separate modules, deeper integration through neural-symbolic computing can unify probabilistic reasoning with logical inference. Hybrid architectures, which integrate symbolic rules to regulate LLM outputs while neural models are trained to perform KG queries, can enhance both reliability and flexibility in scientific reasoning.
- **Multimodal Knowledge Graph Foundation Models.** Just as LLMs are foundation models for text, we anticipate the emergence of multimodal scientific foundation models that natively embed and reason over multimodal SciKGs. These models would be jointly trained on text, structures, spectra, and images, with built-in grounding to structured knowledge, enabling seamless cross-modal understanding and generation.
- **Toward Autonomous Scientific Agents.** Future systems could evolve into autonomous scientific agents that iteratively generate hypotheses using LLMs, validate them against SciKGs, design experiments, and update knowledge bases with new findings. Such agents would operate in a closed-loop cycle of "propose–test–learn", enabling self-driving discovery in domains like drug design or materials optimization.
- **Open and Interoperable SciKG Ecosystems.** A fragmented landscape of isolated knowledge graphs limits scalability. Future efforts should promote open standards (e.g., domain-agnostic ontologies, FAIR principles) and federated architectures that allow distributed, privacy-preserving querying across multiple SciKGs, therefore enabling large-scale, collaborative knowledge ecosystems.

In summary, these directions point toward a future in which LLM–SciKG integration evolves from a tool for information retrieval into a foundational framework for trustworthy, autonomous, and knowledge-driven scientific discovery.

References

[1] Qiang Zhang, Keyan Ding, Tianwen Lv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. Scientific large language models: A survey on biological & chemical domains. *ACM Computing Surveys*, 57(6):1–38, 2025.

[2] Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D Bocarsly, Andres M Bran, Stefan Bringuier, L Catherine Brinson, Kamal Choudhary, Defne Circi, et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital discovery*, 2(5):1233–1250, 2023.

[3] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

[4] Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

[5] Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. A comprehensive survey of scientific large language models and their applications in scientific discovery. *arXiv preprint arXiv:2406.10833*, 2024.

[6] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[7] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pages 15696–15707. PMLR, 2023.

[8] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[9] Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*, 2023.

[10] Sören Auer, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal. Towards a knowledge graph for science. In *Proceedings of the 8th international conference on web intelligence, mining and semantics*, pages 1–6, 2018.

[11] Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. Knowledge graphs: Opportunities and challenges. *Artificial intelligence review*, 56(11):13071–13102, 2023.

[12] Jiaoyan Chen, Hang Dong, Janna Hastings, Ernesto Jiménez-Ruiz, Vanessa López, Pierre Monnin, Catia Pesquita, Petr Škoda, and Valentina Tamma. Knowledge graphs for the life sciences: Recent developments, challenges and opportunities. *arXiv preprint arXiv:2309.17255*, 2023.

[13] Finlay MacLean. Knowledge graphs and their applications in drug discovery. *Expert opinion on drug discovery*, 16(9):1057–1069, 2021.

[14] Xue Zheng, Bing Wang, Yunmeng Zhao, Shuai Mao, and Yang Tang. A knowledge graph method for hazardous chemical management: Ontology design and entity identification. *Neurocomputing*, 430:104–111, 2021.

[15] Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. A comprehensive survey on automatic knowledge graph construction. *ACM Computing Surveys*, 56(4):1–62, 2023.

[16] Marvin Hofer, Daniel Obraczka, Alieh Saeedi, Hanna Köpcke, and Erhard Rahm. Construction of knowledge graphs: Current state and challenges. *Information*, 15(8):509, 2024.

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

[18] Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323, 2022.

[19] Hanieh Khorashadizadeh, Fatima Zahra Amara, Morteza Ezzabady, Frédéric Ieng, Sanju Tiwari, Nandana Mihindukulasooriya, Jinghua Groppe, Soror Sahri, Farah Benamara, and Sven Groppe. Research trends for the interplay between large language models and knowledge graphs. *arXiv preprint arXiv:2406.08223*, 2024.

[20] Alireza Ghafarollahi and Markus J Buehler. Sciagents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, page 2413523, 2024.

[21] DaiFeng Li and Fan Xu. Synergizing knowledge graphs with large language models: a comprehensive review and future prospects. *arXiv preprint arXiv:2407.18470*, 2024.

[22] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[23] Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, and Marinka Zitnik. Kgarevion: an ai agent for knowledge-intensive biomedical qa. In *ICLR*, 2025.

[24] Qinyu Ma, Yuhao Zhou, and Jianfeng Li. Automated retrosynthesis planning of macromolecules using large language models and knowledge graphs. *Macromolecular Rapid Communications*, page 2500065, 2025.

[25] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021.

[26] Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. Kg-gpt: A general framework for reasoning on knowledge graphs using large language models. *arXiv preprint arXiv:2310.11220*, 2023.

[27] Xuefeng Bai, Song He, Yi Li, Yabo Xie, Xin Zhang, Wenli Du, and Jian-Rong Li. Construction of a knowledge graph for framework material enabled by large language models and its application. *npj Computational Materials*, 11(1):51, 2025.

[28] Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.

[29] Yuan Sui, Yufei He, Zifeng Ding, and Bryan Hooi. Can knowledge graphs make large language models more trustworthy? an empirical study over open-ended question answering. *arXiv preprint arXiv:2410.08085*, 2024.

[30] Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3091–3110, 2024.

[31] Yiyan Deng, Shen Zhao, Yongming Miao, Junjie Zhu, and Jin Li. Medka: A knowledge graph-augmented approach to improve factuality in medical large language models. *Journal of Biomedical Informatics*, page 104871, 2025.

[32] Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yu He Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, et al. Kg-rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. *arXiv preprint arXiv:2403.05881*, 2024.

[33] Yichun Feng, Lu Zhou, Chao Ma, Yikai Zheng, Ruikun He, and Yixue Li. Knowledge graph–based thought: a knowledge graph–enhanced llm framework for pan-cancer question answering. *GigaScience*, 14:giae082, 2025.

[34] Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129, 2024.

[35] Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Huajun Chen, and Wen Zhang. Mygo: Discrete modality information as fine-grained tokens for multi-modal knowledge graph completion. *arXiv e-prints*, pages arXiv–2404, 2024.

[36] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18798–18806, 2024.

[37] Ramón AA Erhardt, Reinhard Schneider, and Christian Blaschke. Status of text-mining techniques applied to biomedical text. *Drug discovery today*, 11(7-8):315–325, 2006.

[38] Zhihao Yang, Hongfei Lin, and Yanpeng Li. Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature. *Computational biology and chemistry*, 32 (4):287–291, 2008.

[39] Paul Thompson, John McNaught, Simonetta Montemagni, Nicoletta Calzolari, Riccardo Del Gratta, Vivian Lee, Simone Marchi, Monica Monachini, Piotr Pezik, Valeria Quochi, et al. The biolexicon: a large-scale terminological resource for biomedical text mining. *BMC bioinformatics*, 12(1):397, 2011.

[40] Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim, Hideki Mima, and Junichi Tsujii. The genia corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the human language technology conference*, pages 73–77. Morgan Kaufmann Publishers Inc. San Francisco, 2002.

[41] Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36 (suppl_1):D344–D350, 2007.

[42] Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. Vocabulary modifications for domain-adaptive pretraining of clinical language models. In *HEALTHINF*, pages 180–188, 2022.

[43] Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52, 2023.

[44] Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. Pre-trained language model for biomedical question answering. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 727–740. Springer, 2019.

[45] Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023.

[46] Junho Kim, Yeachan Kim, Jun-Hyung Park, Yerim Oh, Suho Kim, and SangKeun Lee. Melt: Materials-aware continued pre-training for language model adaptation to materials science. *arXiv preprint arXiv:2410.15126*, 2024.

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[48] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[49] Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1): 102, 2022.

[50] Xidong Wu, Yiming Zeng, Arun Das, Sumin Jo, Tinghe Zhang, Parth Patel, Jianqiu Zhang, Shou-Jiang Gao, Dexter Pratt, Yu-Chiao Chiu, et al. Regulogpt: Harnessing gpt for knowledge graph construction of molecular regulatory pathways. *bioRxiv*, 2024.

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

[52] Stefano Marchesin, Gianmaria Silvello, and Omar Alonso. Large language models and data quality for knowledge graphs, 2025.

[53] Vamsi Krishna Kommineni, Birgitta König-Ries, and Sheeba Samuel. From human experts to machines: An llm supported approach to ontology and knowledge graph construction. *arXiv preprint arXiv:2403.08345*, 2024.

[54] Chengqi Xu, Krishna C Bulusu, Heng Pan, and Olivier Elemento. Ddi-gpt: Explainable prediction of drug-drug interactions using large language models enhanced with knowledge graphs. *BioRxiv*, pages 2024–12, 2024.

[55] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.

[56] Siru Ouyang, Zhuosheng Zhang, Bing Yan, Xuan Liu, Yejin Choi, Jiawei Han, and Lianhui Qin. Structured chemistry reasoning with large language models. *arXiv preprint arXiv:2311.09656*, 2023.

[57] Di Wu, Wu Sun, Yi He, Zhong Chen, and Xin Luo. Mkg-fenn: A multimodal knowledge graph fused end-to-end neural network for accurate drug–drug interaction prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10216–10224, 2024.

[58] Hanzhu Chen, Xu Shen, Qitan Lv, Jie Wang, Xiaoqi Ni, and Jieping Ye. Sac-kg: Exploiting large language models as skilled automatic constructors for domain knowledge graphs. *arXiv preprint arXiv:2410.02811*, 2024.

[59] Antonio Di Maria, Lorenzo Bellomo, Fabrizio Billeci, Alfio Cardillo, Salvatore Alaimo, Paolo Ferragina, Alfredo Ferro, and Alfredo Pulvirenti. Netme 2.0: a web-based platform for extracting and modeling knowledge from biomedical literature as a labeled graph. *Bioinformatics*, 40(5): btae194, 2024.

[60] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. 2024.

[61] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 3, 2024.

[62] Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. Colake: Contextualized language and knowledge embedding. *arXiv preprint arXiv:2010.00309*, 2020.

[63] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*, 2020.

[64] Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. Multimodal reasoning with multimodal knowledge graph. *arXiv preprint arXiv:2406.02030*, 2024.

[65] Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 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, 2025.

[66] Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. *arXiv preprint arXiv:2305.13269*, 2023.

[67] Tong Zhou, Yubo Chen, Kang Liu, and Jun Zhao. Cogmg: Collaborative augmentation between large language model and knowledge graph. *arXiv preprint arXiv:2406.17231*, 2024.

[68] Markus J Buehler. Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design. *ACS Engineering Au*, 4(2):241–277, 2024.

[69] Daniel Truhn, Jorge S Reis-Filho, and Jakob Nikolas Kather. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nature medicine*, 29(12): 2983–2984, 2023.

[70] Henry W Sprueill, Carl Edwards, Mariefel V Olarte, Udishnu Sanyal, Heng Ji, and Sutanay Choudhury. Monte carlo thought search: Large language model querying for complex scientific reasoning in catalyst design. *arXiv preprint arXiv:2310.14420*, 2023.

[71] Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19162–19170, 2024.

[72] Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*, 2024.

[73] Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A Rossi, Subhabrata Mukherjee, Xianfeng Tang, et al. Retrieval-augmented generation with graphs (graphrag). *arXiv preprint arXiv:2501.00309*, 2024.

[74] Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. Boosting factual correctness of abstractive summarization with knowledge graph. *arXiv preprint arXiv:2003.08612*, 2020.