SciRAGBench: Benchmarking Large Language Models for Retrieval-Augmented Generation in Scientific Domains

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) enhances the performance of Large Language Models (LLMs) by integrating external knowledge, which is particularly crucial in scientific domains that demand precision and up-to-date information. However, there is currently a lack of a comprehensive framework for systematically evaluating RAG in these specialized contexts, as most existing benchmarks focus on general domains and overlook the complexities of scientific data. To address this gap, we propose SciRAGBench, the first benchmark designed to assess the RAG capabilities of LLMs in scientific contexts. It comprises ten datasets spanning diverse scientific domains, incorporating structured tables, knowledge graphs, and unstructured text as external knowledge sources. SciRAGBench systematically assesses four key competencies: Noise robustness, Negative rejection, Information integration, and Reasoning, with diverse question formats. Through extensive evaluation of state-of-the-art LLMs on SciRAGBench, we benchmark their capabilities across these four dimensions, revealing their limitations in processing various scientific data.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities across a variety of tasks, yet they often struggle with domain-specific knowledge, particularly in the scientific domain (Mann et al., 2020). While general-purpose LLMs can generate coherent and plausible answers, their inherent limitations in accessing and leveraging precise and up-to-date scientific knowledge necessitate the use of Retrieval-Augmented Generation (RAG) (Guu et al., 2020; Izacard and Grave, 2020; Khattab et al., 2021; Borgeaud et al., 2022; Ren et al., 2023; Shi et al., 2023). By integrating external corpora, RAG provides the potential to bridge the knowledge gaps of LLMs and significantly enhance the quality and reliability of generated answers (Lee et al., 2019).

Despite these advancements, RAG faces significant challenges when applied to scientific domains. Unlike general-domain corpora, scientific corpora often exhibit highly technical language, diverse data modalities, and complex interdependencies between data points (Beltagy et al., 2019). Furthermore, RAG systems are susceptible to issues such as inaccurate retrieval, noise amplification, and error propagation, which can significantly degrade the performance of LLMs (Maynez et al., 2020; Gao et al., 2023; Liu et al., 2023). These challenges are particularly critical in scientific fields, where even small inaccuracies can lead to significant misinterpretations and diminished credibility of the outputs.

Existing RAG benchmarks (Lyu et al., 2024; Saad-Falcon et al., 2023; Es et al., 2023; Gao et al., 2023; Trivedi et al., 2022; Li et al., 2022) are primarily tailored to general domains, making them ill-suited for evaluating and improving the performance of LLMs in scientific contexts. They also fail to reflect the heterogeneous nature of scientific data, which spans across structured tables, knowledge graphs, and textual descriptions. This gap highlights the urgent need for a dedicated benchmarking framework that can rigorously assess the RAG capabilities of LLMs in handling scientific corpora.

To address this need, we propose **SciRAGBench**, the first comprehensive benchmarking suite specifically designed to evaluate the RAG performance of LLMs in scientific domains. As illustrated in Figure 1, SciRAGBench incorporates ten curated datasets from a wide range of scientific fields, including biology, chemistry, physics, biomedicine, and materials science. These datasets are sourced from diverse data types, such as plain text, tables, and knowledge graphs, ensuring comprehensive coverage of the data modalities encoun-

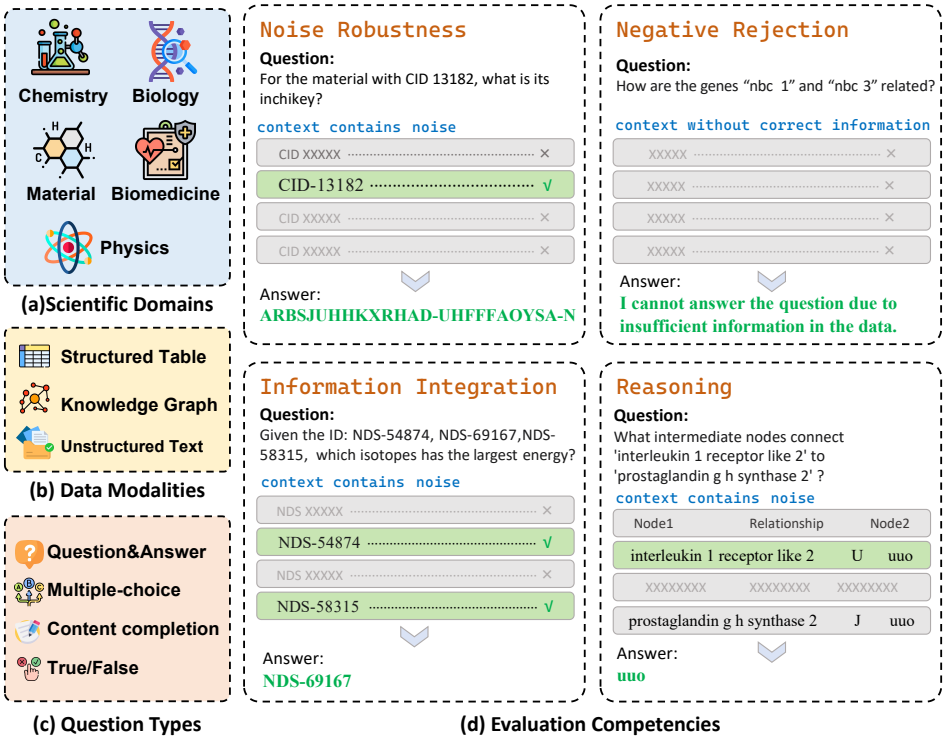

Figure 1: Overview of SciRAGBench, illustrating its data sources, modalities, question types, and evaluation competencies. Our benchmark covers multiple scientific domains, including chemistry, biology, materials science, biomedicine, and physics. It incorporates structured tables, knowledge graphs, and unstructured text as external knowledge sources. SciRAGBench supports diverse question types such as Q&A, multiple-choice, content completion, and true/false validation. Finally, it assesses four key competencies in RAG: noise robustness, negative rejection, information integration, and reasoning.

tered in scientific research. SciRAGBench evaluates four key dimensions of RAG performance: *Noise Robustness* (handling noisy or irrelevant information), *Negative Rejection* (rejecting incorrect information), *Information Integration* (synthesizing data from multiple sources), and *Reasoning* (answering queries through thinking and reasoning). Moreover, SciRAGBench supports a variety of question types, including Q&A, multiple-choice, content completion, and true/false validation, providing a holistic evaluation of LLM capabilities in scientific tasks.

The contributions of this paper can be summarized as follows:

- **Establishing the first scientific RAG benchmark:** We introduce the first benchmark specifically designed to evaluate RAG capabilities of LLMs in scientific domains, serving as a standardized evaluation suite for assessing the performance of RAG-based LLMs.
- **Developing a diverse set of domain-specific RAG datasets:** We construct ten RAG datasets spanning multiple disciplines, encompassing various data modalities and a wide range of question types to ensure comprehensive evaluation.
- **Comprehensive evaluation and analysis of LLMs:** We systematically evaluate and analyze the performance of various state-of-the-art LLMs on SciRAGBench, highlighting their strengths and limitations, and offering insights for improvement.

2 Related Works

Retrieval-Augmented Generation Large language models (LLMs) sometimes generate hallucinations (Cao et al., 2020; Raunak et al., 2021; Ji et al., 2023), producing responses that appear plausible but are ultimately factually incorrect (Xiong et al., 2024), and they also face challenges with outdated knowledge (He et al., 2022). To address these issues, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) was introduced to incorporate external, retrieved knowledge, enhancing the performance of LLMs on knowledge-intensive tasks

(Shuster et al., 2021; Huo et al., 2023). Furthermore, RAG strengthens LLMs' transparency by firmly establishing their arguments in the documents that were obtained (Mialon et al., 2023; Xiong et al., 2024). However, in real-world applications, the RAG process faces challenges such as inaccurate retrieval and noise, which can negatively impact the quality of the model's output. In this study, we systematically assess the performance of retrieved-augmented generation in LLMs, particularly in the field of science discovery.

Retrieval-Augmented Generation Benchmarks The development of robust and comprehensive benchmarks for evaluating RAG systems has gained increasing attention in recent research (Chen et al., 2024; Friel et al., 2024; Lyu et al., 2024; Saad-Falcon et al., 2023; Wang et al., 2024; Xiong et al., 2024; Yang et al., 2024; Pipitone and Alami, 2024). For instance, ARES (Saad-Falcon et al., 2023) presents a framework using LLMs and statistical methods to efficiently evaluate RAG systems, minimizing human annotations while ensuring robust assessment. CRUD-RAG (Lyu et al., 2024) establishes a Chinese benchmark for evaluating RAG systems across CRUD operations, utilizing large datasets and multi-dimensional metrics for performance evaluation. Beyond general-purpose RAG evaluation, domain-specific benchmarks have also emerged. OmniEval (Wang et al., 2024) offers an automated benchmark for evaluating RAG systems in the banking industry, concentrating on domain-specific retrieval and generation difficulties. MIRAGE (Xiong et al., 2024) assesses medical RAG systems utilizing a dataset of 7,663 questions from medical QA sources and the MEDRAG toolkit, which integrates several corpora, retrievers, and LLMs for performance measurement. However, assessing the RAG performance of LLMs in scientific domains (e.g., biology, chemistry, and physics) remains underexplored. To fill this gap, this work introduces a systematic benchmark for scientific RAG evaluation.

3 Datasets

This section presents the dataset construction process in SciRAGBench, which involves formulating evaluation competencies, collecting scientific data, generating a wide range of questions, and conducting rigorous human verification.

3.1 Evaluation Competencies

Inspired by the RAG ability definition in (Chen et al., 2024), we formulate four capabilities essential for evaluating RAG-based LLMs in scientific contexts:

- **Noise Robustness**: LLMs must effectively distinguish between relevant information and extraneous noise in retrieved scientific data. In real-world scenarios, retrievers often introduce contextually related but non-essential data. Robust models should filter out such noise to ensure correct responses.

- **Negative Rejection**: The ability to abstain from responding when all retrieved data is irrelevant or unreliable. Scientific queries often require precise evidence, and when no valid information is retrieved, LLMs should refrain from generating speculative or hallucinated responses.

- **Information Integration**: Scientific queries often require synthesizing data from multiple sources. LLMs must aggregate and compare information across different retrieved entries to generate precise and contextually grounded answers.

- **Reasoning**: The capability to perform logical inference based on retrieved data. Since retrieved information in scientific domains may be fragmented or incomplete, LLMs need to analyze relationships, deduce implicit knowledge, and generate accurate answers.

These four competencies form the foundation of SciRAGBench and provide a systematic evaluation framework for RAG performance in scientific domains.

3.2 Source Data Collection

Scientific Domains To comprehensively evaluate RAG in scientific contexts, we curate data from diverse scientific domains, including Biology, Chemistry, Physics, Biomedicine, and Materials Science. These disciplines are fundamental to modern science, encompassing a wide range of knowledge from theoretical principles to experimental data, ensuring a broad and representative assessment of RAG capabilities in scientific applications.

Data Modalities To support a broad evaluation of RAG capabilities, we consider three distinct data modalities: (1) *Unstructured Text*, (2) *Structured Tables*, and (3) *Semi-structured Knowledge Graphs* (KGs). Each modality presents unique challenges, enabling a holistic assessment of LLMs' retrieval, synthesis, reasoning, and integration capabilities in scientific domains.

Specifically, unstructured text corpora consist of scientific literature, allowing LLMs to retrieve, synthesize, and infer domain knowledge from textual sources. We collect thousands of recent research papers and experimental protocols from open-access repositories such as arXiv[1]. Structured tables contain numerical and categorical data, testing LLMs' capacity to interpret structured knowledge, recognize contextual dependencies, and perform quantitative reasoning. We collect nuclear data from IAEA[2], material properties from Material Project[3], and molecular and protein properties from PubChem[4]. Knowledge graphs encode scientific knowledge as interconnected entities and relational networks, enabling the assessment of LLMs' abilities in relational reasoning, hierarchical knowledge traversal, and cross-domain knowledge synthesis. We collect well-established scientific KGs, including Gene Ontology[5] for gene-function relationships, HIPPIE (Alanis-Lobato et al., 2016) for protein-protein interactions, PharmKG (Zheng et al., 2021) for drug-target interactions, and PrimeKG (Chandak et al., 2023) for clinical entity relationships.

3.3 Test Data Generation

Building on the collected source data, we construct corresponding datasets tailored to assess the proposed four RAG competencies outlined in Sec. 3.1. Our data generation pipeline involves (1) question generation, (2) noise injection, and (3) answer verification.

Question Generation To generate high-quality evaluation questions and corresponding answers, we employ cutting-edge LLMs (e.g., GPT-4o) to automate question generation, incorporating manual prompt engineering to ensure the questions are reasonable and aligned with the required competencies. The detailed prompts are provided in Appendix B. We define distinct question generation strategies for each of the four RAG competencies. Specifically, the noise robustness questions are generated based on isolated database entries to assess whether LLMs can extract relevant information while filtering out irrelevant or misleading noise. The negative rejection questions are achieved by removing correct contextual information from previously generated noise robustness questions, creating scenarios where the retrieved data is entirely irrelevant. The information integration questions are designed to require the synthesis of multiple database entries, focusing on comparative or statistical analysis. The reasoning questions are constructed by retrieving one or more data entries that contain implicit relationships, challenging LLMs to perform thinking beyond direct text extraction. Moreover, these generated questions span various formats, including Q&A, multiple-choice, content completion, and true/false validation, offering a robust assessment of RAG abilities.

Noise Injection Following question generation, we extract relevant contextual information from the source data. To simulate real-world retrieval challenges, we strategically inject noise into the context. Noise is sampled from semantically similar but unrelated database entries, ensuring that the benchmark reflects the imperfections of retrieval systems.

Answer Verification To maintain the rigor of the benchmark, all generated answers must be either explicitly extractable or logically deducible from the provided contexts. We implement a two-stage verification process to ensure data quality: (1) *LLM as a Judge*, where generated answers are validated against source data using LLMs as evaluators, and (2) *Human Expert Validation*, where all data are manually reviewed by domain experts to confirm factual correctness and alignment with competency requirements. Detailed information about data quality verification can be found in Appendix C.

3.4 The SciRAGBench Dataset

Based on the data collection, construction, and quality control processes described above, we construct the SciRAGBench dataset, encompassing ten distinct sub-datasets (two unstructured text datasets, four structured table datasets, and four knowledge graph datasets), covering diverse scientific fields. Each sub-dataset contains approx-

[1] https://arxiv.org
[2] https://www-nds.iaea.org
[3] https://next-gen.materialsproject.org
[4] https://pubchem.ncbi.nlm.nih.gov
[5] https://geneontology.org

imately a thousand high-quality questions, leading to a total of 11,343 questions across the entire benchmark. An overview of the dataset composition is presented in Table 1, summarizing the scientific data source, modality, and question distribution for each sub-dataset. Additionally, representative question examples for each RAG competency are provided in Appendix D.

4 Experiments

In this section, we evaluate the performance of various LLMs on SciRAGBench, and provide a thorough analysis that summarizes these LLMs' capabilities in leveraging external knowledge in scientific domains.

4.1 Experimental Setup

Models We select 14 advanced LLMs, including 3 proprietary models (GPT-4o (OpenAI et al., 2024), Claude-3.5-Sonnet (Anthropic, 2024), GPT-4o-mini (OpenAI et al., 2024)), 7 open-source general-purpose models (Deepseek-V3 (DeepSeek-AI et al., 2024), Qwen2.5-7B-Instruct (Qwen et al., 2025), Llama3.1-8B-Instruct (Dubey et al., 2024), Llama3.1-70B-Instruct (Dubey et al., 2024), Ministral-8B-Instruct (Jiang et al., 2023), GLM4-9B-Chat (GLM et al., 2024)), Gemma2-9B-it (Team et al., 2024), 4 open-source scientific models (SciGLM-6B (Zhang et al., 2024a), LlaSMol-Mistral-7B (Yu et al., 2024), ChemLLM-7B-Chat (Zhang et al., 2024b), ChemDFM-v1.5-8B (Zhao et al., 2024)). The proprietary models are accessed via their official APIs. The open-source models are deployed on a server equipped with two NVIDIA GeForce RTX 3090 GPUs. Detailed information about the models is provided in Appendix F.

Settings To ensure a fair evaluation across all models, we adopt a unified prompting template that standardizes input formatting. Specifically, each input consists of a system prompt that specifies the question type and defines answer format requirements, retrieved contexts potentially relevant to the question, and a question designed to assess one of the four core competencies in SciRAGBench. Notably, noise robustness is a fundamental prerequisite for effective RAG performance, and all other competencies involve some level of noise. In our datasets, we introduce 200–300 noise entries for structured tables and KGs and 5 noise entries for unstructured text data. This design ensures that models are evaluated under realistic retrieval conditions, where extraneous or misleading information must be filtered to generate accurate and reliable responses.

Evaluation Criteria Given that each question in SciRAGBench has a deterministic answer, we adopt accuracy as the evaluation metric for Q&A, multiple-choice, content completion, and true/false validation tasks across the categories of Noise Robustness, information integration, and reasoning. For negative rejection, we use the rejection rate as the evaluation metric. All evaluations are conducted automatically, with manual spot checks performed to ensure the correctness and reliability of the assessments.

4.2 Overall Results on SciRAGBench

Table 2, 3, and 4 show the RAG performance of LLMs on SciRAGBench across ten datasets, four competencies, and three modalities, respectively.

Model Scale Correlates with RAG Capability. The closed-source model GPT-4o achieves the highest overall score on SciRAGBench, and the open-source model Deepseek-V3 attains competitive performance, approaching that of proprietary counterparts. Our results reveal a strong positive correlation between model size and RAG effectiveness. Large-scale models (e.g., GPT-4o, Deepseek-V3, and Llama3.1-70B-it) consistently outperform their smaller counterparts (e.g., GPT-4o-mini, Llama3.1-8B-it, and Ministral-8B-it) across all domains. This phenomenon suggests that larger models exhibit superior capability in dynamically integrating external knowledge through RAG frameworks.

Scientific LLMs underperform in RAG Domain-specific scientific LLMs (e.g., ChemDFM-v1.5-8B and SciGLM-6B) demonstrate suboptimal performance, with average scores 18.7-39.9% lower than general-domain models. We attribute this to two key limitations: (1) Context Processing Deficiency: When processing lengthy or complex contexts, scientific LLMs show significant limitations in managing and integrating multiple retrieved knowledge chunks. (2) Structural Output Constraints: Scientific models have notable difficulties in generating well-formatted structured outputs, resulting in higher error rates in producing standardized responses compared to general models. Case studies of the model's responses and error analysis are provided in Appendix G.

Table 1: Statistic of our SciRAGBench dataset, which comprises ten sub-datasets derived from diverse scientific data. The detailed data sources are listed in Appendix E.

Sub-dataset	Domain	Source	Modality	# Noise Rob.	# Negative Rej.	# Info. Int.	# Reasoning	# Total
MatText	Materials	arXiv	Text	216	146	222	356	940
BioText	Biology	Biorxiv	Text	236	97	318	317	968
MatTab	Materials	Material Project	Table	299	150	287	200	936
IaeaTab	Physics	IAEA	Table	442	222	286	180	1130
ProtTab	Biology	Pubchem	Table	496	249	327	180	1305
MolTab	Chemistry	Pubchem	Table	516	259	350	180	1305
GoKG	Biology	Gene Ontology	KG	507	254	239	180	1180
HipKG	Biology	HIPPLE	KG	470	236	319	140	1165
PhaKG	Biomedicine	PharmKG	KG	512	256	281	168	1217
PriKG	Biomedicine	PrimeKG	KG	410	205	382	253	1250

Table 2: Performance of LLMs across ten datasets on SciRAGBench. Underline results indicate the best results among all models. **Bold results** indicate the best results in each category.

Models	MatTab	IaeaTab	MolTab	ProtTab	PhaKG	PriKG	HipKG	GoKG	BioText	MatText	Overall
GPT-4o	**68.79**	**56.55**	55.79	**52.64**	**55.71**	**54.80**	68.50	74.32	79.03	64.57	**61.52**
GPT-4o-mini	40.71	38.85	46.67	44.57	40.59	52.64	65.20	73.14	**79.24**	**65.00**	54.57
Claude-3.5-Sonnet	48.48	42.03	**67.91**	52.22	50.94	45.96	**75.78**	**84.07**	58.06	61.49	59.20
Deepseek-V3	**56.62**	**54.07**	**59.85**	**52.08**	**52.18**	**51.92**	**63.42**	**72.29**	66.74	45.31	**57.50**
Llama3.1-70B-it	38.25	39.73	44.44	41.29	44.70	44.00	59.31	70.17	66.53	51.91	49.80
Qwen2.5-7B-it	28.10	32.65	43.30	39.46	36.15	45.60	53.99	62.46	**68.18**	**59.68**	46.62
GLM4-9B-Chat	31.41	25.84	47.82	43.45	36.03	44.56	57.94	60.51	67.77	50.96	46.46
Llama3.1-8B-it	28.85	34.34	42.76	39.78	38.29	46.56	52.62	59.32	64.26	49.36	45.50
Gemma2-9B-it	32.91	32.21	42.91	37.22	37.39	50.48	56.57	57.29	37.77	29.67	42.21
Ministral-8B-it	23.08	19.12	35.56	37.38	22.76	37.92	48.51	52.88	48.14	45.32	37.58
ChemDFM-v1.5-8B	**33.65**	**31.15**	**35.56**	**36.82**	**40.43**	**30.72**	**49.70**	**56.44**	26.11	18.91	**36.80**
SciGLM-6B	11.86	11.50	17.70	14.94	19.56	20.88	21.46	28.31	**44.17**	**31.35**	21.58
LlaSMol-Mistral-7B	13.35	12.83	16.55	14.70	21.54	19.84	22.83	29.92	33.13	20.98	20.42
ChemLLM-7B-chat	3.42	6.02	8.81	8.15	13.45	5.92	5.15	15.51	39.94	22.67	12.16

Table 3: Performance of LLMs across four competencies on SciRAGBench.

Models	Noise Rob.	Negative Rej.	Info. Int.	Rea.	Overall
GPT-4o	**89.72**	19.51	**54.90**	**65.97**	**61.52**
GPT-4o-mini	77.81	14.71	47.54	57.68	54.57
Claude-3.5-Sonnet	82.95	**49.10**	50.85	47.29	59.20
Deepseek-V3	**90.80**	6.05	**49.80**	**60.53**	**57.50**
Llama3.1-70B-it	81.05	6.87	45.44	47.76	49.80
Qwen2.5-7B-it	69.92	**9.02**	42.95	50.34	46.62
GLM4-9B	71.48	2.53	50.78	43.82	46.46
Llama3.1-8B-it	75.34	5.88	41.47	39.99	45.50
Gemma2-9B-it	66.97	2.38	28.74	48.22	42.20
Ministral-8B-it	56.80	4.76	31.32	39.62	37.58
ChemDFM-v1.5-8B	**45.49**	**19.31**	**22.23**	**46.40**	**36.80**
SciGLM-6B	33.24	9.01	18.00	29.48	21.58
LlaSMol-Mistral-7B	31.96	6.83	14.63	26.59	20.42
ChemLLM-7B-Chat	20.29	4.09	16.85	7.57	12.16

RAG significantly enhances model performance. Table 6 (in Appendix A) presents a comparison between direct answering and answering with RAG on SciRAGBench. The results clearly demonstrate that the integration of RAG consistently enhances performance. This underscores the efficacy of RAG, especially for tasks requiring specialized knowledge, like those in scientific domains, where accurate results depend on access to and integration of external information.

4.3 Results of Four Competencies

Noise Robustness As shown in Table 3, larger models demonstrate superior performance in handling noisy context, indicating that smaller models are more sensitive to the quality of input data and struggle with filtering irrelevant or misleading information. In contrast, scientific LLMs significantly underperform general models in this competency. The best-performing scientific model, ChemDFM-v1.5-8B, achieves only 45.49, while other domain-specific models perform even worse. This highlights a challenge: while scientific models are trained on domain-specific data, they may not have been optimized for retrieval-based noise filtering, making them less effective in distinguishing relevant from irrelevant information.

Negative Rejection This measures the ability of a model to refrain from answering when provided with entirely irrelevant information. Claude-3.5-Sonnet significantly outperforms all other models, demonstrating a strong tendency to reject unreliable or misleading retrieved content. Notably, all general-purpose models struggle in this aspect. This highlights the risk of "overconfidence" in current models, which may pose potential safety risks in the scientific domain.

Information Integration This assesses a model's capability to synthesize and aggregate knowledge from multiple retrieved sources. GPT-4o achieves the best performance, followed by DeepSeek-V3 and Claude-3.5-Sonnet, indicating that these models are better equipped to combine multiple data points into coherent, accurate answers. Among general open-source models, GLM4-9B shows a competitive score, even surpassing DeepSeek-V3 in this category. However, scientific models significantly lag behind, indicating that domain-specific models, while specialized in scientific text, struggle with multi-source retrieval synthesis, which is critical for real-world scientific applications.

Reasoning This competency reflects a model's ability to infer and deduce answers beyond direct retrieval. GPT-4o and DeepSeek-V3 exhibit a strong ability to perform logical inference using retrieved knowledge. Interestingly, GPT-4o-mini also performs well in reasoning, despite weaker performance in other competencies. In contrast, scientific models show limited reasoning ability, with ChemDFM-v1.5-8B performing the best, but others like ChemLLM-7B-Chat and LlaSMol-Mistral-7B scoring significantly lower. This indicates that while domain models may specialize in factual scientific knowledge, they do not exhibit strong reasoning capabilities when required to infer or synthesize information.

4.4 Results of Three Modalities

Table 4 reports the performance of LLMs across three modalities on SciRAGBench, highlighting notable differences in how models handle different knowledge representations.

Unstructured Text Unstructured text represents free-form scientific literature and articles, requiring models to extract, synthesize, and infer knowledge from raw textual content. GPT-4o and GPT-4o-mini achieve relatively high performance in this experiment. Qwen2.5-7B-it also performs competitively among open-source models, surpassing DeepSeek-V3 and Claude-3.5-Sonnet. For scientific models, SciGLM-6B exhibits superior performance compared to other domain-specific models, yet it remains notably less effective than general-purpose LLMs. This suggests that while domain-specific models are trained on specialized corpora, they may not generalize well in handling diverse textual retrieval and reasoning tasks compared to general-purpose LLMs.

Table 4: Performance of LLMs across three modalities on SciRAGBench.

Models	Text	Table	KG	Overall
GPT-4o	71.91	**55.91**	63.13	**61.52**
GPT-4o-mini	**72.22**	42.98	55.84	54.57
Claude-3.5-Sonnet	59.75	53.41	**64.99**	59.20
Deepseek-V3	56.18	**55.68**	**59.77**	**57.50**
Llama3.1-70B-it	59.33	41.19	54.30	49.80
Qwen2.5-7B-it	**69.93**	36.58	49.38	46.24
GLM4-9B-Chat	59.49	37.94	49.46	46.46
Llama3.1-8B-it	56.92	37.08	49.06	45.50
Gemma2-9B-it	34.27	36.73	50.23	42.21
Ministral-8B-it	46.75	29.50	41.24	37.58
ChemDFM-v1.5-8B	22.92	**34.44**	**44.05**	**36.80**
SciGLM-6B	**38.48**	14.25	22.51	21.58
LlaSMol-Mistral-7B	27.74	14.49	23.45	20.42
ChemLLM-7B-chat	32.28	6.86	10.01	12.16

Structured Table Processing structured tabular data requires models to interpret numerical and categorical relationships, extract relevant information, and perform reasoning over structured scientific datasets. Overall, most models show inferior performance in this modality compared to their performance on unstructured text. Specifically, GPT-4o achieves the best performance, followed closely by DeepSeek-V3. In contrast, open-source models generally struggle with tabular data. Notably, scientific LLMs show severe limitations in this modality, with ChemDFM-v1.5-8B being the best among them, but still far behind general-purpose models, indicating that current domain-specific models lack the necessary adaptations to effectively process structured scientific tables.

Knowledge Graph KGs require models to understand relational structures, traverse multi-hop connections, and integrate information across entity-relation networks. Claude-3.5-Sonnet achieves the highest performance, followed by

GPT-4o. DeepSeek-V3 also performs strongly, demonstrating robust knowledge synthesis capabilities. Among open-source models, Llama3.1-70B-it achieves the best performance, while Ministral-8B-it lags behind, indicating that knowledge graph reasoning remains a challenging task for smaller models. Scientific models perform better in this modality compared to their performance on structured tables. ChemDFM-v1.5-8B leads the scientific models, but still falls behind most general-purpose models, highlighting the need for improved adaptation to structured relational reasoning in scientific contexts.

4.5 Further Discussions

Our experimental results highlight three key discrepancies in the performance of LLMs on scientific RAG tasks, underscoring fundamental challenges that require further advancements.

Modality Discrepancy LLMs exhibit relatively better performance on unstructured text compared to structured tables and KGs. This suggests that existing models rely heavily on linguistic patterns and semantic context rather than structured reasoning and multi-modal data integration. The weaker performance on tables and KGs indicates a bottleneck in structured data comprehension, where models struggle to extract, synthesize, and infer information effectively from unstructured data. To bridge this gap, models need improved cross-modal alignment, integrating structured data reasoning into their training paradigm. Techniques such as joint pretraining on text, tables, and graphs could enhance structured data understanding.

Competency Discrepancy The results reveal uneven performance across the four core RAG competencies. While top-performing models demonstrate relatively strong noise filtering and reasoning abilities, they struggle with negative rejection—the ability to abstain from answering when faced with unreliable or insufficient evidence. This suggests that models prioritize generating responses over ensuring accuracy, increasing the risk of hallucinations in scientific applications where factual correctness is critical. To address this, models should incorporate uncertainty quantification techniques, such as confidence-based rejection mechanisms and calibrated probability outputs, to enhance their ability to detect and reject misleading retrievals. Furthermore, reinforcement learning with human feedback and verification-based prompting strategies could help improve the model's reliability in rejecting incorrect information.

Specialized vs. General Model Discrepancy While scientific LLMs are designed to excel in domain-specific tasks, our results indicate that their performance in RAG-based tasks does not consistently surpass that of general-purpose models. They lack the retrieval and reasoning optimizations necessary for effective RAG, which limits their ability to integrate external knowledge efficiently. To enhance the RAG performance of specialized scientific models, methods such as fine-tuning models with retrieved scientific evidence, and domain-aware prompt engineering can enable scientific LLMs to balance specialization with flexibility, ensuring they remain effective in various scientific scenarios.

5 Conclusion

In this work, we introduced SciRAGBench, a comprehensive benchmark for evaluating retrieval-augmented generation capabilities in large language models within scientific domains. SciRAGBench encompasses multiple data modalities (structured tables, knowledge graphs, and unstructured text), spanning diverse scientific disciplines. By systematically assessing four key competencies (Noise Robustness, Negative Rejection, Information Integration, and Reasoning), we provide a rigorous framework for understanding how well LLMs leverage external knowledge in science-intensive tasks. Our experimental results demonstrate that while RAG enhances LLM performance in scientific contexts, existing models exhibit notable limitations in effectively utilizing retrieved information. The primary challenge lies in the inherent complexity of scientific data, particularly structured formats such as tables and knowledge graphs, which demand high specialization, precise contextual understanding, and the ability to synthesize fragmented and implicitly related information. Even state-of-the-art LLMs struggle to filter noise, reject unreliable sources, and integrate multi-source evidence, indicating that significant advancements are required to improve their RAG capabilities in scientific applications. Moving forward, we envision SciRAGBench as a foundation for guiding future improvements in LLMs, driving the development of more reliable and knowledge-grounded RAG systems for science discovery.

Limitations

While SciRAGBench provides a comprehensive evaluation framework for assessing RAG capabilities in scientific domains, it has certain limitations. First, all included datasets are text-based, excluding non-textual modalities such as images and 3D structures, which are crucial in many scientific tasks. Incorporating more multi-modal data would provide a more holistic assessment of RAG. Second, since our primary goal is to analyze the effectiveness of LLMs in processing retrieved scientific information, the quality of retrieval itself is not considered in this benchmark. However, retrieval quality significantly impacts RAG performance, and future work will explore end-to-end RAG evaluation, incorporating both retrieval and generation processes.

References

Gregorio Alanis-Lobato, Miguel A Andrade-Navarro, and Martin H Schaefer. 2016. Hippie v2. 0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic acids research*, page gkw985.

AI Anthropic. 2024. The Claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. *arXiv preprint arXiv:2010.08712*.

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

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

DeepSeek-AI, Aixin Liu, Bei Feng, and Bing Xue et al. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv:2407.21783*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Team GLM, Aohan Zeng, and Bin Xu et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Siqing Huo, Negar Arabzadeh, and Charles LA Clarke. 2023. Retrieving supporting evidence for llms generated answers. *arXiv preprint arXiv:2306.13781*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv:2310.06825*.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Relevance-guided supervision for openqa with colbert. *Transactions of the association for computational linguistics*, 9:929–944.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. 2024. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systems*.

Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

OpenAI, :, Aaron Hurst, and et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*.

Qwen, An Yang, and Baosong Yang et al. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. *arXiv preprint arXiv:2104.06683*.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Gemma Team, Morgane Riviere, Shreya Pathak, and Pier Giuseppe Sessa et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2024. Omnieval: An omnidirectional and automatic rag evaluation benchmark in financial domain. *arXiv preprint arXiv:2412.13018*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, et al. 2024. Crag–comprehensive rag benchmark. *arXiv preprint arXiv:2406.04744*.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. LlaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv:2402.09391*.

Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. SciGLM: Training scientific language models with self-reflective instruction annotation and tuning. *arXiv:2401.07950*.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, et al. 2024b. ChemLLM: A chemical large language model. *arXiv:2402.06852*.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, et al. 2024. ChemDFM: Dialogue foundation model for chemistry. *arXiv:2401.14818*.

Shuangjia Zheng, Jiahua Rao, Ying Song, Jixian Zhang, Xianglu Xiao, Evandro Fei Fang, Yuedong Yang, and Zhangming Niu. 2021. Pharmkg: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in bioinformatics*, 22(4):bbaa344.

Appendix

A More Results on SciRAGBench

Table 5 presents the more detailed results of SciRAGBench, including four competencies in the ten datasets. Table 6 shows the performance comparison between direct answering and answering with RAG.

B Prompts for Constructing the Dataset

We present distinct prompt templates for each of the four RAG capabilities below.

- **A prompt for generating questions about noise robustness**

> **System Message:**
> You're a brilliant in scientific domain.
> **User Message:**
> You will be provided with several triples from PriKG that form a path connecting a starting point to an endpoint. Based on this path, you need to generate a scientific question designed to test the respondent's ability to find the correct answer in the noise, with information from the knowledge graph. The question types can be Q&A or fill-in-the-blank. The answers to QA questions should be simple, concise, and easily verifiable phrases, not long sentences.
>
> Start Node: start_node
> End Node: end_node
> Path: data['path']
>
> Triples:
> data['triplets']
>
> Please generate a scientific question based on this information. Ensure that the question requires the respondent to find the correct answer in the noise in the knowledge graph and the difficulty level should be moderate. Please output the question in JSON format only. Do not output anything other than the JSON format. The JSON format should look like this:
>
> {
> "question_type": "[Here is the question type]",
> "question": "[Here is the question, or directly reject if unable to generate]",
> "answer": "[Here is the answer to the question]"
> }
> Next is the triples you need to use: {Triples}

- **A prompt for generating questions about negative rejection**

> **System Message:**
> You're a brilliant in scientific domain.
> **User Message:**
> You will be provided with a noise robustness question and its corresponding correct context, also context. Your task is to remove the correct contextual information from the context.
>
> Do not alter the form of the question. Output the question in JSON format only, without any additional text. The JSON format should adhere to the following structure:
> {
> "question": "[Here is the question, or directly reject if unable to generate]",
> "answer": "[Here is the answer to the question]"
> }
> Next is the context you need to use: {Contexts}

- **A prompt for generating questions about information integration**

> **System Message:**
> You're a brilliant in scientific domain.
> **User Message:**
> You will be provided with several data entries describing various properties of different materials. Based on these properties, you need to generate a scientific question that tests the respondent's ability to retrieve, integrate, and

Table 5: Performance of LLMs across four competencies in ten datasets of SciRAGBench

Models	MatTab					IaeaTab				
	Noise	Negative	Info.	Reasoning	All	Noise	Negative	Info	Reasoning	All
GPT-4o	88.16	58.00	29.97	64.00	68.79	81.00	26.13	44.76	52.78	56.55
GPT-4o-mini	71.91	39.33	10.10	39.00	40.71	57.01	5.86	37.06	37.78	38.85
Claude-3.5-Sonnet	77.21	56.46	19.57	41.00	48.48	56.31	23.29	37.68	37.78	42.03
Deepseek-V3	93.31	28.00	24.74	69.00	56.62	79.86	5.41	44.76	65.56	54.07
Llama3.1-70B	69.57	26.00	6.97	45.50	38.25	62.44	6.31	30.77	39.44	39.73
Qwen2.5-7B-instruct	48.83	9.33	3.14	47.00	28.10	45.70	8.11	32.87	30.56	32.65
GLM4-9B-Chat	52.84	2.67	18.82	39.00	31.41	29.64	0.45	40.56	24.44	25.84
Llama3.1-8B	58.53	4.67	4.88	37.00	28.85	53.85	6.31	25.52	34.44	34.34
Gemma2-9B-it	65.55	2.00	6.97	44.50	32.91	50.23	2.70	16.78	48.89	32.21
Ministral-8B-it	44.15	2.00	3.14	36.00	23.08	18.78	9.01	9.79	47.22	19.12
ChemDFM-v1.5-8B	42.81	10.67	9.41	72.00	33.64	36.65	14.41	21.33	53.89	31.15
SciGLM-6B	5.69	1.33	2.79	42.00	11.86	11.99	0.90	3.85	35.56	11.50
LlaSMol-Mistral-7B	7.69	5.33	2.79	43.00	13.35	16.29	6.31	1.75	30.00	12.83
ChemLLM-7B-Chat	1.67	2.67	0.00	11.50	3.42	9.28	0.45	4.90	6.67	6.02

Models	MolTab					ProtTab				
	Noise	Negative	Info.	Reasoning	All	Noise	Negative	Info.	Reasoning	All
GPT-4o	91.09	9.27	30.29	71.11	55.79	90.52	14.46	18.96	62.22	52.64
GPT-4o-mini	68.99	13.13	32.29	58.89	46.67	72.58	9.24	19.27	62.22	44.57
Claude-3.5-Sonnet	92.84	58.59	42.53	60.56	67.91	77.08	31.43	20.12	70.00	52.22
Deepseek-V3	94.38	5.79	39.71	77.78	59.85	96.57	6.02	22.94	46.11	52.08
Llama3.1-70B	71.90	1.93	30.29	54.44	44.44	79.44	2.81	14.07	38.89	41.29
Qwen2.5-7B-instruct	57.75	5.41	37.71	68.33	43.30	61.69	9.64	19.88	55.00	39.46
GLM4-9B	66.86	0.39	50.29	56.67	47.82	60.89	2.01	44.34	51.11	43.45
Llama3.1-8B-instruct	70.74	0.00	31.14	46.67	42.76	67.74	2.41	21.10	48.33	39.78
Gemma2-9B-it	63.57	0.39	36.86	56.67	42.91	62.70	0.80	18.65	51.11	37.22
Ministral-8B-it	51.55	1.16	31.71	46.67	35.56	59.07	1.20	23.24	53.33	37.38
ChemDFM-v1.5-8B	38.95	33.20	22.29	55.00	35.56	38.91	34.94	19.27	65.56	36.82
SciGLM-6B	23.26	2.32	8.86	41.11	17.70	19.76	3.21	6.42	33.33	14.94
LlaSMol-Mistral-7B	22.48	5.79	10.00	27.78	16.55	22.18	5.62	3.98	26.11	14.70
ChemLLM-7B-Chat	11.63	0.00	14.00	3.33	8.81	13.31	0.00	9.48	2.78	8.15

Models	PriKG					HipKG				
	Noise	Negative	Info.	Reasoning	All	Noise	Negative	Info.	Reasoning	All
GPT-4o	70.98	24.88	41.62	72.73	54.80	97.02	33.90	43.26	88.57	68.50
GPT-4o-mini	74.88	27.80	31.15	69.17	52.64	83.51	20.34	43.26	67.86	65.20
Claude-3.5-Sonnet	63.66	28.29	29.66	56.13	45.96	97.85	75.32	43.26	77.14	75.78
Deepseek-V3	73.90	1.95	40.58	73.91	51.92	94.47	9.32	52.66	75.00	63.42
Llama3.1-70B-it	74.88	9.27	20.68	57.31	44.00	91.06	14.41	42.95	65.71	59.31
Qwen2.5-7B-it	68.05	22.93	27.75	54.55	45.60	60.64	3.39	48.90	62.14	53.99
GLM4-9B-Chat	70.98	6.83	27.23	58.50	44.56	85.32	5.51	52.98	65.71	57.94
Llama3.1-8B-it	75.37	18.54	27.75	50.99	46.56	82.55	10.59	45.14	40.00	52.62
Gemma2-9B-it	78.78	6.83	30.37	70.36	50.48	91.49	3.81	36.99	72.86	56.57
Ministral-8B-it	55.12	13.66	23.04	52.17	37.92	66.60	1.27	31.03	47.50	48.51
ChemDFM-v1.5-8B	42.20	34.15	10.47	39.92	30.72	50.85	72.46	14.73	87.14	49.70
SciGLM-6B	42.93	4.39	4.71	22.92	20.88	33.19	1.27	5.33	52.86	21.46
LlaSMol-Mistral-7B	33.90	10.73	5.24	26.48	19.84	26.38	8.47	9.72	65.00	22.83
ChemLLM-7B-Chat	8.05	3.41	7.59	1.98	5.92	3.62	1.27	8.15	10.00	5.15

Models	GoKG					PhaKG				
	Noise	Negative	Info.	Reasoning	All	Noise	Negative	Info.	Reasoning	All
GPT-4o	91.91	11.02	87.45	96.67	74.32	88.09	16.80	45.91	32.74	55.71
GPT-4o-mini	90.34	23.23	76.57	90.56	73.14	63.87	8.20	38.08	23.21	40.5
Claude-3.5-Sonnet	91.91	82.28	70.71	82.22	84.07	88.77	28.63	64.00	14.88	50.94
Deepseek-V3	91.72	1.57	88.70	95.56	72.29	86.91	0.39	46.62	34.52	52.18
Llama3.1-70B-it	92.70	3.54	74.48	95.00	70.17	72.07	2.73	41.28	30.95	44.70
Qwen2.5-7B-it	88.17	12.60	45.61	82.78	62.46	54.30	17.19	31.32	17.86	36.15
GLM4-9B-Chat	87.57	5.12	51.46	74.44	60.51	62.89	2.34	32.38	15.50	36.05
Llama3.1-8B-it	91.32	2.76	39.75	75.00	59.32	56.84	12.89	36.65	23.21	38.29
Gemma2-9B-it	84.42	0.39	47.28	74.44	57.29	61.13	3.12	38.43	19.00	37.39
Ministral-8B-it	82.05	4.33	27.62	72.78	52.88	40.43	5.47	15.66	7.14	22.76
ChemDFM-v1.5-8B	72.58	43.31	39.75	51.67	56.44	58.01	41.80	25.27	15.00	40.43
SciGLM-6B	43.39	9.84	15.48	28.89	28.31	40.62	3.52	3.91	5.95	19.56
LlaSMol-Mistral-7B	50.69	9.45	13.39	22.22	29.92	47.85	7.03	1.07	1.50	21.54
ChemLLM-7B-Chat	19.53	14.96	10.46	11.67	15.51	30.08	3.52	0.36	2.00	13.45

Models	MatText					BioText				
	Noise	Negative	Info.	Reasoning	All	Noise	Negative	Info.	Reasoning	All
GPT-4o	99.07	0.68	97.30	49.44	64.57	99.58	0.00	97.48	69.40	79.03
GPT-4o-mini	96.76	0.00	91.44	55.90	65.00	98.31	0.00	96.23	72.24	79.24
Claude-3.5-Sonnet	99.07	54.11	96.40	19.94	61.49	84.75	52.58	84.59	13.25	58.06
Deepseek-V3	98.15	2.05	42.34	32.87	45.31	98.73	0.00	94.97	35.02	66.74
Llama3.1-70B-it	98.15	0.68	97.30	16.57	51.59	98.31	1.03	95.60	33.75	66.53
Qwen2.5-7B-it	68.05	22.93	27.75	54.55	59.68	60.64	3.39	48.90	62.14	66.18
GLM4-9B-Chat	98.61	0.00	93.24	16.57	50.96	99.15	0.00	96.54	36.28	67.77
Llama3.1-8B-it	98.15	0.68	89.64	14.61	49.36	98.31	0.00	93.08	29.65	64.26
Gemma2-9B-Chat	56.94	0.68	10.53	26.12	29.67	54.85	3.09	55.03	18.30	37.77
Ministral-8B-it	83.33	19.18	76.58	13.48	45.32	66.95	18.56	71.38	19.87	48.14
ChemDFM-v1.5-8B	34.72	6.12	24.53	13.76	18.91	39.24	16.49	35.22	10.09	26.11
ChemLLM-7B-Chat	45.83	11.56	54.72	8.43	31.35	59.92	3.09	58.81	17.35	44.17
LlaSMol-Mistral-7B	44.91	5.44	49.06	8.71	20.98	47.26	4.12	49.37	15.14	33.13
SciGLM-6B	54.63	29.25	66.04	12.92	22.67	56.96	34.02	62.58	19.24	39.94

Table 6: Comparison of Performance on SciRAGBench: Direct Answering vs. Answering with RAG. Underline results indicate the best results among all models. **Bold results** indicate the best results in each category.

Model	**MatTab**		**IaeaTab**		**MolTab**		**ProtTab**		**PhaKG**	
	Direct	RAG	Direct	RAG	Direct	RAG	Direct	RAG	Direct	RAG
GPT-4o	14.64	**68.79**	15.31	**56.55**	26.82	55.79	23.64	**52.64**	16.81	**55.71**
GPT-4o-mini	**15.38**	40.71	18.67	38.85	25.52	46.67	24.84	44.57	14.01	40.59
Claude-3.5-Sonnet	15.22	48.48	**23.45**	42.03	**32.95**	**67.91**	**31.07**	52.22	**26.62**	50.94
Deepseek-V3	**14.82**	**56.62**	**22.65**	**54.07**	**31.88**	**59.85**	**26.20**	**52.08**	15.80	**52.18**
Llama3.1-70B-it	10.04	38.25	16.19	39.73	21.30	44.44	22.60	41.29	13.15	44.70
Qwen2.5-7B-it	14.64	28.1	18.94	32.65	21.07	43.30	20.69	39.46	15.93	36.15
GLM4-9B-Chat	12.61	31.41	16.73	25.84	22.53	47.82	22.28	43.45	13.53	36.03
Llama3.1-8B-it	14.21	28.85	14.78	34.34	18.54	42.76	17.97	39.78	**16.35**	38.29
Ministral-8B-it	0.82	23.08	8.29	19.12	8.00	35.56	4.66	37.38	15.05	22.76
Gemma2-9B-it	13.25	32.91	10.27	32.21	12.87	42.91	13.74	37.22	11.45	37.39
ChemDFM-v1.5-8B	14.38	**33.65**	**13.54**	**31.15**	**26.36**	**35.55**	**28.63**	**36.82**	**44.53**	**40.43**
SciGLM-6B	12.18	11.86	10.44	11.50	15.56	17.70	13.58	14.94	13.56	19.56
LlaSMol-Mistral-7B	11.97	13.35	10.88	12.83	13.71	16.55	11.98	14.70	23.62	21.54
ChemLLM-7B-chat	**17.52**	3.42	13.45	6.02	19.16	8.81	15.73	8.15	18.01	13.45

Model	**PriKG**		**HipKG**		**GoKG**		**BioText**		**MatText**	
	Direct	RAG	Direct	RAG	Direct	RAG	Direct	RAG	Direct	RAG
GPT-4o	17.44	**54.80**	14.42	68.50	43.47	74.32	53.41	79.03	41.28	64.57
GPT-4o-mini	16.48	52.64	10.99	65.20	42.80	73.14	**55.68**	**79.24**	**48.09**	**65.00**
claude-3.5-sonnet	**26.80**	45.96	**21.55**	**75.78**	**45.59**	**84.07**	55.68	58.06	41.60	61.49
Deepsee-V3	**17.33**	**51.92**	14.76	**63.42**	**39.75**	**72.29**	**60.07**	66.74	**51.18**	45.31
Llama3.1-70B-it	14.40	44.00	**15.88**	59.31	32.12	70.17	49.80	66.53	40.21	51.91
Qwen2.5-7B-it	16.56	45.60	9.87	53.99	33.64	62.46	47.11	**68.18**	36.81	**59.68**
GLM4-9B-Chat	16.72	44.56	11.93	57.94	30.17	60.51	47.52	67.77	36.38	50.96
Llama3.1-8B-it	16.24	46.56	14.51	52.62	35.51	59.32	47.31	64.26	37.34	49.36
Gemma2-9B-it	15.36	50.48	9.96	56.57	31.27	57.29	51.81	37.77	35.88	29.67
Ministral-8B-it	15.05	37.92	13.24	48.51	28.27	52.88	41.84	48.14	32.23	45.32
ChemDFM-v1.5-8B	**33.66**	**30.72**	**30.21**	**49.70**	**39.84**	**56.44**	**50.88**	26.11	**30.83**	18.91
SciGLM-6B	15.20	20.88	18.80	21.46	25.93	28.31	33.44	**44.17**	21.63	**31.35**
LlaSMol-Mistral-7B	15.52	19.84	20.17	22.83	23.39	29.92	33.85	33.13	23.58	20.98
ChemLLM-7B-chat	16.80	5.92	23.86	5.15	27.80	15.51	45.92	39.94	30.44	22.67

analyze information from the table.
Please follow the instructions below to generate the question and answer:
1. The question should be in Q&A format, starting with sentence like "Given the following four materials: mp-xxxxx, mp-xxxxx, mp-xxxxx, mp-xxxxx" or "Which of the following materials, mp-xxxxx, mp-xxxxx, mp-xxxxx, mp-xxxxx".
2. The question should focus on a single numeric property of the materials that is representative of the material and comparable.
3. The question should involve comparing the values of this property and identifying the result.
4. The answer should be the material ID of the material with the correct value, and the answer must be one of the materials listed in the question.
Please output the question in JSON format only. Do not output anything other than the JSON format. The JSON format should look like this:
{
"question": "[Here is the question, or directly reject if unable to generate]",
"answer": "[Here is the answer to the question]"
}
Next is the data entries you need to use:

Material ID, Formula ... Sites ... Volume, Density
mp-xxxxx .....
mp-xxxxx .....
mp-xxxxx .....
mp-xxxxx .....

- **A prompt for generating questions about reasoning**

System Message:
You're a brilliant in scientific domain.
User Message:
Please write a scientific reasoning question based on the following article. Treat the paper as consisting of two parts. The first part includes the introduction, background, methods, and experimental results. The second part contains the conclusions and analysis derived from the first part. The goal of the question is to test the ability to infer the second part based on the summary of the first part, without knowing the premises of the first part. Therefore, the question should be based on the first part.
Please follow the instructions below to generate the question and answer:
1. The question should be a multiple-choice question with four options, one or more of which is correct, and the others are incorrect.
2. The difficulty level of the question is high and should involve summarizing, generalizing, and reasoning, rather than simple information retrieval or verification. The question should require at least a university-level education to answer.
3. The answer to the question should not be directly available from the first part paragraphs. It should not be directly deducible but should require complex reasoning to arrive at the correct answer.
4. Incorrect options should contain errors or deviations from the original content. The incorrect options should sound reasonable, but the content must be wrong.
5. If you feel you cannot generate a question or are uncertain about the correctness of the question, please output "[Unable to generate question]".
6. The question should be very difficult. If you feel you cannot provide a high-difficulty question, please output "[Unable to generate question]".
Please output the question in JSON format only. Do not output anything other than the JSON format. The JSON format should look like this:
{ "question": "[Here is the question, or directly reject if unable to generate]",
"options": {
"A": "[Option A]",

"B": "[Option B]",
"C": "[Option C]",
"D": "[Option D]"
},
"answer": "[A or B or C or D]"
}
Next is the full text of the article:
{Papers}

C Data Quality Verification

LLM as a Judge: We use advanced LLMs (e.g., GPT-4o) as automated evaluators to verify that each generated answer is both extractable and logically deducible from the relevant context, ensuring factual consistency and relevance. The prompt is presented below.

System Message:
You're a highly capable evaluator in scientific domain.
User Message:
Below is a question, its relevant context, and an answer. Your task is to verify whether the answer meets the following standard:
1. The answer must be explicitly extractable or logically deducible from the provided context.
2. The answer must adhere strictly to the relevant information in the context and be factually correct.
3. If the answer meets the standard, output "Yes". If it does not meet the standard, output "No".
[Relevant Context start]
{Context}
[Relevant Context end]

[Question start]
{Question}
[Question end]

[Answer start]
{Answer}
[Answer end]
Please evaluate and output either "Yes" or "No" based on the above criteria.

Human Expert Evaluation: To further ensure the quality and accuracy of the generated data, we subjected the data that passed the initial LLM validation to manual review by two domain experts. These experts were tasked with thoroughly evaluating each instance based on the following three criteria: (1) whether the designated competency aligns with the actual capability tested by the question, (2) the clarity and logical consistency of the question's semantics, and (3) whether the answer is fully contained within the provided context and factually accurate based on that context. Each evaluation was scored on a scale from 0 to 2: 0 indicated a faulty answer that required removal from the dataset, 1 denoted an answer with partial validity that needed manual corrections, and 2 signified high-quality, correct responses. A total score of 5 or more was considered as high-quality. After the experts reviewed all instances, the results revealed that 90.83% of the instances met the required high quality standards.

D Dataset Examples

In this part, we demonstrate several examples of questions aligned with four core competencies. For each competency, we present three examples corresponding to three distinct data modalities.

Noise Robustness

- **Unstructured Text** We randomly sampled 5 articles to form a noisy context and selected one article as the correct context. Then, we generated questions based on the correct context to evaluate LLMs' noise robustness ability on text. Below is an example we generated from BioText.

Example of Unstructured Text Question

System Message:
Please answer the scientific questions based on the content. Your answer only needs to include the one or more correct option labels, not the full options. You should give your answer directly without any other characters.

User Message:
What is the primary objective of the statistical framework proposed in the paper 'Augmented Doubly Robust

Post-Imputation Inference for Proteomic Data'?
(A) To develop a method for directly measuring protein abundances without missing values.
(B) To create a statistical framework that offers valid and efficient inference for proteomic data by addressing the challenge of missing values.
(C) To replace the Plugin method with a simpler imputation strategy that discards missing values.
(D) To develop a tool that solely relies on low-dimensional covariates for analyzing proteomic data.

Corpus 1 ..... (Irrelevant Content)

Corpus 2 ..... (Irrelevant Content)

Corpus 3 ..... (Correct Content)

Corpus 4 ..... (Irrelevant Content)

Corpus 5 ..... (Irrelevant Content)

Expected Answer:
B

- **Table**: We randomly selected 200–300 rows of noisy data and one row of data entry as the correct context. Then, we gave questions based on the correct context to assess LLMs' noise robustness ability on table. Below is an example we generated from ProtTab.

Example of Table Question

System Message:
Please answer the scientific questions based on the content. You should give your answer directly without any other characters.

User Message:
For the material with CID 13182, what is its inchikey?

cid, mw, mf, xlog... inchikey ... exactmass
CID XXXXX ..... ×
CID 13182 ..... ✓
CID XXXXX ..... ×
CID XXXXX ..... ×

Expected Answer:
ARBSJUHHKXRHAD-UHFFFAOYSA-N

- **KG**: We randomly selected 200–300 rows of KG data to form a noisy context and one row as the correct context. Then, we gave questions based on the correct context to evaluate LLMs' noise robustness ability on KG. Below is an example we generated from PriKG.

Example of KG Question

System Message:
Please answer the scientific questions based on the content. You should give your answer directly without any other characters.

User Message:
How is the gene or protein known as 'GDPD3' connected to the anatomical structure called the 'lymph node'?

relation... x_type,x_name... y_type,y_name
XXXXX ..... ×
anatomy_protein_present... GDPD3... lymph node ...✓
XXXXX ..... ×
XXXXX ..... ×

Expected Answer:
anatomy_protein_present

Negative Rejection

- **Unstructured Text** We randomly selected 5 articles to form a noisy context. Then we asked questions based on an article that doesn't appear in the noisy context. We anticipate that LLMs will produce responses that explicitly indicate rejection of the question, such as "I cannot answer the question due to insufficient information in the retrieved data." Below is an example we generated from MatText.

Example of Unstructured Text Question

System Message:
Please answer the scientific questions based on the content. Your answer

only needs to include the one or more correct option labels, not the full options. You should give your answer directly without any other characters.

User Message:
What key feature of elliptically geared isostatic metamaterials enables their nonlinear topological transitions?
(A) The unique soliton-induced mechanical deformation in linear gear mechanisms.
(B) The nonlinear Berry phase transition facilitated by geometric nonlinearity.
(C) The presence of circular gear geometry that allows reversible elastic deformation.
(D) The linear topological index change due to minor gear rotations.

Corpus 1 ..... (Irrelevant Content)

Corpus 2 ..... (Irrelevant Content)

Corpus 3 ..... (Irrelevant Content)

Corpus 4 ..... (Irrelevant Content)

Corpus 5 ..... (Irrelevant Content)

Expected Answer:
I cannot answer the question due to insufficient information in the retrieved data.

- **Table**: 200-300 rows of table data entries were randomly selected from the database to compose the noisy context. And we asked question based on a data entry which doesn't appear in the noisy context. We anticipate that LLMs will produce responses that explicitly indicate rejection of the question, such as "I cannot answer the question due to insufficient information in the retrieved data." Below is an example we generated from MatTab.

Example of Table Question

System Message:
Please answer the scientific questions based on the content. You should give your answer directly without any other characters.

User Message:
For the material with ID mp-768851, what is its number of site?

Material ID, Formula ... Sites ... Volume, Density
mp-xxxxx ..... ×
mp-xxxxx ..... ×
mp-xxxxx ..... ×
mp-xxxxx ..... ×

Expected Answer:
I cannot answer the question due to insufficient information in the retrieved data.

- **KG**: 200-300 instances of KG data entries were randomly sampled from the database to construct the noisy context. Queries were formulated based on a specific entry which doesn't appear in the noisy KG context. We anticipate that LLMs will produce responses that explicitly indicate rejection of the question, such as "I cannot answer the question due to insufficient information in the retrieved data." Below is an example we generated from PhaKG.

Example of KG Question

System Message:
Please answer the scientific questions based on the content. You should give your answer directly without any other characters.

User Message:
How are the genes "nbc 1" and "nbc 3" related?

Entity1_name... relationship_type,Entity2_name...
XXXXX ..... ×
XXXXX ..... ×
XXXXX ..... ×
XXXXX ..... ×

Expected Answer:
I cannot answer the question due to insufficient information in the retrieved data.

Information Integration

- **Unstructured Text** We randomly sampled 5 articles to form a noisy context and selected two articles as the correct context. Then, we asked questions based on the correct context, requiring LLMs to integrate details from different fragments, thereby evaluating LLMs' information integration ability on text. Below is an example we generated from BioText.

Example of Unstructured Text Question

System Message:
Please answer the scientific questions based on the content. Your answer only needs to include the one or more correct option labels, not the full options. You should give your answer directly without any other characters.

User Message:
Based on the findings of the study, what is the primary long-term effect of local SBRT/IL-12 therapy on the bone marrow of treated mice?
(A) A permanent increase in hematopoietic stem cells (HSCs).
(B) A transient increase in IL-12 levels followed by long-term activation of myeloid cells.
(C) A significant reduction in hematopoietic stem cells (HSCs) accompanied by skewing toward a myeloid lineage bias.
(D) A substantial increase in IL-12 and IFNγ concentrations in the bone marrow.

Corpus 1 ..... (Irrelevant Content)

Corpus 2 ..... (Correct Content)

Corpus 3 ..... (Correct Context)

Corpus 4 ..... (Irrelevant Content)

Corpus 5 ..... (Irrelevant Content)

Expected Answer:
C

- **Table**: We randomly selected 200-300 rows of table data to create a noisy context and two or more rows as the correct context. Then, we asked questions based on the correct context, requiring LLMs to integrate distinct data entries or compare their values, in order to assess LLMs' information integration ability on Table. Below is an example we generated from IaeaTab.

Example of Table Question

System Message:
Please answer the scientific questions based on the content. You should give your answer directly without any other characters.

User Message:
Given the following isotopes ID: NDS-54874, NDS-30453, NDS-69167, NDS-58315, tell me which isotopes has the largest energy?

id, Z, N, symbol... energy[kev]... relative intensity
NDS-XXXXX ..... ×
NDS-30453 ..... ✓
NDS-58315 ..... ✓
NDS-XXXXX ..... ×
NDS-69167..... ✓
NDS-XXXXX ..... ×
NDS-54874..... ✓

Expected Answer:
NDS-69167

- **KG**: We randomly sampled 200-300 entries of KG data to construct a noisy context while designating two or more entries as the correct context. Then we formulated queries based on the correct context, requiring LLMs to synthesize distinct KG entries, thereby evaluating their information integration ability on KG. Below is an example we generated from HipKG.

Example of KG Question

System Message:
Please answer the scientific questions based on the content. You should give your answer directly without any other characters.

User Message:
Could you list the substances that have the potential to interact with DB131_HUMAN?

Name1:DB131_HUMAN... Name2:LRC8A_HUMAN..... ✓
XXXXX ..... ×
Name1:DB131_HUMAN... Name2:RBM12_HUMAN..... ✓
XXXXX ..... ×
Name1:DB131_HUMAN... Name2:AHNK2_HUMAN..... ✓

Expected Answer:
"LRC8A_HUMAN",
"AHNK2_HUMAN",
"RBM12_HUMAN"

Reasoning

- **Unstructured Text** We randomly sampled 5 articles to form a noisy context and designated one article as the correct context. Then, we asked questions based on the correct context, requiring LLMs to perform logical analysis and multi-step reasoning, thereby evaluating LLMs' reasoning ability on text. Below is an example we generated from MatText.

Example of Unstructured Text Question

System Message:
Please answer the scientific questions based on the content. Your answer only needs to include the one or more correct option labels, not the full options. You should give your answer directly without any other characters.

User Message:
Based on the methods and results described in the first part of the study on epitaxial growth of GaAs on Si(001), which of the following is the most plausible reasoning for the effectiveness of the GaSb buffer layer in reducing defect densities such as threading dislocations and antiphase boundaries in the GaAs layer?
(A) The antimonides, such as GaSb, have a significant lattice mismatch with silicon, leading to the generation of interfacial misfit dislocation arrays that efficiently alleviate strain without forming threading dislocations.
(B) The presence of the GaSb buffer layer increases the thickness of the overall film, which inherently reduces the formation of threading dislocations and antiphase boundaries in the GaAs layer.
(C) The GaSb buffer layer chemically reacts with silicon to form a new compound at the interface, which serves as an ideal seed layer for epitaxial GaAs growth, minimizing defect densities.
(D) The GaSb buffer layer promotes planar defects, such as stacking faults, that counterbalance and neutralize threading dislocations and antiphase boundaries in the GaAs layer.

Corpus 1 ..... (Irrelevant Content)

Corpus 2 ..... (Correct Content)

Corpus 3 ..... (Irrelevant Content)

Corpus 4 ..... (Irrelevant Content)

Corpus 5 ..... (Irrelevant Content)

Expected Answer:
A

- **Table**: We randomly selected 200-300 rows of table data to create a noisy context and one or more rows as the correct context. Then, we asked questions based on the correct context, requiring LLMs to perform logical analysis and multi-step reasoning, thereby evaluating LLMs' reasoning ability on table. Below is an example we generated from MatTab.

Example of Table Question

System Message:
Please answer the scientific questions based on the content. Your answer only needs to include the one or more correct option labels, not the full options. You should give your answer directly without any other characters.

User Message:
Comparing materials mp-760154 and mp-1208151, which statement is cor-

rect?
(A) Both materials have identical band gaps and belong to the same crystal system.
(B) The material mp-1208151 has a much larger volume and higher density than mp-760154.
(C) The material mp-760154 is metallic, while mp-1208151 is semiconducting.
(D) Both materials are predicted to be stable with similar formation energies.

Material ID, Formula ... Sites ... Volume, Density
mp-xxxxx ..... ×
mp-760154 ..... ✓
mp-xxxxx ..... ×
mp-1208151 ..... ✓
mp-xxxxx ..... ×

Expected Answer:
B

- **KG** We randomly selected 200-300 rows of KG data to form a noisy context and one or more rows as the correct context. Then we posed questions based on the correct context, requiring LLMs to perform logical analysis about the relationships between various substances, thereby evaluating LLMs' reasoning ability on KG. Below is an example we generated from GoKG.

Example of KG Question

System Message:
Please answer the scientific questions based on the content. You should give your answer directly without any other characters.

User Message:
Given that there exists a shared intermediate term, fill in the blank: GO:0003399 (cytoneme morphogenesis) _____ GO:0048858 (cell projection morphogenesis).

Term id: GO:0003399..... ✓
Term id: GO:XXXXX ..... ×
Term id: GO:XXXXX ..... ×
Term id: GO:0120039..... ✓
Term id: GO:0048858..... ✓

Expected Answer:
is_a

E Detailed Data Source

Table 7 provides detailed information on all databases we used to construct our SciRAGBench, including their URL, description, and license.

F Detailed Model Descriptions

We have selected 14 high-performing LLMs with different scales for this paper. The detailed information of these models are shown in Table 8.

G Case Studies

In this section, we provide several typical bad cases.

Ability: Noise Robustness

Question:
Could you determine the chemical formula for the compound identified as mp-775760?

Correct Answer:
"LiFeF3"

Prediction of GPT-4o-mini:
"C17H20ClN3O2S"

GPT-4o-mini correctly identified the target column and returned a chemical formula as an answer; however, it incorrectly retrieved the context data row in table, resulting in a formula that did not match the Material ID, rendering the final answer incorrect.

Ability: Negative Rejection

Question:
Can you enumerate all the PMIDs related to the interaction between id: 25840 and id:

Table 7: Detailed URL, description, and license of the source data involved in this paper.

Dataset Name	URL	Database Description	License
MatText	arxiv.org	A compilation of material domain research publications.	Open Source
BioText	bio-protocol.org	A peer-reviewed, open-access journal publishing step-by-step life science protocols.	CC BY 4.0
MatTab	next-gen.materialsproject.org	Offer data on over 160,000 inorganic compounds, like crystal structures.	CC BY 4.0
IaeaTab	www-nds.iaea.org	Provide data on evaluated nuclear structure and decay data, including energy levels.	Open Source
ProtTab	pubchem.ncbi.nlm.nih.gov-protein	Offer chemical property information of more than 320,000 common compounds.	Open Source
MolTab	pubchem.ncbi.nlm.nih.gov-chemical	Offer protein information of more than 60,000 common proteins.	Open Source
GoKG	geneontology.org	A standardized framework for biological knowledge, covering molecular function, cellular component, and biological process.	CC BY 4.0
HipKG	cbdm-01.zdv.uni-mainz.de	Offer confidence scored and functionally annotated human protein-protein interactions.	CC BY 4.0
PhaKG	zenodo.org/records	A biomedical KG comprising over 500,000 interconnections between genes, drugs, etc.	CC BY-NC 4.0
PriKG	dataverse.harvard.edu	A KG integrating 20 biomedical resources to describe over 17,000 diseases and 4,000,000 relationships across ten biological scales.	MIT License

Table 8: Overview of the LLMs assessed in our experimental framework.

Model Name	Creator	Domain	#Parameters	Access	URL
GPT-4o	OpenAI	General	undisclosed	API	https://chat.openai.com
GPT-4o-mini	OpenAI	General	undisclosed	API	https://chat.openai.com
Claude-3.5-Sonnet	Anthropic	General	undisclosed	API	https://claude.ai
Deepseek-V3	Deepseek	General	671B	Weights	https://www.deepseek.com
Llama3.1-70B-it	Meta	General	70B	Weights	https://llama.meta.com/llama3
Qwen2.5-7B-it	Alibaba	General	7B	Weights	https://qwenlm.github.io/
GLM4-9B-Chat	Tsinghua&Zhipu	General	9B	Weights	https://huggingface.co/THUDM/glm-4-9b-chat
Llama3.1-8B-it	Meta	General	8B	Weights	https://llama.meta.com/llama3
Gemma2-9B-it	Google	General	9B	Weights	https://ai.google.dev/gemma
Ministral-8B-it	Mistral	General	8B	Weights	https://mistral.ai
ChemDFM-v1.5-8B	SJTU	Chemistry	8B	Weights	https://github.com/OpenDFM/ChemDFM
SciGLM-6B	Tsinghua	Science	6B	Weights	https://github.com/THUDM/SciGLM
LlaSMol-Mistral-7B	OSU	Chemistry	7B	Weights	https://huggingface.co/osunlp/LlaSMol-Mistral-7B
ChemLLM-7B-chat	ShanghaiAILab	Chemistry	7B	Weights	https://huggingface.co/AI4Chem/ChemLLM-7B-Chat

1528?

Correct Answer:
"I cannot answer the question due to insufficient information in the retrieved data."

Prediction of Claude-3.5-Sonnet:
"16239215, 15604093"

Claude-3.5-Sonnet failed to detect the absence of question-relevant context in context. Instead, it identified an incorrect Context Row in KG as the Relevant Context, and thus did not refuse to answer the question, but rather provided an incorrect answer.

Ability: Information Integration

Question:
What are all the pairs of entity names that have a Gene-Gene relationship type?

Correct Answer:
"cyp4f2,ggcx", "hras,kdr", "cyb5r3,cyb5a"

Prediction of SciGLM-6B:
"Gene", "Gene"

SciGLM-6B failed to provide the correct answer and merely repeated the vocabulary from the question. It also failed to output the response in the required format.

Ability: Information Integration

Question:
Among the molecules with cid: 138031, 91721881, 131783619, and 104741, which one possesses the highest heavycnt?

Correct Answer:
131783619

Prediction of ChemLLM-7B-Chat:
49,36 That2811,64,0585

The result from ChemLLM-7B-Chat is entirely unrelated to the question. For large language models with weaker RAG capabilities and instruction-following abilities, the occurrence of such responses is a key reason for their poor performance.

Ability: Reasoning

Question:
Based on the findings of the study, how do ovarian hormones in females influence the metabolic reprogramming effects of Schistosoma mansoni infection on bone marrow-derived macrophages (BMDM)?

Options:
A.Ovarian hormones enhance the metabolic reprogramming...
B.Ovarian hormones do not affect the metabolic...
C.Ovarian hormones inhibit the metabolic reprogramming...
D.Ovarian hormones cause an increase in glycolysis...

Correct Answer:
C

Prediction of ChemDFM-v1.5-8B:
None

Error:
This model's maximum context length is 8192 tokens. However, you requested 13432 tokens in the messages, Please reduce the length of the messages.

Some individual papers exceed the maximum length limit of certain models. In such cases, we can only classify them as errors.

Ability: Reasoning

Question:
Based on the first part of the article, what conclusions can be inferred about the role of surface imperfections in the anisotropic Rashba effect observed in the 2D Janus XA2Y monolayers, and what implications might this have for spintronic applications?
Options:
...

Correct Answer:

A, B

Prediction of GPT-4o:
A

GPT-4o demonstrates some reasoning ability and selected a correct answer; however, it failed to identify all the correct answers in a multiple-choice question.