SciAssess: Benchmarking LLM Proficiency in Scientific Literature Analysis

Hengxing Cai[1*], Xiaochen Cai[1*], Junhan Chang[1*], Sihang Li[1*], Lin Yao[1], Changxin Wang[1], Zhifeng Gao[1], Hongshuai Wang[1], Yongge Li[1], Mujie Lin[1], Shuwen Yang[1], Jiankun Wang[1], Mingjun Xu[1], Jin Huang[1], Xi Fang[1], Jiaxi Zhuang[1], Yuqi Yin[1], Yaqi Li[1], Changhong Chen[1], Zheng Cheng[2], Zifeng Zhao[2], Linfeng Zhang[1,2] and Guolin Ke[1]

[1]DP Technology [2]AI for Science Institute, Beijing

Abstract

Recent breakthroughs in Large Language Models (LLMs) have revolutionized scientific literature analysis. However, existing benchmarks fail to adequately evaluate the proficiency of LLMs in this domain, particularly in scenarios requiring higher-level abilities beyond mere memorization and the handling of multimodal data. In response to this gap, we introduce SciAssess, a benchmark specifically designed for the comprehensive evaluation of LLMs in scientific literature analysis. It aims to thoroughly assess the efficacy of LLMs by evaluating their capabilities in Memorization (L1), Comprehension (L2), and Analysis & Reasoning (L3). It encompasses a variety of tasks drawn from diverse scientific fields, including biology, chemistry, material, and medicine. To ensure the reliability of SciAssess, rigorous quality control measures have been implemented, ensuring accuracy, anonymization, and compliance with copyright standards. SciAssess evaluates 11 LLMs, highlighting their strengths and areas for improvement. We hope this evaluation supports the ongoing development of LLM applications in scientific literature analysis. SciAssess and its resources are available at https://github.com/sci-assess/SciAssess.

1 Introduction

Recent advances in Large Language Models (LLMs), such as GPT (OpenAI, 2023; Brown, 2020), Gemini (Google, 2023), and Llama (Touvron et al., 2023), have attracted considerable attention due to their profound capabilities in natural language understanding and generation (Bubeck et al., 2023). Evaluating these models is crucial for exploring their capability boundaries and limitations, thereby driving technological advancements. In response, a variety of benchmarks tailored for LLMs have been proposed for extensive evaluation, covering a wide range of skills (Zhong et al., 2023; Huang et al., 2023) and diverse tasks (Srivastava et al., 2022; Suzgun et al., 2023a).

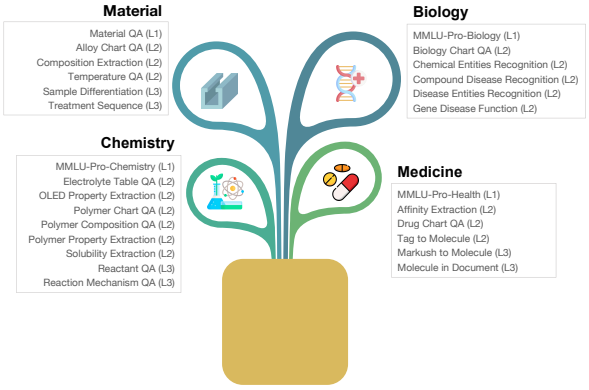

Figure 1: Overview of SciAssess. It spans over 4 subdomains and encompasses 27 tasks.

Despite LLMs not yet fully replacing scientific researchers in generating creative discoveries, they have demonstrated substantial potential in enhancing researchers' efficiency in scientific literature analysis (AI4Science and Quantum, 2023). Specific applications such as automatic literature summarization and knowledge extraction have seen practical deployments, significantly boosting researchers' productivity and expanding the range of literature that can be effectively utilized (Zheng et al., 2023). Inspired by Bloom's Taxonomy (Krathwohl, 2002), we systemize the requirements for scientific literature analysis assistants into three progressive levels: (1) **Memorization (L1)**: Establishing an extensive foundational knowledge base to accurately address common factual questions in various scientific domains; (2) **Comprehension (L2)**: Identifying, extracting, and understanding the core content of provided documents; and (3) **Analysis & Reasoning (L3)**: Integrating extracted information with the existing knowledge base to perform logical reasoning and analysis.

Existing comprehensive LLMs benchmarks, such as MMLU-Pro (Hendrycks et al., 2021; Wang

*Equal Contribution

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2335–2357
April 29 - May 4, 2025

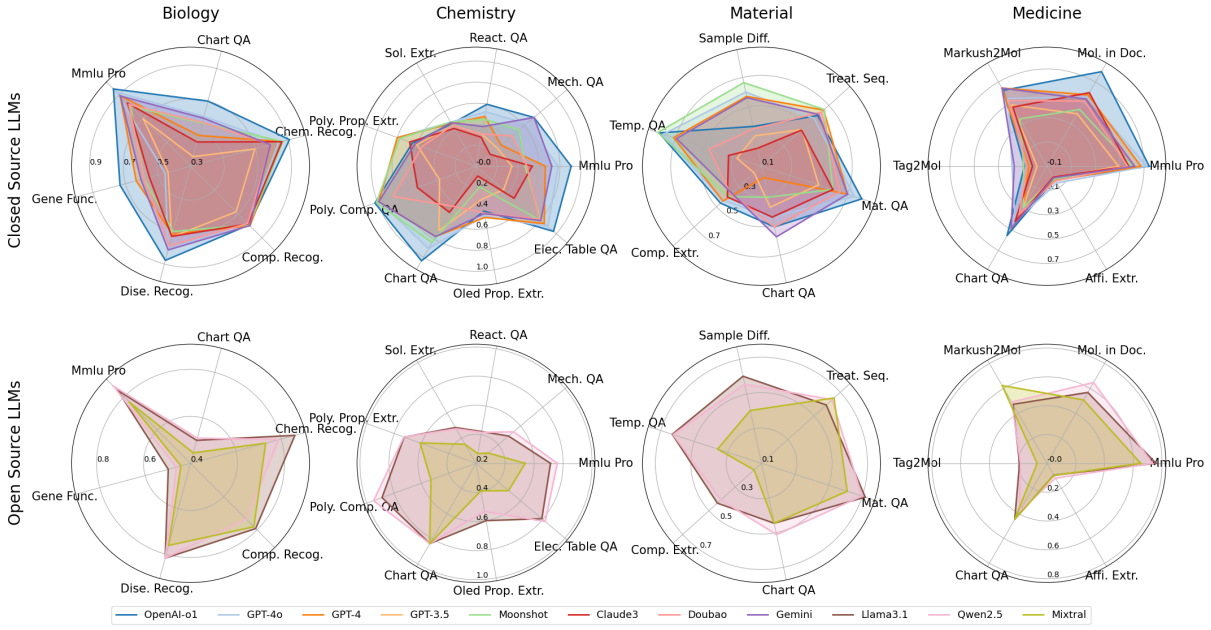

Figure 2: Performance overview of leading open and closed source LLMs on SciAssess. Each column represents a scientific domain. LLMs are evaluated on multiple tasks within each domain, with task details provided in Table 1. For closed source LLMs (first row), GPT-4o and GPT-4 are the leading models. For open source LLMs (second row), Llama3 and Qwen2 emerge as the top models.

et al., 2024), include some tasks related to scientific data. However, these sub-tasks have two limitations: (1) they mostly focus on Memorization, neglecting higher-level abilities such as L2 and L3; (2) these tasks lack the evaluation of various multimodal inputs (*e.g.,* charts, molecular structures, and tables), which are crucial in scientific literature.

In light of these existing limitations, we introduce **SciAssess** (*cf.* Figure 1) – a benchmark specifically designed for scientific literature analysis. SciAssess not only broadens the evaluation scope to encompass a wider range of LLM capabilities but also extends beyond text to include the extraction and interpretation of multimodal contents. Moreover, meticulous design is essential to creating evaluations that yield deep insights, ensure fairness across different LLMs. Consequently, SciAssess is founded on three critical considerations:

Model Ability. A benchmark must clearly delineate the desired capabilities and model the intrinsic relationships among them, facilitating a diagnostic understanding. Thus, SciAssess evaluates across three progressive levels (*i.e.,* Memorization (L1), Comprehension (L2), and Analysis & Reasoning (L3)) and five modalities (*i.e.,* texts, charts, chemical reactions, molecular structures, and tables). Consequently, SciAssess yields nuanced and informative evaluation outcomes, pinpointing specific aspects where the examined models may fall short.

Scope & Task. Benchmarks should encompass a broad array of scientific domains to ensure comprehensiveness. Within each domain, the selected tasks must authentically represent the typical challenges and scenarios characteristic of that field. Consequently, SciAssess spans over 4 subdomains (*i.e.,* biology, chemistry, material, and medicine) and encompasses 27 tasks, each carefully suggested or designed by domain experts according to their professional experience.

Scale & Quality Control. The scale and quality of the benchmark must be impeccable to serve as a dependable basis for deriving accurate, actionable, and applicable insights. SciAssess contains 6,938 questions in total to ensure adequate scale. Each question is transformed from existing datasets or manually curated by domain experts hired by us [1]. Subsequently, expert cross-validation is performed to ensure correctness and reliability.

Overall, SciAssess aims to reveal the performance of LLMs as a scientific literature analysis assistant, thereby identifying their strength and weaknesses. The insights gained from SciAssess could

[1]All data collection, annotation, and quality control tasks were carried out by the authors (who are also employees of the company) as part of their job responsibilities, and therefore, they were not provided with any additional compensation.

hopefully catalyze further enhancing the capabilities of LLMs in scientific literature analysis, ultimately contributing to the acceleration of scientific discovery and innovation.

2 Benchmark Dataset

We begin by outlining the ability assessment framework in Section 2.1, which serves as the backbone of our evaluation framework. Moving forward, we provide detailed description of evaluation scopes and tasks in Section 2.2. Lastly, we present the quality control measures implemented to ensure the integrity and reliability in Section 2.3.

2.1 Ability Assessment Framework

Guided by the widely accepted cognitive learning processes outlined in Bloom's Taxonomy (Krathwohl, 2002), we propose that the evaluation of LLMs in scientific literature analysis should be classified into three core levels:

Memorization (L1) refers to the model's extensive knowledge base, which allows it to accurately answer common factual questions in science autonomously. **Comprehension (L2)** is the ability to precisely identify and extract key information and facts within a given text, and to comprehend them. **Analysis & Reasoning (L3)** demonstrate the model's advanced capability to amalgamate extracted information with its existing knowledge base for logical reasoning and analysis, leading to well-founded conclusions or predictions.

Inspecting existing LLM benchmarks in science field (See Section 4) through three-level ability assessment framework, we find that they mostly focus on Memorization (L1) – the foundational knowledge base for scientific facts – while overlooking the higher-level abilities of Comprehension (L2) and Analysis & Reasoning (L3).

Given the significant potential of leveraging LLMs as scientific literature analysis assistants to boost scientific discovery, we propose SciAssess as a more comprehensive benchmark, in terms of tasks, scopes, and modalities.

2.2 Scope & Task

After categorizing the ability of of LLMs into three levels, we proceed to introduce how we choose the tasks in SciAssess. First, we include four vertical domains: biology, chemistry, material, and medicine, as shown in Figure 1. This categorization ensures that SciAssess captures the unique challenges and requirements of each specific field. Then, as mentioned above, Memorization (L1), being the extensive foundation for other higher-level abilities, should encompass as large a knowledge base as possible. Thus, SciAssess includes factual questions in MMLU-Pro (Wang et al., 2024) and MaScQA (Zaki et al., 2023), covering fundamental knowledge in each field. For the evaluation of Comprehension (L2) and Analysis & Reasoning (L3), we identify realistic demands by consulting domain experts and curate corresponding tasks. The reason is that solving tasks in these domains require finer-grained abilities, such as understanding tables and molecular structures. For instance, crucial composition information in material science literature is often found in tables, whereas key information extraction in drug discovery necessitates the accurate recognition of molecular structures.

SciAssess, as presented in Table 1, comprises 6,888 questions across 27 tasks in five scientific domain, encompassing three ability levels: Memorization (L1), Comprehension (L2), and Analysis & Reasoning (L3). Of these tasks, 7 out of 27 are transformed from existing public datasets (gray tasks in Table 1) , and the other 21 tasks curated by us are based on contents from academic papers, specifically designed to assess the ability to analyze scientific literature. We show the token lengths (GPT-4 tokenizer) of questions and answers for each task in Figure 3. SciAssess also includes five types of questions (*i.e.,* true/false questions, multiple-choice questions, table extraction, text extraction, and molecule generation) with four metrics (*i.e.,* accuracy, recall, F1-score, and molecule similarity). For detailed descriptions and concrete examples, please refer to Appendix A. We also provide general prompt template and specific prompt for each task in Appendix B and C, respectively.

2.2.1 Biology

Biological literature encompasses a wealth of specialized terminology and complex concepts, as well as a significant amount of non-textual information such as tables and figures. Effectively extracting and integrating these elements presents a crucial challenge. Given that tasks in the biological domain typically require precise identification and understanding of intricate biological entities, processes, and relationships, we have selected a set of representative tasks, including the recognition of specialized terminology, the comprehension of chart information, and the extraction of entity rela-

Domain	Task	Ability	# Questions	Question Type	Metric	Modality
Biology	MMLU-Pro-Biology	L1	717	Multiple Choice	Accuracy	Text only
	Biology Chart QA	L2	199	Multiple Choice	Accuracy	Chart
	Chemical Entities Recognition	L2	500	Text Extraction	F1-score	Text only
	Compound Disease Recognition	L2	775	Text Extraction	F1-score	Text only
	Disease Entities Recognition	L2	500	Text Extraction	F1-score	Text only
	Gene Disease Function	L2	24	Text Extraction	F1-score	Text only
Chemistry	MMLU-Pro-Chemistry	L1	1,132	Multiple Choice	Accuracy	Text only
	Electrolyte Table QA	L2	200	Multiple Choice	Accuracy	Table
	OLED Property Extraction	L2	13	Table Extraction	Recall	Mol., Table
	Polymer Chart QA	L2	15	Multiple Choice	Accuracy	Chart
	Polymer Composition QA	L2	209	Multiple Choice	Accuracy	Text only
	Polymer Property Extraction	L2	109	Table Extraction	Recall	Table
	Solubility Extraction	L2	100	Table Extraction	Recall	Table
	Reactant QA	L3	195	Multiple Choice	Accuracy	Reaction
	Reaction Mechanism QA	L3	22	Multiple Choice	Accuracy	Reaction
Material	Material QA	L1	263	Multiple Choice	Accuracy	Text only
	Alloy Chart QA	L2	15	Multiple Choice	Accuracy	Chart
	Composition Extraction	L2	244	Table Extraction	Recall	Table
	Temperature QA	L2	207	Multiple Choice	Accuracy	Text only
	Sample Differentiation	L3	237	Multiple Choice	Accuracy	Text only
	Treatment Sequence	L3	202	True/False	Accuracy	Text only
Medicine	MMLU-Pro-Health	L1	818	Multiple Choice	Accuracy	Text only
	Affinity Extraction	L2	40	Table Extraction	Recall	Mol., Table
	Drug Chart QA	L2	15	Multiple Choice	Accuracy	Chart
	Tag to Molecule	L2	50	Mol. Generation	Mol. Similarity	Mol.
	Markush to Molecule	L3	37	Mol. Generation	Mol. Similarity	Mol.
	Molecule in Document	L3	50	True/False	Accuracy	Mol.

Table 1: Statistics of the SciAssess. It comprises 6,888 questions across 27 tasks in five sub-domains. Tasks are categorized into three ability levels: Memorization (L1), Comprehension (L2), and Analysis & Reasoning (L3). Tasks that are gray are transformed from existing datasets, while others are curated by domain experts hired by us.

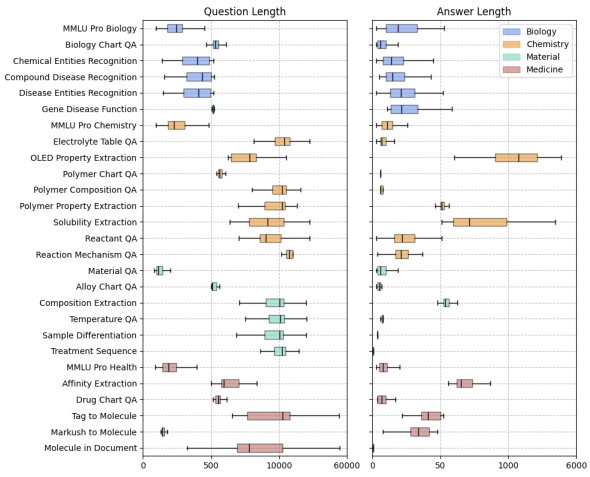

Figure 3: Distribution of token length for questions and answers in each task.

tionships, to evaluate the performance in this field.

In this domain, following tasks are devised: MMLU-Pro-Biology, biology chart QA, chemical entities recognition, compound disease recognition, disease entities recognition, and gene disease function. Detailed descriptions and prompts are provided in Appendix C.1.

2.2.2 Chemistry

The field of chemistry involves a vast array of complex molecular structures, chemical reactions, and properties, alongside a substantial amount of data presented in formulas, reaction equations, and diagrams. Effectively processing and interpreting these components is a significant challenge for language models. Tasks in the chemical domain demand precise understanding of molecular compositions, reaction mechanisms, and material properties. To evaluate the performance of LLMs in this domain, we have selected representative tasks such as the recognition of chemical compounds, the interpretation of reaction pathways, and the extraction of relationships between chemical entities.

We devise following tasks for organic materials: MMLU-Pro-Chemistry, electrolyte table QA, OLED property extraction, polymer chart QA, polymer composition extraction, polymer property extraction, solubility extraction, reactant QA, and reaction mechanism QA. Detailed descriptions and prompt templates are provided in Appendix C.2.

2.2.3 Materials

Materials science encompasses a broad range of substances, including metals, ceramics, polymers, and composites, each with distinct properties and applications. These materials are widely used across industries such as aerospace, automotive, and construction. By fine-tuning their composition, structure, and processing techniques, materials can be engineered to meet specific performance requirements (Caron and Khan, 1983). Accurately

extracting material compositions, structural characteristics, and process parameters from the literature is essential for advancing material design and optimization.

Specifically, following tasks are devised: material QA, Alloy Chart QA, composition extraction, temperature QA, sample differentiation, and treatment sequence. Detailed descriptions and prompt templates are provided in Appendix C.3.

2.2.4 Medicine

Medicine focuses on developing new therapeutics. Leveraging advanced intelligent tools, especially LLMs, can significantly enhance the efficiency and effectiveness of discovering and developing new drugs. To evaluate the capability of LLMs in this domain, it is imperative to develop specialized tasks that reflect the complexities and nuances of biomedical research. By designing targeted tasks, we can better assess the ability of LLMs to navigate and interpret the wealth of information critical to the development of new therapeutics.

Specifically, we devise: MMLU-Pro-Health, affinity extraction, drug chart QA, tag to molecule, markush to molecule, molecule in document. Detailed descriptions and prompts are provided in Appendix C.4.

2.3 Data Quality, Privacy, and Copyright

To safeguard the quality and ethical standards, meticulous steps were undertaken in its preparation and validation:

Distractor Construction: Our data points are human-annotated, as well as the distractors. And how the distractors are determined depends on the specific task. For example, in value-type multiple-choice questions, the distractors are values near the ground truth. For extraction tasks, the distractors are other targets in the given context, except for the ground truth target.

Expert Validation: Each data point (as indicated by black tasks in Table 1) is independently labeled by two annotators who are domain experts in the relevant fields. If their labels agree, the label is accepted; if not, they engage in a discussion to determine the final label. Their initial annotations have a Cohen's Kappa value (McHugh, 2012) of 0.75, which indicates high reliability or agreement.

Screening and Anonymization: Our annotators were instructed not to use any data samples containing sensitive information when building the benchmark. For example, data samples including personal health information or specific drug details were carefully reviewed. If such sensitive information was identified, it was either anonymized by removing personal identifiers or replacing specific details with general terms, or the entire sample was excluded from the benchmark.

Copyright Compliance: Our benchmark includes two types of data: some are adopted from existing benchmarks, and others are constructed from scratch by our team. For the data adopted from existing benchmarks, we provide the corresponding sources. For the data we created, we have obtained the necessary copyrights for the files used. To ensure full compliance with copyright laws, our repository only provides the Digital Object Identifier (DOI) for papers or patent number, and does not distribute the actual documents. Researchers need to download the necessary files independently. Detailed instructions is included in the codebase to guide researchers on how to place the downloaded documents into the designated folder.

3 Experiment

3.1 Experiment Setup

Baseline LLMs. To measure how leading LLMs perform on SciAssess, we benchmark extensively. For closed-source LLMs, we test OpenAI-o1, GPT-4o, GPT-4, GPT-3.5, Gemini-1.5-Pro, Claude 3 Opus, Moonshot-v1 and Doubao. For open-source LLMs, we test Llama-3.1-70B, Mixtral-8x22B-Instruct-v0.1, and Qwen-2.5-72B. Briefs about all models are provided in Appendix F.

Experiment Workflow. For closed-source models, we utilize the official API calls provided by the model developers, while for open-source models, we obtain these models from HuggingFace (Wolf et al., 2019), deploy them locally with vllm (Kwon et al., 2023), and then perform the tests. Tasks curated by us require real context from papers, thus the PDF content needs to be converted to text as inputs for LLMs. If the LLM includes a built-in PDF parsing interface (*e.g.,* Gemini and Moonshot), we simply use the interface; otherwise, we employ PyPDF2 [2], a widely-used open-source PDF parsing tool. Our aim is to explore the ability boundary of LLMs, thus strategies that enhance LLMs' inference ability (*i.e.,* in-context learning (Brown, 2020) and chain-of-thought (Wei et al., 2022)) are adopted. Specifically, due to the input length limitations of the LLMs, tasks requiring long context of

[2] https://pypdf2.readthedocs.io/en/3.x/

Domain	Task	o1	GPT-4o	GPT-4	GPT-3.5	Moonshot	Claude3	Doubao	Gemini	Llama3.1	Qwen2.5	Mixtral
Biology	MMLU-Pro-Biology*	**0.901**	0.874	0.845	0.650	0.755	0.781	0.770	0.842	0.815	**0.840**	0.743
	Biology Chart QA	**0.653**	0.558	0.442	0.312	0.518	0.402	0.523	0.548	0.477	**0.487**	0.422
	Chemical Entities Recognition*	**0.862**	0.795	0.817	0.649	0.821	0.815	0.749	0.745	**0.836**	0.764	0.707
	Compound Disease Recognition*	0.745	0.733	**0.753**	0.636	0.745	0.737	0.733	0.751	**0.768**	0.712	0.757
	Disease Entities Recognition*	**0.831**	0.763	0.670	0.688	0.654	0.684	0.742	0.767	**0.793**	**0.793**	0.737
	Gene Disease Function*	**0.687**	0.410	0.587	0.391	0.538	0.506	0.539	0.558	**0.474**	0.438	0.418
Chemistry	MMLU-Pro-Chemistry*	**0.868**	0.745	0.621	0.303	0.428	0.496	0.446	0.683	0.676	**0.723**	0.501
	Electrolyte Table QA	**0.925**	0.855	0.810	0.305	0.765	0.435	0.745	0.765	0.755	**0.785**	0.455
	OLED Property Extraction	0.394	0.438	**0.455**	0.280	0.160	0.055	0.413	0.419	**0.563**	0.499	0.355
	Polymer Chart QA	**1.000**	0.867	0.733	0.667	0.800	0.467	0.400	0.733	**0.800**	**0.800**	**0.800**
	Polymer Composition QA	**0.986**	0.938	0.947	0.330	0.971	0.555	0.804	0.947	0.852	**0.914**	0.493
	Polymer Property Extraction	0.606	**0.759**	0.758	0.562	0.736	0.634	0.508	0.580	0.690	**0.692**	0.573
	Solubility Extraction	0.427	0.444	0.431	0.408	**0.445**	0.375	0.409	0.440	**0.447**	0.437	0.314
	Reactant QA	**0.559**	0.487	0.441	0.272	0.415	0.241	0.272	0.344	**0.385**	0.379	0.231
	Reaction Mechanism QA	**0.682**	0.591	0.273	0.227	0.500	0.136	0.409	**0.682**	0.455	**0.500**	0.273
Material	Material QA	**0.821**	0.768	0.722	0.521	0.620	0.620	0.669	0.722	**0.738**	0.719	0.631
	Alloy Chart QA	0.533	0.467	0.200	0.400	0.333	0.467	0.533	**0.600**	0.467	**0.533**	0.467
	Composition Extraction	**0.488**	0.462	0.467	0.189	0.423	0.427	0.398	0.389	**0.457**	0.430	0.177
	Temperature QA	0.836	0.807	0.734	0.295	**0.845**	0.353	0.488	0.715	**0.652**	0.647	0.382
	Sample Differentiation	0.392	0.624	0.595	0.329	**0.688**	0.245	0.376	0.586	**0.624**	0.578	0.426
	Treatment Sequence	0.624	0.594	0.678	0.485	**0.683**	0.480	0.658	0.634	0.614	0.658	**0.673**
Medicine	MMLU-Pro-Health*	**0.784**	0.763	0.715	0.531	0.644	0.614	0.605	0.663	**0.710**	0.685	0.603
	Affinity Extraction	0.068	**0.101**	0.076	0.055	0.063	0.045	0.081	0.052	0.047	**0.071**	0.049
	Drug Chart QA	**0.600**	0.467	0.333	0.333	0.333	0.467	0.400	0.533	**0.400**	0.333	**0.400**
	Tag2Mol	0.127	**0.229**	0.092	0.023	0.133	0.061	0.105	0.211	**0.143**	0.136	0.021
	Markush2Mol	0.662	0.585	**0.684**	0.523	0.391	0.503	0.565	0.683	0.425	0.443	**0.576**
	Mol In Document	**0.840**	0.600	0.620	0.440	0.480	0.640	0.560	0.580	0.520	**0.600**	0.460

Table 2: Performance comparison of LLMs across various scientific domains. **Orange** and **green** indicate the best in closed and open source LLMs, respectively. Chain-of-thought prompt is implemented for each task and model, except for OpenAI-o1. * indicates 3-shot.

a PDF document are executed in a zero-shot manner. Tasks that do not require such long context (*e.g.,* MMLU-Pro, entities recognition) are evaluated using 3-shot settings. And chain-of-thought prompt is implemented in every task by prompting the model to think step-by-step before concluding. The only exception is OpenAI-o1, whose official prompt guideline suggests users to "avoid chain-of-thought prompts". We also provide complete performance evaluation results with chain-of-thought prompts in Appendix D.

3.2 Results and Analysis

In this section, we analyze the performance of LLMs on SciAssess. The overall performance comparison, as summarized in Table 2, reveals the distinct strengths and weaknesses of each model in science literature analysis.

3.2.1 Performances of Different Ability Levels

Table 3 presents the performance of evaluated LLMs across three progressive ability level. Tasks are further categorized according to their question types, with average results and rankings provided for each ability levels. We observe the following: (1) **Memorization (L1):** OpenAI-o1 and Qwen-2.5 demonstrates the highest average accuracy of 0.843 and 0.742, respectively, indicating consistently superior performance in memorization tasks. (2) **Comprehension (L2):** OpenAI-o1 excels in multiple-choice and text extraction comprehension with accuracy of 0.790 and 0.781, respectively, and maintains the top average rank of 3.25. Notably, the only L2-level molecule generation task, *Tag to Molecule*, reveals poor performance across all LLMs. As illustrated in Figure 4, current PDF parsing technologies, whether open-source like PyPDF or proprietary like Gemini or Moonshot, fail to effectively parse molecular structures in documents. Consequently, LLMs struggle with the *Tag to Molecule* task. We propose that a critical advancement for future LLM-based literature understanding assistant is the integration of PDF parsing solutions capable of recognizing molecular structures. (3) **Analysis & Reasoning (L3):** The average rank reveals OpenAI-o1, GPT-4, and Gemini are the top performers with ranks of 2.00, 3.33, and 3.33, respectively.

Overall, OpenAI-o1 consistently ranks high across all ability levels. GPT-4o and Gemini also demonstrate strong overall performance, especially in memorization and reasoning.

Based on these observations, we suggest the following recommendations: (1) For tasks heavily reliant on memorization, OpenAI-o1 and Qwen2.5 are recommended due to their high accuracy and ranking; (2) For comprehension tasks, particularly those involving complex data extraction and generation, OpenAI-o1 is an ideal choice. (3) For analysis and reasoning tasks, OpenAI flagship models (*i.e.,* o1, GPT4, GPT-4o), and Gemini provide reliable performance and should be considered.

Ability Level	Question Type	Metric	o1	GPT-4o	GPT-4	GPT-3.5	Moonshot	Claude3	Doubao	Gemini	Llama3.1	Qwen2.5	Mixtral
Memorization (L1)	Multiple Choice	Accuracy	**0.843**	0.788	0.726	0.501	0.612	0.628	0.622	0.728	0.735	**0.742**	0.619
	Average Rank		**1.000**	2.000	6.000	11.000	10.000	7.000	8.000	5.000	4.000	**3.000**	9.000
Comprehension (L2)	Multiple Choice	Accuracy	**0.790**	0.708	0.600	0.377	0.652	0.449	0.556	0.692	0.629	**0.643**	0.488
	Table Extraction	Recall	0.397	**0.441**	0.437	0.299	0.365	0.307	0.362	0.376	**0.441**	0.426	0.294
	Text Extraction	F1-score	**0.781**	0.675	0.707	0.591	0.690	0.686	0.691	0.705	**0.718**	0.677	0.655
	Mol. Generation	Mol. Similarity	0.127	**0.229**	0.092	0.023	0.133	0.061	0.105	0.211	**0.143**	0.136	0.021
	Average Rank		**3.250**	3.375	5.250	10.500	5.500	8.750	7.000	3.750	**3.125**	5.250	10.250
Analysis & Reasoning (L3)	Multiple Choice	Accuracy	0.544	**0.567**	0.436	0.276	0.534	0.207	0.352	0.537	**0.488**	0.486	0.310
	Mol. Generation	Mol. Similarity	0.662	0.585	**0.684**	0.523	0.391	0.503	0.565	0.683	0.425	0.443	**0.576**
	True/False	Accuracy	**0.732**	0.597	0.649	0.462	0.582	0.560	0.609	0.607	0.567	**0.629**	0.566
	Average Rank		**2.000**	3.667	3.333	9.333	7.333	9.667	6.000	3.333	7.667	**6.000**	7.667

Table 3: Performance on Memorization (L1), Comprehension (L2), and Analysis & Reasoning (L3) tasks.

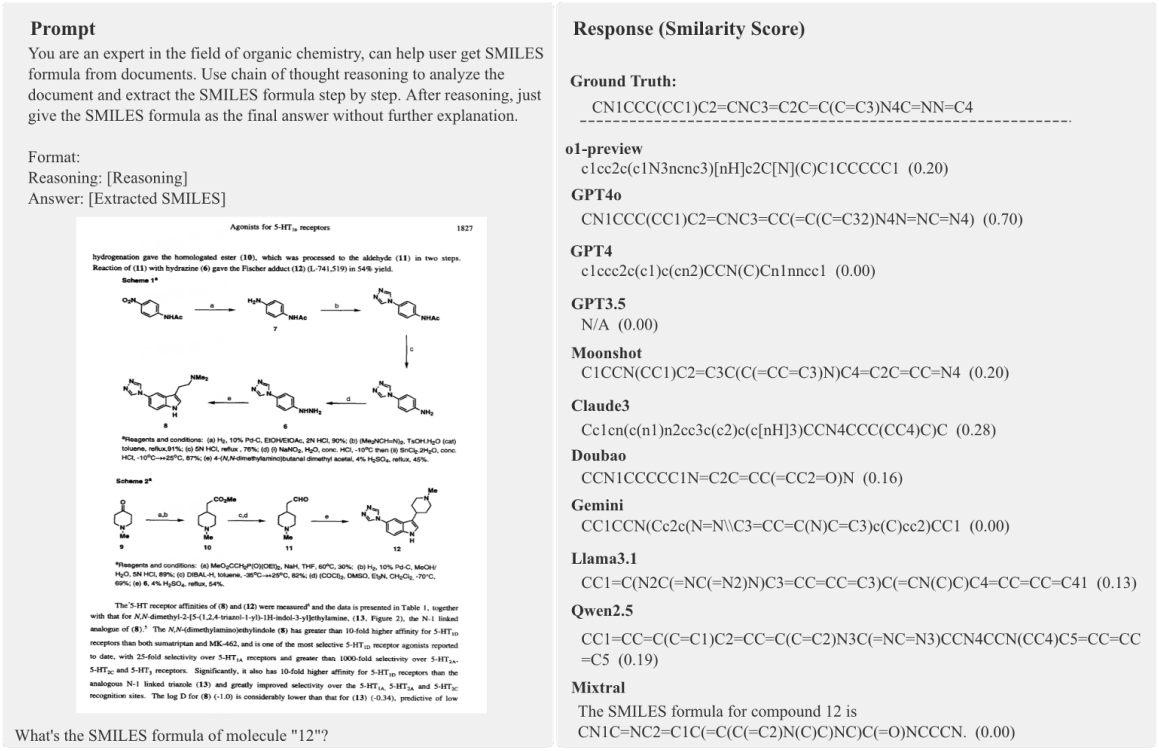

Figure 4: Example of Tag to Molecule task.

3.2.2 Performance on Multimodal Contents

Table 4 summarizes the performance of LLMs on multimodal content tasks. For each modality, performances are averaged over different question types. We observe the following: (1) **Text-only tasks:** GPT-4 achieves the highest average rank (2.00). (2) **Chart tasks:** OpenAI-o1 exhibit the highest accuracy (0.696). (3) **Chemical reaction tasks:** OpenAI-o1 stands out with high accuracy in multiple-choice questions (0.62). (3) **Molecule tasks:** GPT-4o excels with average ranks of 3.17, particularly in table extraction task. (5) **Table tasks:** GPT-4o lead with the highest table extraction recall (0.44).

Overall, OpenAI models consistently rank as top performers across most modalities. Gemini also demonstrate strong performance, especially in molecule generation tasks.

Based on these observations, we suggest the following recommendations: (1) For text-only tasks, OpenAI-o1 and GPT-4o are highly recommended due to their superior accuracy and ranking. (2) For chart and chemical reaction tasks, OpenAI-o1 excels, making it suitable for such specialized applications. (3) For molecule structure and tabular tasks, GPT-4o is the preferred model, given it remarkable performance.

3.2.3 Error Patterns and Performance Analysis

To analyze the challenges LLMs face in handling multimodal contents, we conducted an error analy-

Modality	Question Type	Metric	o1	GPT-4o	GPT-4	GPT-3.5	Moonshot	Claude3	Doubao	Gemini	Llama3.1	Qwen2.5	Mixtral
Text Only	Multiple Choice	Accuracy	**0.798**	0.788	0.740	0.423	0.707	0.523	0.594	0.737	0.724	**0.729**	0.540
	Text Extraction	F1-score	**0.781**	0.675	0.707	0.591	0.690	0.686	0.691	0.705	**0.718**	0.677	0.655
	True/False	Accuracy	0.624	0.594	0.678	0.485	**0.683**	0.480	0.658	0.634	0.614	0.658	**0.673**
	Average Rank		3.000	6.667	**2.667**	10.667	4.667	9.333	5.833	4.667	**5.333**	5.833	7.333
Chart	Multiple Choice	Accuracy	**0.696**	0.590	0.427	0.428	0.496	0.451	0.464	0.604	0.536	**0.538**	0.522
	Average Rank		**1.000**	3.000	11.000	10.000	7.000	9.000	8.000	2.000	5.000	**4.000**	6.000
Reaction	Multiple Choice	Accuracy	**0.620**	0.539	0.357	0.250	0.458	0.188	0.340	0.513	0.420	**0.440**	0.252
	Average Rank		**1.000**	2.000	7.000	10.000	4.000	11.000	8.000	3.000	6.000	**5.000**	9.000
Mol.	Table Extraction	Recall	0.231	**0.270**	0.266	0.168	0.112	0.050	0.247	0.236	**0.305**	0.285	0.202
	Mol. Generation	Mol. Similarity	0.394	0.407	0.388	0.273	0.262	0.282	0.335	**0.447**	0.284	0.290	**0.298**
	True/False	Accuracy	**0.840**	0.600	0.620	0.440	0.480	0.640	0.560	0.580	0.520	**0.600**	0.460
	Average Rank		3.667	**3.167**	3.667	10.000	10.000	7.333	5.667	4.333	5.667	**4.500**	8.000
Table	Multiple Choice	Accuracy	**0.925**	0.855	0.810	0.305	0.765	0.435	0.745	0.765	0.755	**0.785**	0.455
	Table Extraction	Recall	0.397	**0.441**	0.437	0.299	0.365	0.307	0.362	0.376	**0.441**	0.426	0.294
	Average Rank		3.000	**1.750**	3.000	10.500	6.250	9.500	8.000	5.750	4.250	**4.000**	10.000

Table 4: Performance on multimodal contents.

sis on the *Tag to Molecule* task. In this task, models generate a SMILES formula based on extracted textual and visual information from scientific literature. The process involves three key steps: (1) identifying the correct molecular reference from the text, (2) comprehending the corresponding molecular structure diagram, and (3) generating a chemically valid SMILES representation.

Error Analysis. Figure 4 presents an example of this task, along with model-generated outputs, revealing three primary types of errors: (1) Misalignment Between Text and Diagram. Some models failed to associate textual descriptions (e.g., *"molecule 12"*) with the correct molecular structure. For instance, Moonshot generated a SMILES sequence that deviated significantly from the ground truth. (2) Failure to Comprehend Molecular Diagrams. Most models struggled to interpret molecular structures accurately. Even GPT-4o, with the highest similarity score (0.7), exhibited structural inaccuracies, such as missing functional groups and incorrect bond placements. (3) SMILES Syntax Errors. Some models produced syntactically invalid SMILES formulas. GPT-4 and Gemini, for example, generated outputs with missing stereochemical information or incorrect bonding details, as confirmed by RDKit validation.

Potential Improvements The observed errors suggest that LLMs still face significant challenges in processing and comprehending over multimodal scientific information. To enhance performance in this task, improvements are needed in the following areas: (1) Comprehending Molecular Diagrams. Many models struggle to correctly interpret molecular structures from diagrams, often missing key functional groups or misidentifying structural elements. Enhancing the ability to extract fine-grained details from molecular representations is crucial. (2) Aligning Diagrams with Textual Descriptions. Establishing accurate correspondences between molecular diagrams and their textual references remains challenging. Models need better mechanisms to associate entity mentions in text with the correct visual structures, reducing misalignment errors. (3) Domain-Specific Expertise. Accurate SMILES generation requires deep chemical knowledge, particularly in recognizing stereochemistry, bond configurations, and structural constraints. Further domain adaptation and fine-tuning on specialized chemical corpora could improve model reliability.

3.2.4 SciAssess vs. Existing Benchmarks

To better understand the positioning of SciAssess within existing evaluation frameworks, we compare it with two representative benchmarks: (1) MMLU-Pro (Wang et al., 2024), a general-purpose benchmark, and (2) SciKnowEval (Feng et al., 2024), a domain-specific benchmark focused on scientific knowledge. Detailed ranking comparisons can be found in Appendix E.

Comparison with MMLU-Pro. The ranking comparison between MMLU-Pro and SciAssess reveals both similarities and notable deviations: (1) Consistency in High-Performing Models: GPT-4o consistently ranks at the top across both benchmarks (1st in MMLU-Pro; 1st in SciAssess L1). Similarly, Qwen2.5 maintains a strong position (2nd in MMLU-Pro; 2nd in SciAssess L1), consistently ranking among the top-performing models in both benchmarks. (2) Variations Reflecting Benchmark Focus: Certain models exhibit ranking discrepancies. For example, Llama3.1 ranks 7th in MMLU-Pro but performs significantly better in SciAssess (3rd in L1, 1st in L2). This suggests

that SciAssess captures aspects of domain-specific reasoning that MMLU-Pro does not emphasize.

Comparison with SciKnowEval. When compared to SciKnowEval, SciAssess rankings exhibit greater alignment: (1) Domain-specific similarities: Models such as GPT-4o and Gemini rank highly in both benchmarks, indicating their effectiveness in scientific tasks. (2) Distinctions in evaluation scope: While SciKnowEval ranks GPT-4 as 2nd, SciAssess highlights performance variations across its three evaluation levels (3rd in L1, 1st in L3), providing a more detailed assessment of capabilities at different levels of reasoning complexity.

Key Insights on Benchmark Positioning. This comparison highlights the complementary role of SciAssess in existing evaluation frameworks: (1) Bridging the gap between general and scientific benchmarks: Unlike MMLU-Pro, which focuses on general knowledge assessment, SciAssess provides a structured evaluation in the scientific domain, addressing gaps in existing benchmarks. (2) Fine-grained scientific domain evaluation: Similar to SciKnowEval, SciAssess assesses domain-specific tasks but introduces a multi-level framework (L1-L3) that enables a more nuanced analysis of model performance across varying complexities.

4 Related Work

General benchmarks for LLMs. LLMs are evaluated across a variety of benchmarks to comprehensively assess their capabilities. Some benchmarks, such as MMLU (Hendrycks et al., 2021), MMLU-pro (Wang et al., 2024), CMMLU (Li et al., 2023a), and Xiezhi (Gu et al., 2024), are instrumental in evaluating models' world knowledge across diverse domains. For reasoning capabilities, benchmarks like GSM8k (Cobbe et al., 2021) and BBH (Suzgun et al., 2023b) provide rigorous assessments of models' problem-solving and logical reasoning skills. In the realm of programming, benchmarks such as HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) serve as popular testbeds for evaluating models' coding proficiency. Additionally, TruthfulQA (Lin et al., 2022) and HaluEval (Li et al., 2023b) are pivotal in assessing the veracity of models' outputs, ensuring their alignment with factual information.

Although some general benchmarks include a subset of science subjects, they mostly focus on Memorization (L1) and often overlook higher-level abilities such as Comprehension (L2) and Analysis & Reasoning (L3). Furthermore, these benchmarks lack context-involved tasks, for example, understanding and reasoning over a scientific paper.

Scientific literature benchmarks. Prior works have made significant strides in developing LLM benchmarks to assess the understanding of scientific literature. In the biomedical domain, notable efforts include BLUE (Peng et al., 2019), which provides a set of tasks for evaluating models on various aspects of biomedical text-mining. Building on this, BLURB (Gu et al., 2021) offers an extensive collection of datasets to further refine model performance in this specialized field. More recently, InBoXBART (Parmar et al., 2022) has been introduced, focusing on integrating information across multiple biomedical documents. SciRIFF (Wadden et al., 2024) is designed to extract and synthesize information from research literature across various scientific disciplines.

Compared with existing scientific literature benchmarks, SciAssess focuses more on tasks for interpreting multi-modal content (*e.g.,* molecular structures and tables), which are common in scientific literature. Moreover, it features a real-world application scenarios that LLMs digest parsed PDF contents with parsing errors.

5 Conclusion and Future Work

SciAssess rigorously assesses the capabilities of LLMs for scientific literature analysis. It focuses four specialized areas: biology, chemistry, material, and medicine. The benchmark focuses on assessing LLMs' core competencies in Memorization (L1), Comprehension (L2), and Analysis & Reasoning (L3) within the context of scientific literature analysis. Through detailed evaluations of 11 LLMs, SciAssess highlights their strengths and identifies areas needing improvement across various ability levels, content modalities, and contextual scenarios. Additionally, we emphasize the urgent need for PDF parsing algorithms tailored to handle content of various modalities, such as molecular structures and chemical reactions. We hope that SciAssess supports the ongoing development of LLMs in scientific literature analysis. Looking ahead, we plan to broaden the range of scientific domains included in SciAssess and incorporate more vertical domains. These enhancements aim to improve the benchmark's utility and efficacy, providing clearer guidance and fostering the advancement of LLMs in scientific literature analysis.

Limitation

While SciAssess provides a comprehensive and valuable benchmarking suite across four primary domains – biology, chemistry, material, and medicine – there are several limitations to consider. Firstly, the scope of SciAssess is currently constrained to these four domains, with potential future extensions to other vertical domains such as physics and engineering.

Secondly, the creation and curation of high-quality, domain-specific training data are essential for the effective evaluation and improvement of LLMs. However, due to the high cost associated with manual labeling, SciAssess does not provide additional training data for these tasks. This absence of supplementary data can limit the ability of researchers to fine-tune and enhance LLMs specifically for the tasks included in SciAssess. Consequently, the benchmark results might reflect the inherent capabilities of the models rather than their optimized performance for each specific domain.

Lastly, while SciAssess aims to provide a rigorous evaluation framework, the complexity and diversity of scientific domains present challenges in ensuring comprehensive coverage and fairness. Some tasks may inherently favor certain types of models or architectures, leading to potential biases in performance evaluation.

Broader Impact

Our work on benchmarking scientific literature analysis aligns with the scope of existing LLM benchmarks such as MMLU-pro. This paper represents progress in calibrating LLMs for specific domains, thereby amplifying the impacts that LLM benchmarks have had (and will continue to have) on the broader world. Additionally, we have not identified any ethical concerns or potential risks associated with this work.

References

Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using GPT-4. *CoRR*, abs/2311.07361.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Pierre Caron and T Khan. 1983. Improvement of creep strength in a nickel-base single-crystal superalloy by heat treatment. *Materials Science and Engineering*, 61(2):173–184.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. 2024. Sciknoweval: Evaluating multi-level scientific knowledge of large language models.

Gemini Team Google. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, Qianyu He, Rui Xu, Wenhao Huang, Jingping Liu, Zili Wang, Shusen Wang, Weiguo Zheng, Hongwei Feng, and Yanghua Xiao. 2024. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *AAAI*, pages 18099–18107. AAAI Press.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*. OpenReview.net.

José Luis Hernández-Rivera, Esperanza Elizabeth Martínez Flores, Emmanuel Ramírez Contreras, Jorge Garcia Rocha, Jose de Jesus Cruz-Rivera, and Gabriel Torres-Villasenor. 2017. Evaluation of hardening and softening behaviors in zn–21al–2cu alloy processed by equal channel angular pressing. *Journal of Materials Research and Technology*, 6(4):329–333.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *CoRR*, abs/2305.08322.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Dong-Cho Kim, Tomo Ogura, Ryosuke Hamada, Shotaro Yamashita, and Kazuyoshi Saida. 2021. Prediction of reversible α/γ phase transformation in multi-pass weld of fe-cr-ni ternary alloy by phase-field method. *Journal of Advanced Joining Processes*, 4:100067.

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. CMMLU: measuring massive multitask language understanding in chinese. *CoRR*, abs/2306.09212.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. Halueval: A large-scale hallucination evaluation benchmark for large language models. *Preprint*, arXiv:2305.11747.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL (1)*, pages 3214–3252. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024. Openai o1.

Sizhuo Ouyang, Xinzhi Yao, Yuxing Wang, Qianqian Peng, Zhihan He, and Jingbo Xia. 2022. Text mining task for "gene-disease" association semantics in chip 2022. In *China Health Information Processing Conference*, pages 3–13. Springer.

Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, Murad Mohammad, and Chitta Baral. 2022. Inboxbart: Get instructions into biomedical multi-task learning. In *NAACL-HLT (Findings)*, pages 112–128. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas

Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023a. Challenging big-bench tasks and whether chain-of-thought can solve them. In *ACL (Findings)*, pages 13003–13051. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023b. Challenging big-bench tasks and whether chain-of-thought can solve them. In *ACL (Findings)*, pages 13003–13051. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Francesca Villa, Adelaide Nespoli, Carlo Fanciulli, Francesca Passaretti, and Elena Villa. 2020. Physical characterization of sintered nimnga ferromagnetic shape memory alloy. *Materials*, 13(21):4806.

David Wadden, Kejian Shi, Jacob Morrison, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, et al. 2024. Sciriff: A resource to enhance language model instruction-following over scientific literature. *arXiv preprint arXiv:2406.07835*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Jiao Li, Thomas C. Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database*, 2016.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Mohd Zaki, NM Krishnan, et al. 2023. Mascqa: A question answering dataset for investigating materials science knowledge of large language models. *arXiv preprint arXiv:2308.09115*.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh T. N. Nguyen, Lauren T. May, Geoffrey I. Webb, and Shirui Pan. 2023. Large language models for scientific synthesis, inference and explanation. *CoRR*, abs/2310.07984.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *CoRR*, abs/2304.06364.

A Question Type

Five types of questions, as illustrated in Figure 5 are devised to evaluate the models. Each question type is accompanied by a detailed description and representative examples, along with the corresponding metrics used for assessment. For convenience, the input in each example is simplified, and its instruction is omitted.

B General Prompt Template

We design following general prompt template for scientific literature analysis. It consists of: a system message defining the role of the assistant, the task description, some optional few-shot examples, and a user prompt of the question.

Prompt Template

Role setting and task description:
You are a highly intelligent assistant who answers the following multiple choice question correctly.

Few-shot examples:
Question: <question 1>
Answer: <answer 1>
...
Question: <question n>
Answer: <answer n>

Question:
Predict the number of lines in the EPR spectrum of a solution of 13C-labelled methyl radical (13CH3•), assuming the lines do not overlap.
a) 4
b) 3
c) 6
d) 24

C Task Prompt

In this section, we detail the prompt templates for all tasks in SciAssess benchmark. We will introduce these templates in the following order: Biology (Section C.1), Chemistry (Section C.2), Material (Section C.3) and Medicine (Section C.4).

C.1 Biology

C.1.1 MMLU-Pro-Biology

Prompt

System Message:
You are a highly intelligent assistant who answers the following multiple choice question correctly. Use chain of thought reasoning and provide reasoning before selecting the correct answer (e.g., a)xxx, or b)xxx).
Format:
Reasoning: [Reasoning]
Answer: [Answer]

User Message:
Which of the following would most likely provide examples of mitotic cell divisions?
a) cross section of muscle tissue
b) longitudinal section of a shoot tip
c) longitudinal section of a leaf vein
d) cross section of a fruit
e) cross section of a leaf
f) longitudinal section of a petal
g) longitudinal section of a seed
h) cross section of an anther (site of pollen production in a flower)

Expected Answer:
b) longitudinal section of a shoot tip

C.1.2 Biology Chart QA

The analysis and understanding of biological properties, compositions, and processing techniques are critical for the discovery and development in life sciences. Often, this information is presented in charts, making it essential to extract and integrate such information with textual data. To assess the retrieval capabilities of models in the context of biological chart information, we have designed multiple-choice questions.

Prompt

System Message:
You are an expert in the field of Biomedical. You are a highly intelligent biology scientist who answers the following multiple choice question correctly. Use chain of thought reasoning and provide reasoning before selecting the correct answer (e.g., a)xxx, or b)xxx).
Format:
Reasoning: [Reasoning]
Answer: [Answer]

User Message:
In Figure 3, which has a higher accurate score, with the graph encoder or without?
a) with graph encoder
b) w/o graph encoder

Expected Answer:
a) with graph encoder

C.1.3 Chemical Entities Recognition

This task involves recognizing chemical entity names using data from B5CDR (Wei et al., 2016) and additional expert-annotated data. It evaluates the performance of LLMs in identifying complex drug names. The prompt template is as follows.

True/False

True/False questions offer a straightforward yet effective means of evaluating a model's basic comprehension and ability to verify factual accuracy. They require models to provide a binary True/False response to a certain statement based on their understanding of the information or their own knowledge.

Example

Input: Does "CNS(=O)(=O)CC1=CC2=C(CC=C1)NC=C2CCN(C)C" appear in the document?

Output: Yes (True)

Metric: Accuracy

Multiple Choice

Multiple Choice questions assess a model's ability to select the correct answer from a set of options, testing its knowledge and reasoning. These questions range from basic factual inquiries to more complex scenarios requiring understanding and analyzing information.

Example

Input: A 0.217 g sample of HgO (molar mass = 217 g) reacts with excess iodide ions according to the reaction shown above. Titration of the resulting solution requires how many mL of 0.10 M HCl to reach equivalence point? A. 1mL B. 10mL C. 20mL D. 50mL

Output: C

Metric: Accuracy

Table Extraction

Table Extraction tasks are designed to evaluate a model's proficiency in extracting, summarizing and structuring data from understanding and memorizing given article. Model is usually given an article and is required to collect certain information from it, and presented in the form of a table.

Example

Input: *Given an article* While reading this paper, please summarize a complete list of the names and abbreviations of all solutions.

Output: (The original return is in CSV format, which is converted to intuitive display here)

Compound	Solvent	Solution	aFC[a]	aFEMC	aFC/aFEMC	XFC/XFEMC[b]
LiPF6	FEC:FEMC	1:04:04 LiPF6:FEC:FEMC	0.56	0.44	1.26	1.42
LiPF6	DFEC:FEMC	1:04:04 LiPF6:DFEC:FEMC	0.13	0.48	0.28	0.23

Metric: Table Recall (The recall rate for each value, which checks whether every cell value in the expected table appears in the output table.)

Text Extraction

Text Extraction tasks are designed to evaluate a model's ability to extract information from the text. The information can be entities or triplets (entity-relationship or event).

Example (Chemical Entities Recognition)

Input: unanesthetized, spontaneously hypertensive rats the decrease in blood pressure and heart rate produced by intravenous clonidine, 5 to 20 micrograms/kg, was inhibited or reversed by nalozone, 0.2 to 2 mg/kg. The hypotensive effect of 100 mg/kg alpha-methyldopa was also partially reversed by naloxone. As naloxone and clonidine do not appear to interact with the same receptor site, the observed functional antagonism suggests the release of an endogenous opiate by clonidine or alpha-methyldopa and the possible role of the opiate in the central control of sympathetic tone.

Output: (naloxone), (clonidine), (nalozone), (alpha-methyldopa)

Metric: F1-score (calculated based on fuzzy matching using Levenshtein Distance and optimal assignment via the Kuhn-Munkres algorithm. Details and implementation can be found in our repository.)

Molecule Generation

Molecule Generation tasks are designed to evaluate a model's ability to generate molecules in SMILES format.

Example (Tag to Molecule)

Input: What's the SMILES formula of molecule "Sumatriptan"?

Output: CNS(=O)(=O)CC1=CC2=C(C=C1)NC=C2CCN(C)C

Metric: Molecular Similarity (Molecule similarity is measured by the Tanimoto similarity between two molecules' Morgan fingerprints)

Figure 5: Question types.

Prompt

System Message:

You are an expert in the field of Biomedical. I'll give you the abstract of literature. Please use chain of thought reasoning to identify all the compound entities in the abstract. First, analyze the abstract step by step, explaining your reasoning for identifying each compound entity. Then, provide a final list of the compound entities you recognized in the format: (compound 1), (compound 2), (compound 3).
Format:
Reasoning: [Reasoning]
Answer: [List of identified compounds]

User Message:

In unanesthetized, spontaneously hypertensive rats the decrease in blood pressure and heart rate produced by intravenous clonidine, 5 to 20 micrograms/kg, was inhibited or reversed by nalozone, 0.2 to 2 mg/kg. The hypotensive effect of 100 mg/kg alpha-methyldopa was also partially reversed by naloxone. Naloxone alone did not affect either blood pressure or heart rate. In brain membranes from spontaneously hypertensive rats clonidine, 10(-8) to 10(-5) M, did not influence stereoselective binding of [3H]-naloxone (8 nM), and naloxone, 10(-8) to 10(-4) M, did not influence clonidine-suppressible binding of [3H]-dihydroergocryptine (1 nM). These findings indicate that in spontaneously hypertensive rats the effects of central alpha-adrenoceptor stimulation involve activation of opiate receptors. <rest of the input>.

Expected Answer:

(naloxone), (clonidine), (nalozone), (alpha-methyldopa)

C.1.4 Compound Disease Recognition

Proposed in B5CDR (Wei et al., 2016), this task evaluates the capability of LLMs to identify and understand associations between compounds and diseases. Examples of process text:

Example Paragraph

Twenty children with acute lymphoblastic leukemia who developed meningeal disease were treated with a high-dose intravenous methotrexate regimen that was designed to achieve and maintain CSF methotrexate concentrations of 10(-5) mol/L without the need for concomitant intrathecal dosing. The methotrexate was administered as a loading dose of 6,000 mg/m2 for a period of one hour followed by an infusion of 1,200 mg/m2/h for 23 hours. Leucovorin rescue was initiated 12 hours after the end of the infusion with a loading dose of 200 mg/m2 followed by 12 mg/m2 every three hours for six doses and then every six hours until the plasma methotrexate level decreased to less than 1 X 10(-7) mol/L. The mean steady-state plasma and CSF methotrexate concentrations achieved were 1.1 X 10(-3) mol/L and 3.6 X 10(-5) mol/L, respectively. <rest of the paragraph>.

We then prompt the model with the following:

Prompt

System Message:
You are a biologist AI. I'll give you the abstract of literature. Please use chain of thought reasoning to identify all the (compound, disease) relations in the abstract. First, analyze the abstract step by step, explaining your reasoning for identifying each relation. Then, provide a final list of the relations in the format: '(compound 1, disease 1),(compound 2, disease 2),(compound 3, disease 3),....', without adding any additional comments or explanations.
Format:
Reasoning: [Reasoning]
Answer: [List of recognized relations]

User Message:
[processed text]

Expected Answer:
(methotrexate, transient hemiparesis), (methotrexate, neutropenia), (methotrexate, seizures), (methotrexate, mucositis)

C.1.5 Disease Entities Recognition

Similarly, this task involves recognizing disease entity names using data from (Wei et al., 2016) and additional expert-annotated data, evaluating the performance of LLMs in identifying specialized disease names:

Prompt

System Message:
You are an expert in the field of Biomedical. You are a biologist AI. I'll give you the abstract of literature. Please use chain of thought reasoning to identify all the disease entities in the abstract. First, analyze the abstract step by step, explaining your reasoning for identifying each disease entity. Then, provide a final list of the disease entities you recognized in the format: (disease 1), (disease 2), (disease 3).
Format:
Reasoning: [Reasoning]
Answer: [List of recognized diseases]

User Message:
In unanesthetized, spontaneously hypertensive rats the decrease in blood pressure and heart rate produced by intravenous clonidine, 5 to 20 micrograms/kg, was inhibited or reversed by nalozone, 0.2 to 2 mg/kg. The hypotensive effect of 100 mg/kg alpha-methyldopa was also partially reversed by naloxone. Naloxone alone did not affect either blood pressure or heart rate. In brain membranes from spontaneously hypertensive rats clonidine, 10(-8) to 10(-5) M, did not influence stereoselective binding of [3H]-naloxone (8 nM), and naloxone, 10(-8) to 10(-4) M, did not influence clonidine-suppressible binding of [3H]-dihydroergocryptine (1 nM). These findings indicate that in spontaneously hypertensive rats the effects of central alpha-adrenoceptor stimulation involve activation of opiate receptors. <rest of the input>.

Expected Answer:
(hypertensive), (hypotensive)

C.1.6 Gene Disease Function

The Gene Disease Text Mining task focuses on "Gene-Disease" association semantics text mining. It evaluates the ability of models to extract and understand relationships between genes and diseases from scientific literature, with a focus on identifying gene and disease entities (Ouyang et al., 2022). Examples of process text:

Example Paragraph

A novel frameshift mutation (+G) at codons 15/16 in a beta0 thalassaemia gene results in a significant reduction of beta globin mRNA values.
AIMS: To identify a novel beta globin gene mutation found in a Chinese family, and also to assess its functional consequences.
METHODS: Haematological analysis was performed on all family members. The 23 common mutations of beta thalassaemia found in Chinese populations were detected by means of a reverse dot blot method. Direct DNA sequencing of polymerase chain reaction (PCR) amplified complete beta globin gene was carried out to identify the novel mutation. A real time, one step reverse transcription PCR assay was used to measure beta globin mRNA in the reticulocytes of heterozygous patients.
RESULTS: A novel frameshift mutation-an insertion of G between codons 15 and 16 in a homonucleotide run of four guanines-was determined, which generates a new premature chain terminator at the 22nd codon. Relative quantitative analysis of the beta globin mRNA in heterozygous subjects demonstrated a 39.83% reduction compared normal controls.
CONCLUSIONS: The significantly lower amounts of beta globin mRNA found in mutation carriers is probably caused by the rapid nonsense mediated degradation of the mutant mRNA. These data, combined with haematological analysis, suggest that this novel mutation of CDs 15/16 (+G) results in a beta(0) thalassaemia phenotype.

For extracting triplets (entities, semantic roles, entities), we prompt the model with:

Prompt

System Message:
You are an expert in the field of Biomedical. In this semantic role regconition task, you need to follow 3 steps, and finally just return me triples that needed. First, you need to identify the entities in the text. Entities can be classified into 2 categories–molecular, and trigger word. 'Molecular' includes disease, gene, protein, and enzyme. 'Trigger word' includes:
1)Variation(Var), which means DNA, RNA, and mutations in proteins and changes in molecular structure, e.g. 'mutations on the Arg248 and Arg282', 'mutant R282W', 'missense mutations';
2)Molecular Physiological Activity (MPA), including molecular activity, gene expression and molecular physiological activity, e.g. 'phosphorylation', 'transcription', 'histone methylation', 'bioactivation of cyclophosphamide';
3)Interaction, molecule-to-molecule or molecule-to-cell connections, e.g. 'bind', 'interaction';
4)Pathway, e.g. 'Bmp pathway','PI3K pathway';
5)Cell Physiological Activity (CPA), Activities at or above the cellular level, including cellular reactivity and cell or organ development and growth, e.g. 'T helper cell responses', 'renal development';
6)Regulation (Reg), a neutral cue word or phrase meaning no loss or gain, e.g. 'resolved in', 'regulated';
7)Positive Regulation (PosReg), a cue word or phrase that indicates the acquisition of a function, e.g. 'facilitates', 'enhanced', 'increased';
8)Negative Regulation (NegReg), a clue word or phrase that indicates a loss of function, e.g. 'suppressed', 'decreased', 'inhibited'.
Second, you need to identify the semantic role labeling objects, including 'ThemeOf'(from the main thing entity to the current entity) and 'CauseOf'(From the current entity to the Cause entity).
Third, please give me tripples that contain entities and semantic role labeling objects(ThemeOf or Causeof).
Use chain of thought reasoning to explain your process of identifying the entities and relations, and then provide the final triples in the format: (. . .), (. . .)
Format:
Reasoning: [Reasoning]
Answer: [List of recognized triples]

User Message:
[processed text]

Expected Answer:
(frameshift, CauseOf, reduction), (caused by, CauseOf, lower), (mutation, CauseOf, results in), (beta(0) thalassaemia, ThemeOf, results in), (beta globin mRNA, ThemeOf, reduction), (beta0 thalassaemia gene, ThemeOf, frameshift), (insertion, CauseOf, generates), (premature chain terminator, ThemeOf, generates), (amounts of beta globin mRNA, ThemeOf, lower), (mutation, CauseOf, caused by), (degradation, ThemeOf, caused by).

C.2 Chemistry

C.2.1 MMLU-Pro-Chemistry

The example of MMLU-Pro-Chemistry is similar to the MMLU-Pro-Biology task in Appendix C.1.1.

C.2.2 Electrolyte Table QA

The composition and properties of organic electrolytes are crucial for battery performance, stability, and safety. To evaluate the model's retrieval capabilities regarding electrolyte information, we posed multiple-choice questions about the components of solution systems and the dissolution reactions, focusing on their physical and chemical properties as presented in the tables within the articles. We prompt the model with the following:

Prompt

System Message:
You are an expert in the electrolytes field. Please answer the following multiple choice question correctly. Use chain of thought reasoning and provide reasoning before selecting the correct answer (e.g., a)xxx, or b)xxx).
Format:
Reasoning: [Reasoning]
Answer: [Answer]

User Message:
In the upper paper, what are the minimum and maximum intramolecular distances (nm) of dimethyl carbonate?
a) 0.41/0.87
b) 0.49/0.67
c) 0.25/0.25
d) 0.25/0.38

Expected Answer:
a) 0.41/0.87

C.2.3 OLED Property Extraction

This task evaluates the LLM's ability to extract information about OLED molecules and their optical properties. It tests several key capabilities, including their understanding of complex and domain-specific language and their ability to interpret and extract data from tables. An example output is shown in Table 5. We prompt the model with the following:

Prompt

System Message:
You are an expert in the field of organic photovoltics. Please give a complete list of Host, Host's SMILES structure (if exists), Dopant, Assistant Dopant (if exists), Td/Tg/ET, Von,max EQE/CE/PE,EQE/CE/PE, and CIE [x, y]
* Output in csv format with columns of those attributes, do not write units only the value like "10.5".
* Quote the column name or Host's Name or Dopant's Name if it contains space or special characters like ",".
* If there are multiple tables, concat them. Don't give me reference or using "...", give me complete table!
* Should return all columns mentioned, if empty just return 'NaN'. "Host" and "Dopant" should not be empty.
* "Host" and "Dopant" should be short name of the organic molecule.
* Should find more information from the whole content, including tables, text.
for example, you should return:
‘‘‘csv
Host,SMILES,Dopant,Td /Tg /ET,
Von,max EQE/CE/PE,EQE/CE/PE,"CIE"
PPO1,O=P(c1ccccc1)(c1ccccc1)c1ccccc1,FCNIr,–/74/3.02,–,
17.1/20.5/14.3,–/–/–,"(0.14, 0.16)"
PPO2,O=P(c1ccccc1)(c1ccccc1)c1ccccc1,FCNIr,–/123/3.02,–,
18.4/21.1/16.6,–/–/–,"(0.14, 0.15)"
‘‘‘
Please use a step-by-step approach to analyze the content and ensure that all relevant information is accurately extracted. Only provide reasoning for how you identified each attribute and output the final csv format.
Format:
Reasoning: [Reasoning]
Answer: [Extracted csv]

User Message:
[docoment.pdf]

C.2.4 Polymer Chart QA

The processing steps and properties of polymer materials are often represented through charts. Extracting information from these charts and integrating it with textual data is crucial. To further assess the retrieval capabilities of models concerning polymer chart information, we designed multiple-choice questions involving polymer composition, processing techniques, and properties.

The example of Polymer Chart QA can be found in a similar format to the Biology Chart QA task in Appendix C.1.2.

C.2.5 Polymer Composition QA

This task involves extracting the blend ratio of donor to acceptor in the most efficient solar cell from the text of scientific literature.

We prompt the model with the following:

Table 5: OLED Property example.

Host	Dopant	Td [°C] / Tg [°C] / ET [eV]	Von [V]	max EQE [%] / CE [cd A-1] / PE [lm W-1]	EQE [%] / CE [cd A-1] / PE [lm W-1]	CIE [x, y]
PPO1	FCNIr	– / 74 / 3.02	–	17.1 / 20.5 / 14.3	– / – / –	(0.14, 0.16)
PPO2	FCNIr	– /123 / 3.02	–	18.4 / 21.1 / 16.6	– / – / –	(0.14, 0.15)
mCPPO1	FCNIrpic	– /– / 3.00	–	25.1 / – / 29.8	23.1 / 28.9 / 15.1	(0.14, 0.18)
CDPO	5CzCN	455 / 89 / 2.84	4.9	13.2 / 31.6 / 18.1	– / – / –	(0.20, 0.38)

Prompt

System Message:
You are an expert in the field of polymer solar cells researcher who answers the following multiple choice question correctly. Use chain of thought reasoning and provide reasoning before selecting the correct answer (e.g., a)xxx, or b)xxx).
Format:
Reasoning: [Reasoning]
Answer: [Answer]

User Message:
In this paper, What is the blend ratio of donor to acceptor in the most efficient solar cell?
a) 1:4
b) 20:8
c) 30:50
d) 2:4

Expected Answer:
a) 1:4

C.2.6 Polymer Property Extraction

This task focuses on extracting vital values such as power conversion efficiency (PCE) and open-circuit voltage (V_{OC}) from tables within the literature.

We prompt the model with the following:

Prompt

System Message:
You are an expert in the field of polymer solar cells researcher.
Please give a complete list of Nickname, PCE_max, PCE_ave, Voc , Jsc, FF; * Output in csv format with columns of those attribution, do not write units only the value like "10.5".
* If there are multiple tables, concat them. Don't give me reference or using "...", give me complete table!
* Should return all columns mentioned, if empty just return 'NaN'. Nickname should not be empty.
* Nickname should be short name of polymers, for example: 'PCBM:PffBT4T-2OD:PC61PM' should return 'PffBT4T-2OD'.
* Only return acceptor 'PC71BM' related records.
* If with different experiment settings for the same nickname, only return the record with 'highest PCE' !
* Should find more information from the whole content, including tables, text.
* For FF use 0.xx instead of xx.x, for example: 63.0 should return 0.63 !
for example, you should return:
‘‘‘csv
Nickname,PCE_max(%),PCE_ave(%),Voc (V),Jsc (mA cm^2),FF
PBTTT-C14,2.34,2.34,0.53,9.37,0.48
‘‘‘
Please use a step-by-step approach to analyze the content and ensure that all relevant information is accurately extracted. Only provide reasoning for how you identified each attribute and output the final csv format.
Format:
Reasoning: [Reasoning]
Answer: [Extracted csv]

User Message:
[docoment.pdf]

C.2.7 Solubility Extraction

Organic electrolytes, extensively used in battery technologies, comprise organic solvents, lithium salts, and additives. Understanding solubility in organic electrolytes is crucial as it impacts the efficiency of electrolytic processes, product selectivity, and equipment design. This task evaluates the LLM's capability in retrieving solubility-related tables. Papers typically select data from various aspects to describe the system, making it challenging to combine multiple tables for fuzzy matching. Therefore, we focus on examining the LLM's semantic understanding ability, enabling the model to select the most relevant and comprehensive table related to "solubility" from numerous alternatives and convert it into the specified format.

We prompt the model with the following:

Prompt

System Message:
You are an expert in the field of chemistry and specialize in the study of solubility. Now you are required to extract tables related to solubility from the article. The extracted information includes solute name, solvent name, temperature, pressure and solubility. Since these properties are temperature-dependent and pressure-dependent, please place the properties at different temperatures or pressure on different rows. The values of temperature and solubility should be output together with their unit. Output the whole table in csv format and satisfy these requirements:
(1) Do not truncate tables using "...". Always output the complete tables.
(2) Keep all the superscripts in the form like "^3", "^+" or "^a".
(3) Do not use "NaN" to replace the blank cells, just leave it empty.
(4) Use "x" to replace all "×", Use "()" to replace all " () "
(5) Always add space before and after operators like " ± ".
As a example, the csv should be like:
‘‘‘csv
solute_name,solvent_name,temperature,pressure,solubility
FLBDOB,PC,298.2 K,1 atm,0.275 ± 0.1 mol/L
‘‘‘
Please use a step-by-step approach to analyze the content and ensure that all relevant information is accurately extracted. Only provide reasoning for how you identified each attribute and output the final csv format.
Format:
Reasoning: [Reasoning]
Answer: [Extracted csv]

User Message:
[docoment.pdf]

C.2.8 Reactant QA

Organic and bio-catalyzed synthetic reactions are vital for the manufacture of drug-like molecules. Therefore, we designed a complex task to test the model's capability in extracting information from schematic diagrams and texts of chemical reactions. The model is required to understand the charts specified in the articles and select the correct answer from the provided multiple-choice descriptions.

Prompt

System Message:
You are an expert in the field of organic chemistry. Use chain of thought reasoning and provide reasoning before selecting the correct answer (e.g., a)xxx, or b)xxx).
Format:
Reasoning: [Reasoning]
Answer: [Answer]

User Message:
Which compound is in the reactants or reagents of the following reaction?
a) c4ccc(B3OB(c1ccccc1)OB(c2ccccc2)O3)cc4
b) O=C(C(C)C(OC)=O)C1CC1
c) CC(=O)OP(=O)([O-])[O-].[NH4+].[NH4+]
d) COC(=O)/C(C)=C(/OS(=O)(=O)c1ccc(C)cc1)C1CC1
The new reaction you should deal with is the second step in the first reaction in Section "2.2 Procedures".

Expected Answer:
b) O=C(C(C)C(OC)=O)C1CC1

C.2.9 Reaction Mechanism QA

Investigating electrolyte reactions helps improve the solid electrolyte interphase (SEI) layer, which directly affects battery performance and lifespan. Studies in this area lead to the development of advanced electrolytes that enhance a robust SEI, resulting in more efficient and durable batteries. We design a complex task to test the capability of extracting information from schematic diagrams of chemical reaction mechanisms. LLM is required to understand the specified reaction diagram and select the correct answer from the provided multiple choices.

We prompt the model with the following:

Prompt

System Message:
You are a highly intelligent organic electrolyte researcher who answers the following multiple choice question correctly. Use chain of thought reasoning and provide reasoning before selecting the correct answer (e.g., a)xxx, or b)xxx).
Format:
Reasoning: [Reasoning]
Answer: [Answer]

User Message:
According to figure 1, which one of these synthetic routes for LTFOP is correct?
a) DTMSO + LiPF6 -> LTFOP + 2 CH3)3SiF
b) 2 DTMSO + LiPF6 -> LTFOP + 4 CH3)3SiF
c) HOOCCOOH + 2/3 CH3)3SiCl + 2/3 CH3)3SiNH)SiCH3) -> LTFOP + 2/3 NH4Cl
d) DTMSO + LiPCl6 -> LTFOP + 2 CH3)3SiCl

Expected Answer:
a) DTMSO + LiPF6 -> LTFOP + 2 CH3)3SiF

C.3 Material

C.3.1 Material QA

The example of Material QA is similar to the MMLU-Pro-Biology task in Appendix C.1.1.

C.3.2 Alloy Chart QA

The processing steps and properties of alloy materials are often presented in charts, such as those comparing the performance of multiple alloys or illustrating how elongation changes with composition. Therefore, extracting information from these charts and integrating it with textual information is crucial. To further evaluate the retrieval capability of models regarding alloy chart information, we have designed multiple-choice questions involving alloy composition, processing techniques, and properties.

The example of Alloy Chart QA is similar to the Biology Chart QA task in Appendix C.1.2.

Alloy	Comp. C	Comp. Cr	Comp. Cu	Comp. Fe	Comp. Mn	Comp. Mo
LeanDSS	0.014%	20.85%	0.09%	73.38%	1.49%	0.30%
StandardDSS	0.012%	22.46%	0.17%	69.94%	1.81%	3.07%
SuperDSS	0.013%	24.98%	0.20%	63.41%	0.48%	4.03%

Table 6: Alloy composition example. **Comp.**: Composition.

C.3.3 Composition Extraction

Extracting alloy composition information from an article's text or tables and unifying it into a structured format helps researchers utilize historical data more effectively and provides valuable guidance for subsequent designs. This comprehensive task evaluates LLMs' ability to extract alloy compositions (including all element contents) from text and tables. Typically, alloy element content is found in two cases: (1) the element content is stored in a table, and (2) the element content is implicitly indicated by the alloy name, such as 'Fe30Co20Ni50', which represents an atomic ratio of 30% Fe, 20% Co, and 50% Ni. The objective of this task is to comprehensively extract this information and organize it into a digestible table. The metric is to calculate the matching score between the standard answer table and the extraction result table. This task showcases the LLM's comprehension ability to integrate, extract, and structure multi-modal information (Kim et al., 2021).

An alloy composition table example is shown as following:

Prompt

System Message:
You are an expert in the field of Alloy Materials. Please give a complete list of alloy names and compositions of all alloys in this paper.
If there is no alloy composition element ratio in the text, try to extract the element ratio from the alloy name from the perspective of alloy experts.
Output in csv format with multiindex (2 headers), The names in first header are 'AlloyName' and 'Composition' forcely. The names in second header are element names of alloy.
Starting on the third row, list the alloy names and their corresponding element content. Based on the number of reference commas, the element name corresponds to the content.
Please write units not in header but in value like "50 wt.%","30 at.%".
Output the data strictly in the CSV format shown below and exclude any other content. Example format:
"'csv
AlloyName,Composition,Composition,Composition
nan,Fe,Co,Al
Fe70Co15Al3,70 wt.%,15 wt.%,3 wt.%
Fe70Co18,70 wt.%,18 wt.%,nan
"'
Please use a step-by-step approach to analyze the content and ensure that all relevant information is accurately extracted. Only provide reasoning for how you identified each attribute and output the final csv format.
Format:
Reasoning: [Reasoning]
Answer: [Extracted csv]

User Message:
[docoment.pdf]

C.3.4 Temperature QA

The properties of an alloy are determined by its composition and the processes it undergoes, including processing and heat treatment. Therefore, extracting heat treatment values is critical. This task aims to determine the maximum temperature value for the heat treatment of the alloy. To ensure easy statistical analysis, questions are designed as multiple-choice. Examples of process paragraphs (Villa et al., 2020):

Example Paragraph

Cast NiMnGa samples, of Ni50Mn30Ga20 nominal composition, were prepared by 5 arc melting cycles of the pure elements (electrolytic Ni 99.97%, electrolytic Mn 99.5% and Ga 99.99%) in stoichiometric ratio, in a non-consumable electrode furnace (Leybold LK6/45) (Leybold, Cologne, Germany). The as-cast ingot was ground to powder in a planetary ball mill (Fritsch Pulverisette 4) (FritschIdar-Oberstein, Germany) and the powder size was selected by means of sieves. Densified pellets were produced by die-pressing alloy powders with different average sizes (lower than 50 um or between 50 and 100 um) at 0.75 GPa at room temperature and sintered by thermal treatment at 925 °C for 24, 72, and 168 h in an Ar atmosphere, followed by slow cooling in the furnace. Sintered pellets had the following dimensions: approximately 3 mm in height and 13 mm in diameter. Table 1 provides a summary of the prepared sintered samples.

We prompt the model with the following:

Prompt

System Message:
You are an expert in the field of Alloy Materials. You are a highly intelligent alloy researcher who answers the following multiple choice question correctly. Use chain of thought reasoning and provide reasoning before selecting the correct answer (e.g., a)xxx, or b)xxx).
Format:
Reasoning: [Reasoning]
Answer: [Answer]

User Message:
In the upper paper, what is the maximum temperature of the heat treatment process for all alloys?
a) 925 C
b) 650 C
c) 700 C
d) 800 C

Expected Answer:
a) 925 C

C.3.5 Sample Differentiation

Alloys with the same composition but treated by different processes are considered different samples because they exhibit different properties. Therefore, distinguishing between different samples and understanding the differences in their processes is essential. This multiple-choice question task is designed to comprehensively judge the number of different alloy samples proposed or studied by the authors. It assesses the LLMs' analysis and reasoning abilities regarding alloy distinctions from text.

The following example is process paragraphs where the sample are treated by different processes (Hernández-Rivera et al., 2017):

Example Paragraph

An induction furnace was used to produce the Zn-21A1-2Cu alloy by melting proper amounts of Zn (99.99%), Al (99.99%), and Cu (99.96%). The alloy was melted in a graphite crucible exposed to air and poured into cylindrical bars of 19 mm in diameter and 35 mm in length. After that, some bars were homogenized at 350 °C for 24h in the air. Cast and homogenized samples were subjected to an equal channel angular extrusion(ECAP) in a die with two cylindrical channels with a diameter of 15.8mm. The inner intersecting angle (y) was 90 and the outer angle (y) was 36°. All samples were extruded by two and sixpasses with a ram velocity of 5 mm/min and by using B. route. The lubricant used was MoS, and it was applied to both channels on each pass.

We prompt the model with the following:

Prompt

System Message:
You are an expert in the field of Alloy Materials. Please answer the following multiple choice question correctly. Use chain of thought reasoning and provide reasoning before selecting the correct answer (e.g., a)xxx, or b)xxx).
Format:
Reasoning: [Reasoning]
Answer: [Answer]

User Message:
Materials with the same components but processed through different techniques are considered as different alloys because they possess distinct properties. In the upper paper, please provide a count of all the alloys proposed and discussed by the authors?
a) 2
b) 0
c) 3
d) 1

Expected Answer:
d) 1

C.3.6 Treatment Sequence

Each alloy treatment process has a clear sequence requirement, so it is necessary to ensure that the extracted heat treatment process sequence is consistent with the experimental sequence. For example, after solution treatment, a sample is further aged to ensure the release of internal stresses. This task aims to objectively analyze and evaluate the sequential relationship between two heat treatments and provide True/False answers. Additionally, if a specific heat treatment name does not exist in the paper, it should be considered False. This task assesses the LLM's comprehension ability to judge treatment order from the text.

Prompt

System Message:
You are an expert in the field of Alloy Materials. You are a specialist in the domain of heat treatment processes, such as homogenization, annealing, aging, solution treatment, quenching, and tempering, among others. Use chain of thought reasoning to analyze the question step by step. After your reasoning, answer the question with 'Yes' or 'No'.
Format:
Reasoning: [Reasoning]
Answer: [Yes/No]

User Message:
In the upper paper, is the processing heat treatment technique before the thermal treatment at 925 C called arc melting?

Expected Answer:
Yes

C.4 Medicine

C.4.1 MMLU-Pro-Health

The example of MMLU-Pro-Health is similar to the MMLU-Pro-Biology task in Appendix C.1.1.

C.4.2 Affinity Extraction

This task evaluates the LLM's ability to extract an affinity table containing molecules' tags, SMILES, and their affinities to different targets in bioassays. It tests several key capabilities of LLMs, including understanding complex and domain-specific language, as well as molecules and tables. Affinity data extraction requires not just surface-level text processing but also a deeper analysis to match different modalities.

An example output is shown in Table 7.

We prompt the model with the following:

Prompt

System Message:
You are an expert in the field of pharmaceutical chemistry, and your task is to summarize the results of activity assays from an article in a tabular format. Please follow these steps to complete the task:
1. Determine if the article includes an activity assay. If it does, locate the section(s) presenting the assay results, which are usually in one or more tables.
2. Compile all the activity assay results into a single table. You may use multiple columns to represent different conditions or outcomes of various experiments.
3. Identify the names or codes used in the table, such as Example 1 or Compound A, and find the corresponding sections in the article that mention these substances. Extract the full name and SMILES notation of each substance.
4. Compile the names and SMILES notations of each substance in the table. Output in csv format with multiindex (Affinities, protein/cell line), write units not in header but in the value like "10.5 μM". Quote the value if it has comma! For example:
```csv
Compound,Name,SMILES,Affinities,Affinities,Affinities,Affinities
,,,5HT1A (IC50),5HT1D (IC50),5HT-UT (IC50),5HT1E (<affinity type>)
5a,Aspirin,CC(=O)Oc1ccccc1C(=O)O,2.0 nM,8.0 nM,12.6 nM, >1000 nM
```
5. If there are multiple tables, concat them. Don't give me reference or using "...", give me complete table! Please use a step-by-step approach to analyze the content and ensure that all relevant information is accurately extracted. Only provide reasoning for how you identified each attribute and output the final csv format.
Format:
Reasoning: [Reasoning]
Answer: [Extracted csv]

User Message:
[docoment.pdf]

C.4.3 Drug Chart QA

The analysis of drug properties, compositions, and processing techniques is critical for drug discovery and development. Often, this information is presented in charts, making it essential to extract and integrate such information with textual data. To further assess the retrieval capabilities of models in the context of drug chart information, we have designed multiple-choice questions focusing on drug composition, processing methods, and properties. The example of Drug Chart QA can be found in a similar format to the Biology Chart QA task in Appendix C.1.2.

C.4.4 Tag to Molecule

This task evaluates the model's ability to find the correct SMILES given its tag in a document. Typically, a molecule is shown with an chart of its structure and a tag below it. The LLM should recognize both the structure and the tag and understand their

Compound	Name	SMILES	Affinities	
			Cytotoxicity in 2.2.15 Cells (IC50)	Anti-HBV Activity in 2.2.15 Cells (EC50)
1	/	C1[C@H](O[C@H]([C@H]1F) N2C=NC3=C(N=CN=C32)N)CO	>200000 nM	>10000 nM
2	/	C1[C@H](O[C@H]([C@H]1F) N2C=CC(=NC2=O)N)CO	>200000 nM	4000 nM
3	/	CC1=CN(C(=O)NC1=O)[C@H]2C [C@@H]([C@H](O2)CO)N=[N+]=[N-]	NA	NA

Table 7: Example output of affinity data extraction task

connection.

Prompt

System Message:
You are an expert in the field of organic chemistry, can help user get SMILES formula from documents. Use chain of thought reasoning to analyze the document and extract the SMILES formula step by step. After reasoning, just give the SMILES formula as the final answer without further explanation.
Format:
Reasoning: [Reasoning]
Answer: [Extracted SMILES]

User Message:
What's the SMILES formula of molecule "Sumatriptan"?

Expected Answer:
"CNS(=O)(=O)CC1=CC2=C(C=C1)NC=C2CCN(C)C"

C.4.5 Markush to Molecule

This task evaluates the model's ability to obtain the correct SMILES given a Markush formula (in CXSMILES pattern) and its substituents.

Prompt

System Message:
You are an expert in the field of chemistry, can help user insert substituents into CXSMILES-type markush formula to get SMILES formula (removing Hs). Use chain of thought reasoning to explain how you insert the substituents step by step, ensuring the correct SMILES is generated. After reasoning, just reply with the SMILES formula without further explanation.
Format:
Reasoning: [Reasoning]
Answer: [Generated SMILES]

User Message:
C(*)CC(*)CC |$A;;Pol_p;;;Q_e;;;M_p$|, A = H, Pol = NH2, Q = OH, M = [Li]

Expected Answer:
"NCCC(O)CC[Li]"

C.4.6 Molecule in Document

This task evaluates the model's ability to determine whether a molecule (represented by SMILES) is mentioned in a document. The LLM should recognize all Markush formulas and their substituents, and then judge whether the required molecule is covered.

Prompt

System Message:
You are an expert in the field of chemistry. You are given a SMILES formula of a molecule, and should judge whether it is in the document. If the molecules are given by Markush formula (containing R group), You need to 1) analyze the skeletons of the provided molecule and the molecule in the literature or patent, and 2) if the compare the variable values of the molecular structure with the range of variable values given in the patent, to determine whether the molecule is covered by the literature or patent. Use chain of thought reasoning to analyze the question step by step. After your reasoning, answer the question with 'Yes' or 'No'.
Format:
Reasoning: [Reasoning]
Answer: [Yes/No]

User Message:
[document.pdf]
Does the molecule "CC(CCCCCCCC1=CC(=C(C(=C1)OC)OC)OC)CCC(C2=CC=CS2)O" appear in the document?

Expected Answer:
Yes

D Performance without Cot

In Table 8, we show the performance comparison of LLMs across various scientific domains.

E Comparison with Existing Benchmarks

We compare SciAssess with two representative benchmarks: (1) MMLU-Pro (Wang et al., 2024), a general-purpose benchmark, and (2) SciKnowEval (Feng et al., 2024), a domain-specific benchmark focused on scientific knowledge. The results are shown in Table 9 and Table 10, respectively.

F Baseline LLMs

We briefly introduce the baseline LLMs and endpoints that we have tested on SciAssess.

- **OpenAI-o1** (OpenAI, 2024) OpenAI's o1 model is designed to reason through complex tasks and solve harder problems in science, coding, and math. The model we tested is `OpenAI-o1-preview`.
- **GPT-4o**[3]: OpenAI's GPT-4o advances human-computer interaction by handling text, audio, image, and video inputs and outputs. It offers improved efficiency and cost compared to previous GPT models. The model we use is `gpt-4o`.

[3] https://openai.com/index/hello-gpt-4o/

Domain	Task	o1-preview	GPT-4o	GPT-4	GPT-3.5	Moonshot	Claude3	Doubao	Gemini	Llama3.1	Qwen2.5	Mixtral
Biology	MMLU-Pro-Biology*	**0.901**	0.824	0.783	0.654	0.748	0.709	0.768	0.826	0.799	**0.802**	0.709
	Biology Chart QA	**0.653**	0.563	0.513	0.377	0.563	0.447	0.482	**0.653**	0.503	**0.533**	0.482
	Chemical Entities Recognition*	**0.862**	0.855	0.845	0.614	0.803	0.826	0.786	0.799	0.824	**0.845**	0.731
	Compound Disease Recognition*	**0.745**	0.659	0.742	0.539	0.679	0.735	0.682	0.733	0.675	**0.700**	0.605
	Disease Entities Recognition*	**0.831**	0.809	0.828	0.697	0.715	0.797	0.789	0.822	**0.814**	0.700	0.783
	Gene Disease Function*	0.687	0.488	0.717	0.523	0.637	0.558	0.649	**0.792**	**0.655**	0.495	0.484
Chemistry	MMLU-Pro-Chemistry*	**0.868**	0.339	0.334	0.228	0.260	0.226	0.342	0.513	0.347	**0.420**	0.314
	Electrolyte Table QA	**0.925**	0.590	0.380	0.170	0.735	0.315	0.640	0.845	0.490	**0.668**	0.337
	OLED Property Extraction	0.394	**0.459**	0.390	0.107	0.165	0.130	0.327	0.365	0.103	**0.371**	0.237
	Polymer Chart QA	**1.000**	0.867	0.600	0.067	0.733	0.133	0.733	0.867	**0.867**	**0.867**	0.800
	Polymer Composition QA	**0.986**	0.756	0.708	0.316	0.967	0.608	0.823	0.952	0.689	**0.894**	0.524
	Polymer Property Extraction	0.606	**0.785**	0.782	0.435	0.705	0.489	0.524	0.701	0.590	**0.689**	0.559
	Solubility Extraction	0.427	0.508	**0.516**	0.326	0.476	0.396	0.407	0.454	0.443	**0.448**	0.303
	Reactant QA	**0.559**	0.374	0.359	0.251	0.256	0.226	0.241	0.338	0.277	**0.472**	0.205
	Reaction Mechanism QA	0.682	0.545	0.500	0.091	0.591	0.409	0.409	**0.773**	**0.682**	0.455	0.409
Material	Material QA	**0.821**	0.757	0.707	0.521	0.586	0.544	0.665	0.715	0.684	**0.738**	0.654
	Alloy Chart QA	0.533	0.467	0.533	0.333	0.333	0.467	**0.733**	0.667	0.467	**0.667**	0.600
	Composition Extraction	**0.488**	**0.488**	0.441	0.100	0.351	0.360	0.336	0.423	0.424	**0.433**	0.198
	Temperature QA	0.836	0.609	0.536	0.261	0.836	0.295	0.425	**0.879**	**0.594**	0.507	0.353
	Sample Differentiation	0.392	0.245	0.333	0.089	**0.662**	0.274	0.207	0.641	0.211	0.194	**0.300**
	Treatment Sequence	0.624	0.599	0.421	0.431	**0.683**	0.470	0.564	0.634	0.604	**0.629**	0.411
Medicine	MMLU-Pro-Health*	**0.784**	0.759	0.681	0.498	0.603	0.549	0.634	0.647	**0.686**	0.660	0.532
	Affinity Extraction	0.068	**0.104**	0.074	0.051	0.041	0.027	0.065	0.087	**0.086**	0.058	0.052
	Drug Chart QA	**0.600**	0.467	0.333	0.400	0.333	0.333	0.333	0.400	**0.400**	**0.400**	0.267
	Tag2Mol	0.127	0.073	0.036	0.000	0.127	0.008	0.123	**0.216**	**0.097**	0.002	0.000
	Markush2Mol	0.662	0.645	**0.675**	0.488	0.664	0.519	0.583	0.671	0.376	**0.475**	0.400
	Mol In Document	**0.840**	0.480	0.580	0.380	0.460	0.460	0.600	0.700	0.480	**0.540**	0.460

Table 8: Performance comparison of LLMs across various scientific domains. **Orange** and **green** indicate the best in closed and open source LLMs, respectively. * indicates 3-shot. The prompts simply require the model to return the final answer without chain-of-thought.

Ranking	MMLU-Pro	SciAssess (L1)	SciAssess (L2)	SciAssess (L3)
o1	-	-	-	-
gpt-4o	1	1	2	3
gpt-4	5	5	4	1
gpt-3.5	-	-	-	-
Moonshot	-	-	-	-
claude3	4	6	6	7
Doubao	-	-	-	-
Gemini	3	4	3	1
Llama3.1	7	3	1	5
Qwen2.5	2	2	5	4
Mixtral	6	7	7	5

Table 9: MMLU-Pro Ranking Comparison (Only overlapping models are shown)

Ranking	SciKnowEval	SciAssess (L1)	SciAssess (L2)	SciAssess (L3)
o1	-	-	-	-
gpt-4o	1	1	1	3
gpt-4	2	3	3	1
gpt-3.5	-	-	-	-
Moonshot	-	-	-	-
claude3	-	-	-	-
Doubao	-	-	-	-
Gemini	3	2	2	2
Llama3.1	-	-	-	-
Qwen2.5	-	-	-	-
Mixtral	-	-	-	-

Table 10: SciKnowEval Ranking Comparison (Only overlapping models are shown)

- **GPT-4** (OpenAI, 2023): OpenAI's GPT-4 excels in text generation and comprehension, augmented with capabilities for image processing, code interpretation, and information retrieval. These features make it adept at handling the complexities of scientific texts, positioning it as a versatile tool for scientific research. The model we use is `gpt-4-turbo`.
- **GPT-3.5**[4]: Preceding GPT-4, GPT-3.5 by OpenAI distinguishes itself with adept language processing skills, enabling effective engagement with complex texts. The model we use is `gpt-3.5-turbo-0125`.
- **Gemini-1.5-Pro** (Google, 2023): Google DeepMind's Gemini model family excels in multimodal comprehension, integrating text, code, image, and audio analysis.
- **Claude 3 Opus**[5]: Claude 3 Opus model excels across major AI benchmarks, demonstrating near-human levels of comprehension and fluency in tasks like analysis, forecasting, and multilingual communication.
- **Moonshot-v1**[6]: Moonshot-v1 is a text generation model proposed by Moonshot AI. We use `moonshot-v1-128k` in this study.
- **Doubao**[7]: Doubao is a set of LLMs developed by ByteDance. The model we use is `Doubao-pro-128k`.

Apart from the closed-source LLMs, we also include some SOTA open-source LLMs:

[4] https://openai.com/blog/chatgpt

[5] https://www.anthropic.com/news/claude-3-family

[6] https://platform.moonshot.cn/docs/intro

[7] https://www.volcengine.com/product/doubao

- **Llama-3.1-70B**[8]: Llama 3-70B is a leading open-source LLMs released by Meta.
- **Mixtral-8x22B** (Jiang et al., 2024): Mixtral-8x22B-Instruct-v0.1 is the latest and largest mixture of experts large language model (LLM) from Mistral AI.
- **Qwen-2.5-72B** (Bai et al., 2023): Qwen2 are series of LLMs developed by Alibaba. The model we test is `Qwen2-72B-Instruct`.

[8] https://ai.meta.com/blog/meta-llama-3/