
Scientists’ First Exam: Probing Cognitive Abilities of MLLM via Perception, Understanding, and Reasoning

PrismaX Team*

Shanghai Artificial Intelligence Laboratory

Dataset: <https://huggingface.co/datasets/PrismaX/SFE>

Website: <https://prismax.opencompass.org.cn/>

Code: <https://github.com/PrismaX-Team/sfe>

Abstract

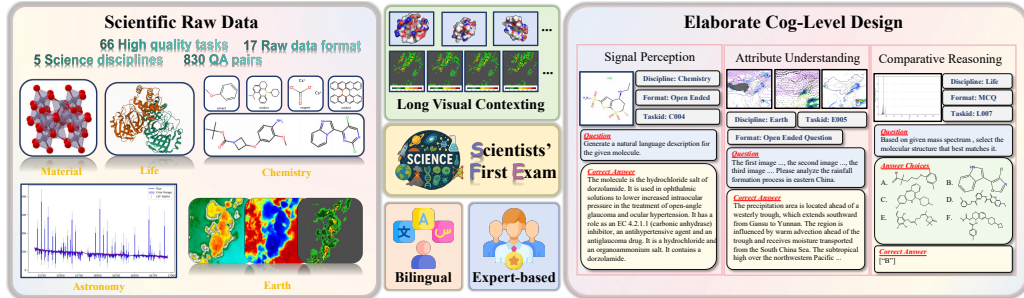
Scientific discoveries increasingly rely on complex multimodal reasoning that integrates information-intensive scientific data and domain-specific expertise. Empowered by expert-level scientific benchmarks, scientific Multimodal Large Language Models (MLLMs) hold the potential to significantly enhance this discovery process in realistic workflows. However, current scientific benchmarks mostly focus on evaluating the knowledge understanding capabilities of MLLMs, leading to an inadequate assessment of their perception and reasoning abilities. To address this gap, we present the Scientists First Exam (SFE) benchmark, designed to evaluate the scientific cognitive capacities of MLLMs through three cognitive levels: *scientific signal perception*, *scientific attribute understanding*, *scientific comparative reasoning*. Specifically, SFE comprises 830 expert-verified VQA pairs across three question types, spanning 66 multimodal tasks across five high-value disciplines. Extensive experiments reveal that current *state-of-the-art* GPT-o3 and InternVL-3 achieve only 34.08% and 26.52% on SFE, highlighting significant room for MLLMs to improve in scientific realms. We hope the insights obtained in SFE will facilitate further developments in AI-enhanced scientific discoveries.

1 Introduction

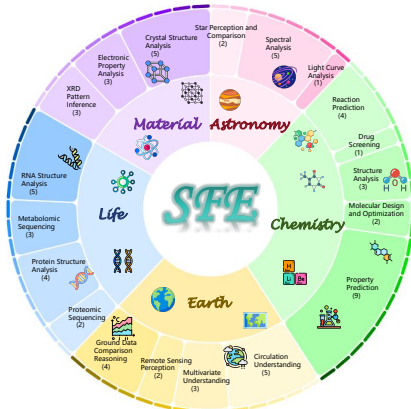
Scientific discoveries rely on investing significant time in analyzing large-scale, complex, and diverse data. Researchers require domain-specific knowledge to interpret scientific data across various modalities, and apply problem-solving skills to address specific scientific challenges [64]. Recent advances in multimodal large language models (MLLMs) have achieved remarkable performance on a wide range of benchmarks, comparable to or even surpassing human-level understanding in both general-level (e.g., MMLU [28], SuperGLUE [67], TriviaQA [31]) and graduate-level (e.g., GPQA [60], HumanEval [12], GSM8K [15]) knowledge domains. As MLLMs continue to progress from general-purpose understanding to domain-specific knowledge, scientific discovery has emerged as a critical frontier for evaluating and extending their abilities [5, 10, 11, 35, 42, 44, 84].

The process of scientific discovery often involves specialized scientific analysis of data modalities (e.g., molecular structures, spectra, protein sequences) from various scientific fields. SuperGPQA [17] extends conventional domains by incorporating long-tail disciplines, ensuring accessibility to real-world professional expertise. CURIE [16] establishes a ten-task benchmark for evaluating scientific reasoning in long-context scenarios. HLE [59] is introduced to evaluate model capabilities

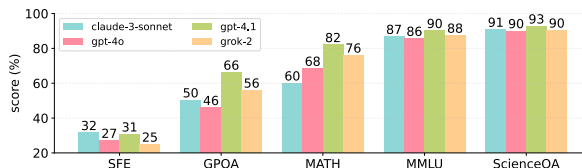
*Please refer to Appendix A for all team members



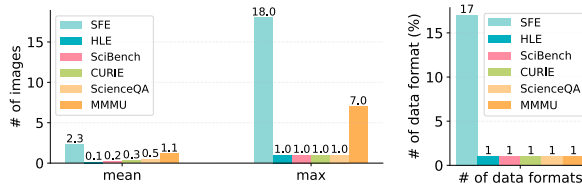
(a) Examples from our proposed SFE. SFE is designed to comprehensively evaluate the scientific capabilities of MLLMs in depth and breadth.



(b) The structure of SFE includes 5 disciplines, 19 scientific directions, and 66 tasks.



(c) Performance of mainstream MLLM on various benchmarks.



(d) Number of images per QA on various benchmarks. (e) Number of released data formats.

Figure 1: Overview of the Scientists' First Exam (SFE) benchmark.

through challenging and expert-authored questions. However, despite the growing interest in the scientific domain, most existing scientific benchmarks extract tasks from secondary sources such as academic materials [16, 27, 78] and textbooks [70, 78]. As a result, they inadequately probe the cognitive abilities (e.g., perception, understanding, and reasoning) required for analyzing scientific data encountered in real-world research. Moreover, these benchmarks tend to focus exclusively on a single ability to interpret domain knowledge from the data, while neglecting the full spectrum from perception to reasoning. This capability gap exposes a fundamental challenge: *How to granularly measure MLLMs' scientific cognitive capabilities across multiple disciplines for scientific discovery?*

To bridge this gap, we introduce the Scientists' First Exam (SFE) benchmark, designed to comprehensively evaluate the scientific cognitive capabilities of MLLMs through three cognitive levels (cog-levels): *Scientific Signal Perception (L1)* characterizes the capacity to discern critical components within visualizations of scientific raw data; *Scientific Attribute Understanding (L2)* demonstrates the ability to interpret domain-expert knowledge; *Scientific Comparative Reasoning (L3)* manifests the ability to derive phenomenological insights through structured comparison of multiple scientific visual sources. SFE encompasses 66 expert-curated, high-value multimodal tasks across five disciplines: Astronomy, Chemistry, Earth, Life, and Materials Sciences (Fig. 1b). Each task is constructed from native scientific raw data formats (Fig. 1a) and formulated as visual question answering (VQA) pairs (Fig. 1e), designed to probe specific levels of scientific cognition. All tasks are bilingual (English & Chinese) to support broad accessibility. These tasks are designed not only to require a deep understanding of domain-specific knowledge and data analysis skills but also to significantly enhance research efficiency and facilitate advancements that benefit society.

We benchmark 16 *state-of-the-art* open and closed weight MLLMs using SFE, as illustrated in Fig. 1c. As observed, while these MLLMs perform well on benchmarks such as MMLU [78] and ScienceQA [48], they all exhibit suboptimal results on the SFE benchmark. This indicates that SFE serves as a challenging frontier for scientific MLLM development. Our contributions:

Table 1: Comparison of large language model (LLM) benchmarks related to science. Astro, Chem, Phy, Bio, CS, QC, Geo, and Mat are abbreviations for Astronomy, Chemistry, Physics, Biology, Computer Science, Quantum Computing, Geospatial Analysis, and Material, respectively. AD, Bus., Sci., Med., HSS, and TE refer to Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering, respectively. The Question Types include MCQ (Multiple-Choice Questions), EM (Exact Match), and OQ (Open Questions). EN/ZH denotes the language of each benchmark (English/Chinese).

Benchmark	Discipline	Multi-Modal	Question Type	Raw Data	Task #Count	Question Source	Language
MMMU[78]	AD, Bus., Sci. Med., HSS, TE	✓	MCQ OQ	.png	✗	Textbooks e-Resources	EN
SCIENCEQA[48]	Natural, social and language science	✓	MCQ	.png	✗	Curricula	EN
SCIBENCH[70]	Math, Chem Physics	✓	OQ	.png	✗	Textbooks	EN
CMMU[27]	7 School subjects	✓	OQ	.png	✗	Exams	ZH
ChemBench[26]	Chem	✗	EM	-	8	Public datasets	EN
SuperGPQA[17]	13 Disciplines	✗	MCQ	-	✗	Experts	EN
SciEval[63]	Chem, Phy, Bio	✗	MCQ, EM judgment	-	✗	Knowledge base	EN
HLE[59]	Math, Phy, Bio HSS, CS, Chem, TE	✓	MCQ, EM	.png	✗	Experts	EN
CURIE[16]	Mat, Phy, QC, Geo Bio, Proteins	✓	OQ	.png	10	Experts	EN
SFE	Astro, Chem, Earth Life, Mat	✓	MCQ EM, OQ	.png, .mgfsto, .txt'	66	Experts	EN, ZH

1. We propose the first benchmark to categorize scientific tasks by cognitive capacity, introducing a three-level taxonomy: Scientific Signal Perception (**L1**), Scientific Attribute Understanding (**L2**), and Scientific Comparative Reasoning (**L3**). This formulation enables fine-grained evaluation of how MLLMs engage with different layers of scientific research.
2. We release the bilingual SFE benchmark, encompassing 66 expert-curated multimodal tasks across five scientific disciplines and covering three question types. All tasks are constructed from native scientific data formats and aligned with three cognitive capacity levels.
3. We comprehensively evaluate 16 state-of-the-art MLLMs, revealing that GPT-o3 achieves the best overall performance, and newer model versions show clear improvements in L3 tasks.

2 Related Works

Science LLMs / MLLMs. Recent advancements in domain-specific large language models (LLMs) have significantly impacted various scientific fields. In the biomedical domain, LLMs have been employed for tasks such as clinical documentation, information retrieval, and hypothesis generation [35, 36, 49–51, 56]. In chemistry, recent studies have focused on tasks such as drug property prediction, molecular discovery, chemical reaction extraction, and protein structure understanding [11, 39, 40, 43–45, 47, 52, 81, 87]. Notably, researchers have developed LLMs trained on scientific corpora to support molecular and protein discovery [5, 14, 20, 54, 66, 71]. For instance, iupacGPT [14] uses IUPAC nomenclature to effectively capture relationships among atoms and chemical groups. Similarly, Progen [54] and ProtGPT2 [20] are trained for protein sequence generation and understanding. Geosciences have benefited from LLMs through their applications in ocean science, extreme weather, and remote sensing [8, 38, 41, 53, 73, 77, 83, 84]. Models like EarthGPT [83] and CLLMate [38] further integrate multimodal knowledge to support scientific question answering.

Science Benchmarks. With the development of large language models (LLMs), recent efforts in LLM benchmarking have increasingly focused on evaluating scientific reasoning capabilities across diverse domains and modalities. Early benchmarks [46] such as ScienceQA [48] and CMMU [27], primarily focus on science level below high school. Several college-level benchmarks [17, 70, 72,

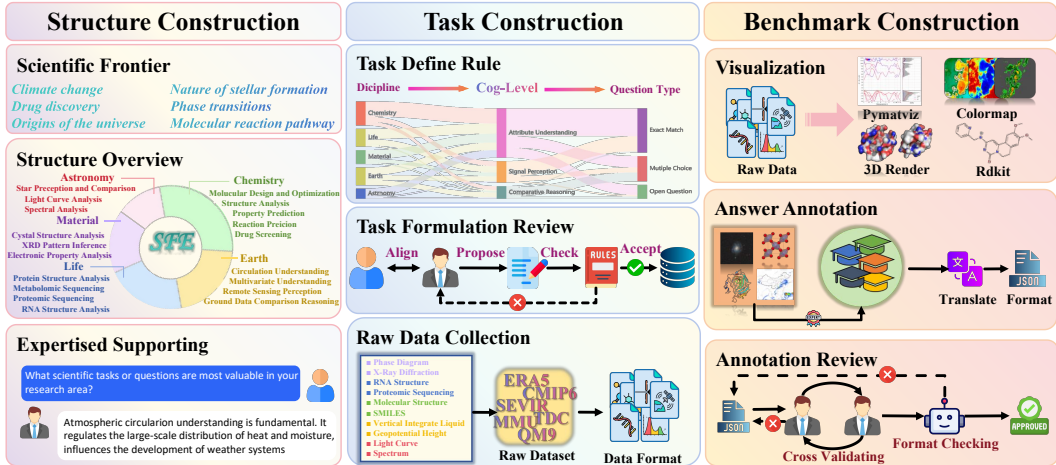


Figure 2: **Data collection framework of SFE.** First, we define 18 scientific directions based on the scientific frontier and domain experts. Building upon this structure, we invite experts to propose tasks and contribute raw data based on three cognitive levels. Finally, we employ visualization techniques and further engage experts to annotate the resulting benchmark.

[78, 79] have emerged to support deeper scientific understanding. As state-of-the-art multimodal LLMs rapidly improve, many of them now achieve strong performance on these benchmarks. This has motivated the creation of more expert-level datasets that feature high-difficulty questions such as HLE [59] and CURIE [16]. However, although some benchmarks offer broad domain coverage, they often lack clearly defined tasks, making it difficult to assess specific model limitations. To address this, works like ChemLLMBench [26] and CURIE [16] have introduced smaller-scale but task-oriented benchmarks that enable more targeted evaluation of scientific capabilities. Nevertheless, nearly all of these benchmarks remain monolingual, limiting their applicability for comprehensively evaluating LLMs in global deployment contexts.

3 SFE Dataset and Tasks

The SFE benchmark consists of 830 multimodal VQA pairs spanning 66 real-world scientific tasks across five disciplines: Astronomy, Chemistry, Earth, Life, and Materials Science (Fig. 1b). Each task is constructed from native disciplinary data formats such as molecular structures, spectra, and radar charts, and is carefully annotated by domain experts. On average, each question contains 2.3 scientific images (ranging from 1 to 18), and the dataset supports bilingual prompts and answers in both English and Chinese. In total, SFE spans 17 distinct scientific data formats and is motivated by scientific frontiers such as drug discovery, celestial body radial velocity estimation and peptide sequence inference, etc. Detailed dataset statistics are presented in Table 2, and a complete task list is provided in Appendix Table 6.

3.1 Data Collection

The construction of SFE is the result of extensive collaboration with domain experts across multiple scientific disciplines and research directions. As shown in Fig. 2, the overall data collection pipeline comprises three key stages: (1) **Structure Construction**, where we collaborate with experts to identify high-value challenges and define scientific directions; (2) **Task Construction**, where the scientific directions are instantiated into concrete tasks with specific question types and cognitive

Table 2: **Statistics of SFE.**

Statistic	Number
# of VQA	830
# of Scientific Tasks	66
# of scientific signal perception VQA	202
# of scientific attribute understanding VQA	503
# of scientific comparative reasoning VQA	125
# of MCQ	284
# of Exact Match	420
# of Open Question	126
Average Question Tokens Length	88 (en) / 86 (zh)
Average Answer Tokens Length	100 (en) / 106 (zh)
Average # of Images	2.3 (1-18)
Released # Scientific Data Format	17

levels through expert-driven design and review; (3) **Benchmark Construction**, where scientific raw data are rendered, visualized, and used to construct expert-authored VQA with high-quality samples.

Structure Construction. The SFE is initiated by identifying foundational scientific directions across five core disciplines: Astronomy, Chemistry, Earth, Life, and Materials Science. These fields are selected for their broad scientific relevance and the need for specialized knowledge. Based on the experts’ consultations, we define a structured set of scientific directions, such as circulation understanding in Earth and reaction prediction in chemistry. These directions serve as the foundation for downstream task formulation and reflect current research frontiers.

Task Construction. Building upon the established scientific directions, domain experts collaboratively define concrete tasks. We first formalize the task definition rules by mapping each scientific direction to an appropriate cognitive level and question type as shown in Fig. 2. This taxonomy ensures that each task is aligned with a distinct level of cognitive capacity and supports evaluation diversity. For each discipline, we collaborated with 2-5 experts to define benchmark tasks that align with high-priority issues in real-world scientific research. Each expert-designed task is expected to satisfy the following core criteria: (1) It reflects a meaningful problem that requires substantial domain knowledge and is commonly encountered in real-world research. (2) It must be solvable through expert-level reasoning that necessarily integrates multimodal scientific inputs such as structured visualizations and symbolic representations. (3) It aligns with one of the proposed cognitive levels (L1-L3). To support each task, experts also provide instructions for collecting raw data that would serve as visual inputs for the VQA pairs. These data sources span a diverse range of formats, including public datasets (e.g., ERA5 [29]), existing benchmarks (e.g., MoleculeNet [75]), and domain-specific databases (e.g., RCSB PDB [7], PDBbind [69], and PubChem [33]). For each task, domain experts identify suitable data entries that are typically used to analyze the corresponding scientific problems, such as RNA sequence files (.stockholm) and protein structure files (.pdb).

Benchmark Construction. In the final stage, each task is instantiated into a set of VQA pairs. To ensure accessibility and visual coherence, all data formats are rendered into image form. During the answer annotation phase, experts construct each VQA pair by composing: (1) rendered multimodal input, (2) a scientifically meaningful question, and (3) an answer grounded in expert reasoning. Then, all VQA pairs are translated into both Chinese and English to provide multilingual support and formatted in JSON for standardization. Finally, a two-stage validation process is applied for quality control. Each VQA pair is first cross-reviewed by domain experts for scientific correctness, clarity, and alignment with the intended reasoning level. This is followed by rule-based validation for format checking. Only VQA pairs that pass both stages are included in the final benchmark.

3.2 Tasks

Astronomy. In Astronomy, analyzing diverse modalities such as spectra is essential for a wide range of scientific tasks, including property estimation and event detection. Therefore, we design 8 tasks to assess the cognitive abilities of MLLMs from 3 cognitive levels. L1 tasks include *galaxy morphology classification*, where MLLMs are required to differentiate perceptual features of galaxies based on the provided image. Second, L2 tasks include *surface temperature estimation*, *gravitational constant estimation*, *light curve classification*, *metallicity estimation*, etc. For example, in the *metallicity estimation* task, MLLMs need to infer the total metal abundance of a target celestial object based on its observed spectrum. Third, L3 tasks target *transient detection*. Given the pre-transient and post-transient images, along with difference images, MLLMs are expected to determine whether a transient has occurred during the process. Refer to the Appendix for more details.

Chemistry. We formulate 19 tasks in the Chemistry domain, spanning three cognitive levels to evaluate MLLMs’ understanding of molecular structures, properties, and interactions. L1 tasks include *elemental composition recognition*, *molecular description generation*, etc. In the *molecular description generation* task, MLLMs are required to produce descriptions of specific molecules, highlighting key attributes such as types of chemical bonds and the number of carbon atoms. L2 tasks involve *Lipinski drug-likeness estimation*, *absorption property prediction*, *distribution property prediction*, etc. For example, in the *absorption property prediction* task, MLLMs need to estimate properties of molecules, such as the plasma protein binding rate. L3 tasks include *virtual screening*, *protein-ligand binding affinity prediction*, etc. In the *virtual screening* task, MLLMs are

expected to identify all molecules capable of binding to a given structure. Refer to the Appendix for more details.

Earth. To systematically evaluate the performance of MLLMs in the Earth science domain, we construct 14 tasks across three cognitive levels based on diverse weather variables and data sources. First, L1 tasks include *thermocline depth recognition*, *perception of extreme precipitation distribution*, *SAR image grounding*, etc. For example, in the *perception of extreme precipitation distribution* task, MLLMs are required to identify multiple locations of extreme precipitation. Second, L2 tasks include *moisture source understanding*, *precipitation event analysis*, *convective weather types identification*, etc. For example, the *precipitation event analysis* requires models to analyze the formation process of precipitation by using information from geopotential height, moisture flux, vertical velocity, etc. Third, L3 tasks consist of *differential prediction comparison*, *temperature sequence comparison*, etc. In the *temperature sequence comparison* task, MLLMs need to compare annual temperature series from two different time periods and describe differences in statistical characteristics. Refer to the Appendix for more details.

Life. In the life science domain, we construct 14 tasks across three cognitive levels to comprehensively evaluate MLLMs, focusing on modalities such as RNA structures and protein sequences. L1 tasks include *fragment ion peaks count*, *protein chain count*, *small molecule count*, etc. For example, in the *fragment ion peaks count* task, MLLMs are asked to determine the number of specific ions present in a given MS/MS spectrum. L2 tasks consist of *molecular composition inference*, *specified protein detection*, etc. For instance, in the *molecular composition inference* task, MLLMs need to infer the elemental composition of the compound based on the provided spectrum. L3 tasks include *spectrum matching*, where MLLMs are required to identify the molecular structure that best corresponds to a given mass spectrum. Refer to the Appendix for more details.

Materials. To evaluate MLLMs’ performance in the realm of materials, we formulate 11 tasks spanning three cognitive levels. L1 tasks include *atomic composition description*, *crystal group identification*, *crystal formula determination*, etc. For example, in the *atomic composition description* task, models are required to select relevant properties of a target lattice from multiple descriptions. Second, L2 tasks cover *band gap classification*, *stability estimation*, *energy band and DOS interpretation*. For example, in the *energy band and DOS interpretation* task, MLLMs are required to infer whether a given material belongs to a metal or a semiconductor and estimate its band gap. Third, L3 tasks involve complicated comparative reasoning, such as *phase identification*. Specifically, given the XRD pattern of a composite material and candidate substances, MLLMs are required to identify three materials that form the composite material. Refer to the Appendix for more details.

4 Experiments and Evaluations

General Settings. We conduct a comprehensive evaluation of the *state-of-the-art* MLLMs on SFE. For models with open weights, we assess InternVL2.5-78B [13], InternVL3-78B [86], Qwen2.5-VL-72B [6], LLaMaVision-90B [22], and LLaVa-Onevision-72B [37]. For models with closed weights, we evaluate GPT-4o-2024-11-20 [57], GPT-4.1-2025-04-14 [57], GPT-o1-2024-12-17 [57], GPT-o3 [57], Claude 3 Opus [4], Claude 3.7 Sonnet [4], Gemini-2.0-Flash [65], Gemini-2.5-Flash [65], Gemini-2.5-Pro [65], Grok-2-Vision-12-12 [76] and Doubao-1.5-Vision-Pro [24]. When benchmarking, we configure all MLLMs’ temperatures to 0 for reduced randomness and employ a standard zero-shot prompt template across all tasks. Specifically, the template begins with a description of the task assigned to the model, followed by the inclusion of question texts with interleaved images. Additionally, we fix the maximal number of generated tokens to 1024, ensuring fairness and cost-effectiveness in our evaluations.

Metrics. We present the BERTScore [82] and the LLM-as-a-Judge score [23] for all tasks, except for the remote sensing perception task in Earth science, where we report the execution success rate and the Intersection over Union (IoU). For BERTScore, we use the F1 score. For LLM-as-a-Judge score, we employ GPT-4o-2024-11-20 as the judge, allowing us to semantically verify the correctness of answers against model predictions. Conversely, the execution success rate is utilized to assess whether MLLMs can accurately follow prompts to produce bounding boxes in the desired format. Finally, the IoU metric evaluates the precision of these bounding boxes compared to ground truths. Without further clarification, each experiment is conducted once to obtain the final results.

Table 3: Experimental results of all models on different disciplines using different languages. The LLM-as-a-Judge score is used as the evaluation metric. ‘Average’ represents the mean score.

Model	Astronomy		Chemistry		Earth		Life		Material		Average	
	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
<i>Closed Weight MLLMs</i>												
Grok-2-Vision-12-12	19.37	18.23	21.29	20.20	33.58	36.23	25.10	25.95	35.57	36.89	24.97	25.10
GPT-4o-2024-11-20	20.38	18.04	21.60	19.25	32.65	29.93	30.39	27.58	49.67	48.36	27.15	24.72
GPT-4.1-2025-04-14	24.18	25.50	24.01	22.11	40.40	44.43	34.90	33.66	47.70	48.85	30.88	31.05
GPT-o1-2024-12-17	22.03	21.84	27.41	27.21	38.61	36.56	33.99	31.31	61.15	61.64	32.19	31.24
GPT-o3	24.24	23.80	28.91	27.89	43.05	36.29	33.59	31.57	63.44	58.20	34.08	31.60
Gemini-2.0-Flash	16.14	12.78	<u>27.82</u>	24.69	34.24	32.91	32.48	27.32	52.79	50.49	29.49	26.33
Gemini-2.5-Flash	24.30	24.11	23.67	23.47	31.99	30.53	25.03	25.10	56.39	55.90	28.03	27.63
Gemini-2.5-Pro	5.13	6.08	2.07	2.28	2.52	3.84	19.73	22.35	28.69	27.70	8.04	8.96
Claude-3-Opus	14.68	16.08	19.63	17.45	36.62	32.12	22.55	23.66	36.72	32.13	23.64	22.15
Claude-3.7-Sonnet	25.89	22.34	27.79	25.14	38.21	37.09	31.24	29.48	49.51	46.72	31.62	29.23
Doubao-1.5-vision-pro	28.35	23.99	25.00	24.46	26.16	25.83	30.07	29.35	51.48	46.39	28.79	27.17
<i>Open Weight MLLMs</i>												
Qwen2.5-VL-72b	26.46	20.89	18.33	15.27	25.03	24.57	25.29	22.75	41.47	42.46	24.17	21.51
InternVL-2.5-78B	15.76	17.09	18.54	15.58	34.77	37.22	27.06	25.49	43.11	39.84	24.43	23.54
InternVL-3-78B	<u>27.09</u>	<u>24.62</u>	19.80	16.29	28.81	28.94	30.85	27.58	40.98	42.30	26.52	24.30
Llama-3.2-Vision-90B	20.63	18.04	16.94	15.14	27.22	30.53	25.29	21.37	32.30	31.80	22.26	20.95
Llava-OneVision-72B	25.19	23.23	15.34	15.17	23.31	27.81	24.58	24.58	37.54	42.62	22.10	23.39
Average	21.23	19.79	21.14	19.51	31.07	30.93	28.27	26.82	45.54	44.52	26.15	24.93

Table 4: Experimental results of all models on different cognitive levels and different question types in both Chinese and English.

Model	L1		L2		L3		Exact Match		Open Question		MCQ	
							IoU		LLM score		LLM score	
	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
<i>Closed Weight MLLMs</i>												
Grok-2-Vision-12-12	28.31	25.66	23.10	25.81	27.44	21.44	6.11	8.30	17.32	17.39	0.707	0.718
GPT-4o-2024-11-20	35.50	32.70	26.46	23.32	17.28	18.32	2.66	2.13	20.42	17.58	0.698	0.710
GPT-4.1-2025-04-14	35.98	36.61	<u>30.02</u>	29.64	26.64	28.32	5.19	3.47	22.10	20.14	0.714	0.723
GPT-o1-2024-12-17	43.65	<u>42.80</u>	29.09	26.60	27.36	32.40	0.77	1.83	22.61	21.82	0.70	0.71
GPT-o3	<u>42.54</u>	<u>41.27</u>	30.30	<u>28.27</u>	36.48	<u>31.60</u>	5.17	5.94	<u>23.75</u>	<u>22.12</u>	0.694	0.693
Gemini-2.0-Flash	41.43	36.46	26.86	23.70	22.00	21.60	0.91	0.44	21.38	19.28	0.70	0.65
Gemini-2.5-Flash	37.04	37.57	27.89	26.98	14.96	15.20	0.99	0.96	24.45	24.45	0.688	0.688
Gemini-2.5-Pro	17.46	18.78	6.16	6.94	1.36	2.24	-	-	8.11	8.60	0.548	0.558
Claude-3-Opus	25.71	23.54	23.80	22.09	19.84	20.32	0.87	0.08	15.99	14.71	0.700	0.708
Claude-3.7-Sonnet	37.46	33.39	29.80	28.15	30.08	27.28	2.27	1.38	21.40	19.09	<u>0.710</u>	0.671
Doubao-1.5-vision-pro	32.70	32.43	26.48	24.31	<u>32.16</u>	30.72	2.19	<u>11.84</u>	23.50	21.00	0.677	0.691
<i>Open Weight MLLMs</i>												
Qwen2.5-VL-72b	31.11	26.08	23.36	21.91	16.96	12.96	24.49	15.37	20.49	17.55	0.645	0.667
InternVL-2.5-78B	25.08	22.91	25.33	24.35	19.84	21.20	3.79	3.65	16.53	14.92	0.690	0.692
InternVL-3-78B	26.67	27.30	26.10	22.82	28.00	25.68	5.25	7.16	19.39	15.99	0.69	0.70
Llama-3.2-Vision-90B	26.14	24.23	22.33	20.34	16.16	18.48	0.27	0.25	15.90	14.17	0.70	0.68
Llava-OneVision-72B	25.13	26.40	19.14	20.68	29.52	29.76	2.84	1.42	14.99	15.87	0.700	0.665
Average	32.00	30.49	24.76	23.49	22.89	22.34	4.26	4.25	19.27	17.79	0.689	0.683

4.1 Main Results

We evaluate the performance of 16 MLLMs across five scientific disciplines and two languages (English and Chinese) on SFE using the LLM-as-a-Judge score for evaluation metric. Our results in Table 3 demonstrate that SFE is capable of revealing fine-grained differences in model capabilities, visual grounding quality, and multilingual reasoning robustness. Below, we detail key observations.

Observation 1. SFE is capable of distinguishing model capability across both proprietary and open-source MLLMs. Among all evaluated models, GPT-o3 achieves the best overall performance, with an average score of 34.08% in English and 31.60% in Chinese. It is consistently strong across all disciplines, particularly in Earth and Materials, where scientific visual interpretation and structured multimodal reasoning are often required. Notably, Gemini-2.5-Pro registers the lowest average score among all evaluated models, with only 8.04% in English and 8.96% in Chinese

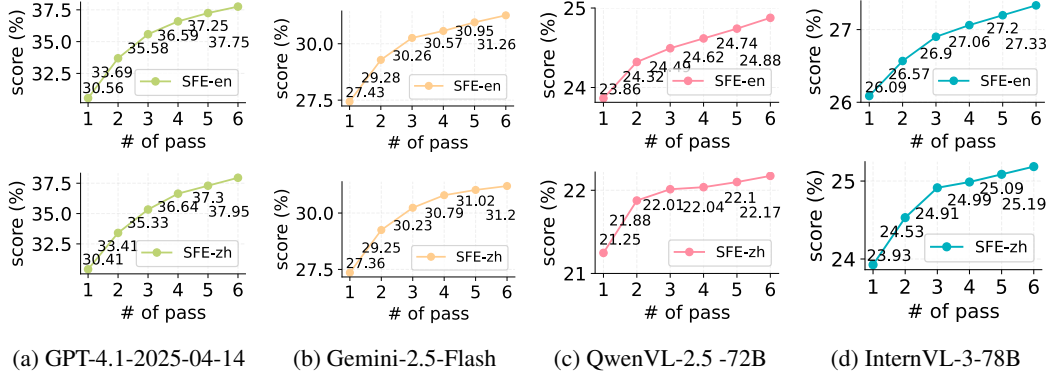


Figure 3: Pass@k scores of four *state-of-the-art* MLLMs on SFE. Closed-weight MLLMs demonstrate superior initial performance and greater scalability than open weight MLLMs.

due to excessive thinking, causing exhausted token budgets and clipped answers². On the other hand, while GPT-o3 reasons as well, it balances reasoning and token usage effectively. The performance gap between GPT-o3 and Gemini-2.5-Pro exceeds 26% on average, showing SFE is comprehensive enough to differentiate the ability across the full range of models. The results also present a systematic divide between closed-weight proprietary models and open-weight models. On average, the best proprietary models (GPT-o3, GPT-o1-2024-12-17, Claude-3.7-Sonnet) outperform the strongest open alternatives (InternVL-3-78B) by 6-8%. Additionally, within SFE, the model series shows clear internal progress. For example, Claude-3.7-Sonnet outperforms Claude-3-Opus by over 7% in both English and Chinese, reflecting measurable architectural or training gains. A similar pattern is also observed within the InternVL series.

Observation 2. SFE exhibits a clear performance gap of MLLMs between disciplines. Our findings reveal that across nearly all models, Material Science emerges as the most tractable domain. The top model, GPT-o3, reaches 63.44% in English and 58.20% in Chinese on it. And even open models like Qwen2.5-VL-72b and InternVL-3-78B achieve over 40% in this domain. This trend reflects the relatively structured visual inputs (*e.g.*, phase diagram, X-Ray diffraction) which require the model to generate structured scientific outputs based on symbolic visual images that are aligned with current models’ comparatively strong skills. In contrast, Astronomy tasks present more substantial challenges. This domain involves spectral analysis tasks, where models estimate numerical astrophysical parameters (*e.g.*, temperature and velocity) from raw or noisy spectral visualizations, which current MLLMs find challenging. This highlights SFE’s role in diagnosing which types of scientific reasoning MLLMs can currently handle and which are still out of reach.

Observation 3. SFE reveals a potential shift in MLLM capabilities from knowledge understanding to high-order reasoning. Through SFE’s three-level cognitive framework, its results provide an analysis of how recent models’ cognitive improvements are distributed. Our findings (in Table 4) show that newer MLLMs exhibit notably higher performance on L3 tasks compared to earlier models, while their L2 performance remains largely similar. This trend aligns with the adoption of advanced reasoning techniques in state-of-the-art models. For example, GPT-o3 improves L3 performance from 26.64% (GPT-4.1-2025-04-14) to 36.48% without an obvious increase in L2 scores (from 30.30% to 30.02%). This pattern reflects OpenAI’s reports that emphasize scaling reinforcement learning for reasoning and learning tool-use strategies rather than knowledge expansion. Similarly, InternVL-3 outperforms InternVL-2.5 by 8% in L3 performance in English, despite only marginal gains in L2. This improvement can be attributed to InternVL-3’s architectural and training advances, particularly its native multimodal pretraining and Mixed Preference Optimization (MPO) to support Chain-of-Thought reasoning, which is a critical skill for success in L3 tasks such as comparative diagram analysis. The negligible gains in L2 tasks suggest that these improvements arise not from broader knowledge acquisition but from enhanced training innovations.

²Experiments with more generated tokens are conducted in Table 17 and Table 18 in the appendix.

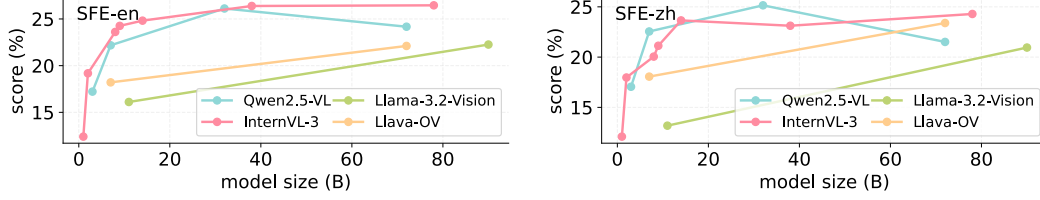


Figure 4: The scaling law of model size on SFE. Clearly, the amount of scientific data has not been scaled proportionally with the increase in model size, resulting in minor overfit for QwenVL-2.5-72B.

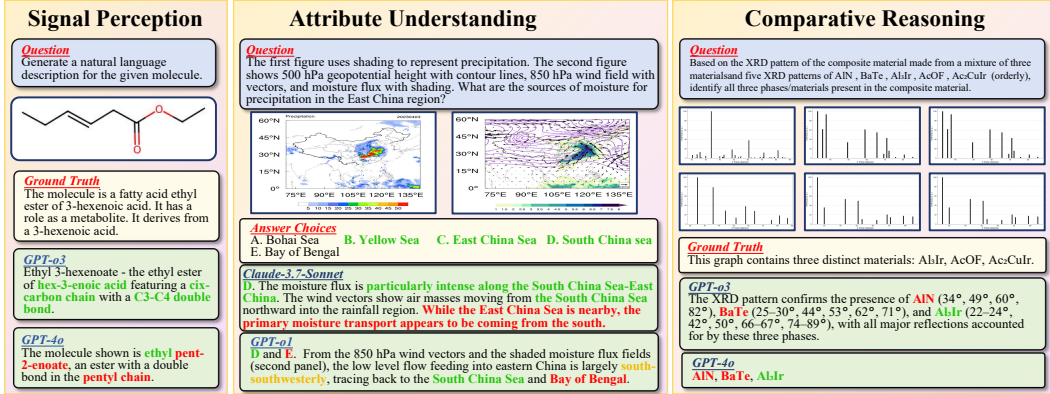


Figure 5: Case studies on SFE across different cognitive levels.

4.2 Analysis

Pass@k Analysis. The Pass@k metric [34] selects the highest quality answers from an MLLM as its final response to a question, indicating the models potential for improvement through post-training (e.g., RLHF [58], GRPO [25], etc.). In Fig. 3, we evaluate pass@k scores of GPT-4.1-2025-04-14, Gemini-2.5-Flash, Qwen2.5-VL-72b and InternVL-3-78B on SFE, with k ranging from 1 to 6. As shown, GPT-4.1-2025-04-14 and Gemini-2.5-Flash outperform *state-of-the-art* open-weight MLLMs by not only exhibiting superior initial performance (30.56% v.s. 26.09%) but also demonstrating strong scalability (30.56% \rightarrow 37.75% v.s. 26.09 \rightarrow 27.33%). This suggests that closed-weight MLLMs may leverage more diverse and expansive raw datasets during pre-training than open-weight MLLMs. Furthermore, their post-training phase may prioritize a balanced approach, emphasizing exploration alongside exploitation, rather than focusing exclusively on exploitation [80].

Scaling Law of Model Size. We benchmarked MLLMs of varying sizes to evaluate their impact on SFE, as depicted in Fig. 4. Although Llama-3.2-Vision and Llava-Onevision series improve with size, they underperform compared to the Qwen2.5-VL and InternVL series. The larger models, Qwen2.5-VL-72B and InternVL-3-78B, do not significantly surpass their smaller counterparts, indicating a lack of proportional scientific data scaling during pre-training. Additionally, Qwen2.5-VL-72Bs performance is lower than Qwen2.5-VL-7B, suggesting potential overfitting. This emphasizes the need for balanced data scaling relative to model size in the scientific domain.

Impact on Temperatures. We analyze the effect of temperature settings on MLLMs’ performance, as shown in Table 5. The results indicate that both excessively high and low temperature values can lead to performance degradation for scientific discoveries. Empirically, maintaining the temperature within the range of 0.4 to 0.6 optimally balances the trade-off between exploration and exploitation, thereby enhancing MLLMs’ efficacy.

Table 5: Scores on different temperatures.

Model	0.0	0.2	0.4	0.6	0.8	1.0
GPT-4.1-2025-04-14	30.88	31.31	31.59	32.09	31.97	31.63
InternVL-3-78B	24.43	26.98	26.93	26.55	25.86	23.12

Case Study. Fig. 5 presents case studies evaluating the performance of MLLMs on SFE across different cognitive levels. In the signal perception task, both GPT-4o-2024-11-20 and GPT-o3

effectively recognize functional groups within molecular images. However, GPT-o3, equipped with reasoning capabilities, offers more precise results for fine-grained tasks such as counting carbon atoms. For attribute understanding, Claude-3.7-Sonnet and GPT-o1-2024-12-17 accurately identify colormaps and wind vectors in geographical images. Despite achieving approximately one-third accuracy in their responses, GPT-o1-2024-12-17 falsely identifies the Bay of Bengal with its wind strength misinterpreted, suggesting an incomplete geographical understanding. In comparative reasoning, GPT-o3 integrates information from multiple images but struggles with comparative selection, revealing an incomplete reasoning chain. Conversely, across similar tasks, MLLMs like GPT-4o-2024-11-20, Grok-2-Vision-12-12, and others primarily focus on the initial four images, indicating limitations in processing long visual contexts.

5 Conclusion

In this paper, we propose the Scientists First Example (SFE) benchmark, aiming to provide a granular assessment of MLLMs’ scientific cognitive capabilities from perception to reasoning. SFE includes 830 expert-verified VQA pairs across 66 tasks in five high-value disciplines, addressing critical needs for more rigorous and diverse evaluation tools for scientific MLLMs. Extensive experiments reveal insights that could contribute to advancements in AI-driven scientific discoveries.

However, the benchmarks scope and depth could be further enhanced, which we intend to improve in future work. Furthermore, while SFE has the potential to significantly advance these discoveries by offering a robust evaluation framework, it also raises concerns about increasing reliance on AI in scientific research. This might inadvertently undermine the value of human intuition and creativity.

Acknowledgments and Disclosure of Funding

This work is supported by Intern Discovery.

References

- [1] A. Aghamoussa, J. Aguilar, S. Ahlen, S. Alam, L. E. Allen, C. A. Prieto, J. Annis, S. Bailey, C. Balland, O. Ballester, et al. The desi experiment part i: science, targeting, and survey design. *arXiv preprint arXiv:1611.00036*, 2016.
- [2] H. Aihara, N. Arimoto, R. Armstrong, S. Arnouts, N. A. Bahcall, S. Bickerton, J. Bosch, K. Bundy, P. L. Capak, J. H. Chan, et al. The hyper supprime-cam ssp survey: overview and survey design. *Publications of the Astronomical Society of Japan*, 2018.
- [3] E. Angeloudi, J. Audenaert, M. Bowles, B. M. Boyd, D. Chemaly, B. Cherinka, I. Ciucă, M. Cranmer, A. Do, M. Grayling, et al. The multimodal universe: enabling large-scale machine learning with 100 tb of astronomical scientific data. In *NeurIPS*, 2024.
- [4] Anthropic. Anthropic ai, 2023. URL <https://www.anthropic.com>.
- [5] V. Bagal, R. Aggarwal, P. Vinod, and U. D. Priyakumar. Molgpt: molecular generation using a transformer-decoder model, 2021.
- [6] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [7] H. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide protein data bank. *Nature structural & molecular biology*, 10(12):980–980, 2003.
- [8] Z. Bi, N. Zhang, Y. Xue, Y. Ou, D. Ji, G. Zheng, and H. Chen. Oceangpt: A large language model for ocean science tasks. *arXiv preprint arXiv:2310.02031*, 2023.
- [9] M. S. Binetti, C. Campanale, C. Massarelli, and V. F. Uricchio. The use of weather radar data: Possibilities, challenges and advanced applications. *Earth*, 2022.
- [10] H. Cao, Z. Liu, X. Lu, Y. Yao, and Y. Li. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*, 2023.

- [11] L. Chen, W. Wang, Z. Bai, P. Xu, Y. Fang, J. Fang, W. Wu, L. Zhou, R. Zhang, Y. Xia, et al. PharmaGPT: Domain-specific large language models for bio-pharmaceutical and chemistry. *arXiv preprint arXiv:2406.18045*, 2024.
- [12] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- [13] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [14] K.-H. Cho, K. T. No, et al. iupacgpt: Iupac-based large-scale molecular pre-trained model for property prediction and molecule generation. 2023.
- [15] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [16] H. Cui, Z. Shamsi, G. Cheon, X. Ma, S. Li, M. Tikhonovskaya, P. C. Norgaard, N. Mudur, M. B. Plomecka, P. Raccuglia, et al. CURIE: Evaluating llms on multitask scientific long-context understanding and reasoning. In *ICLR*, 2025.
- [17] X. Du, Y. Yao, K. Ma, B. Wang, T. Zheng, K. Zhu, M. Liu, Y. Liang, X. Jin, Z. Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.
- [18] Y. Du, A. R. Jamasb, J. Guo, T. Fu, C. Harris, Y. Wang, C. Duan, P. Liò, P. Schwaller, and T. L. Blundell. Machine learning-aided generative molecular design. *Nature Machine Intelligence*, 6(6):589–604, 2024.
- [19] L. L. Ferreira and A. D. Andricopulo. Admet modeling approaches in drug discovery. *Drug discovery today*, 24(5):1157–1165, 2019.
- [20] N. Ferruz, S. Schmidt, and B. Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications*, 2022.
- [21] J. Gong, L. Bai, P. Ye, W. Xu, N. Liu, J. Dai, X. Yang, and W. Ouyang. Cascast: Skillful high-resolution precipitation nowcasting via cascaded modelling. *arXiv preprint arXiv:2402.04290*, 2024.
- [22] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [23] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.
- [24] D. Guo, F. Wu, F. Zhu, F. Leng, G. Shi, H. Chen, H. Fan, J. Wang, J. Jiang, J. Wang, et al. Seed1. 5-v1 technical report. *arXiv preprint arXiv:2505.07062*, 2025.
- [25] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [26] T. Guo, B. Nan, Z. Liang, Z. Guo, N. Chawla, O. Wiest, X. Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *NeurIPS*, 2023.
- [27] Z. He, X. Wu, P. Zhou, R. Xuan, G. Liu, X. Yang, Q. Zhu, and H. Huang. Cmmu: a benchmark for chinese multi-modal multi-type question understanding and reasoning. In *IJCAI*, 2024.
- [28] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- [29] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.
- [30] K. Huang, T. Fu, W. Gao, Y. Zhao, Y. Roohani, J. Leskovec, C. W. Coley, C. Xiao, J. Sun, and M. Zitnik. Artificial intelligence foundation for therapeutic science. *Nature chemical biology*, 18(10):1033–1036, 2022.
- [31] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

- [32] R. Kessler, G. Narayan, A. Avelino, E. Bachelet, R. Biswas, P. Brown, D. Chernoff, A. Connolly, M. Dai, S. Daniel, et al. Models and simulations for the photometric lsst astronomical time series classification challenge (plasticc). *Publications of the Astronomical Society of the Pacific*, 2019.
- [33] S. Kim and E. E. Bolton. Pubchem: A large-scale public chemical database for drug discovery. *Open Access Databases and Datasets for Drug Discovery*, pages 39–66, 2024.
- [34] S. Kulal, P. Pasupat, K. Chandra, M. Lee, O. Padon, A. Aiken, and P. S. Liang. Spoc: Search-based pseudocode to code. *Advances in Neural Information Processing Systems*, 32, 2019.
- [35] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.
- [36] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2020.
- [37] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [38] H. Li, Z. Wang, J. Wang, A. K. H. Lau, and H. Qu. Cllmate: A multimodal llm for weather and climate events forecasting. *arXiv preprint arXiv:2409.19058*, 2024.
- [39] J. Li, Y. Liu, W. Fan, X.-Y. Wei, H. Liu, J. Tang, and Q. Li. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective, 2024.
- [40] Y. Liang, R. Zhang, L. Zhang, and P. Xie. DrugChat: towards enabling chatgpt-like capabilities on drug molecule graphs. *arXiv preprint arXiv:2309.03907*, 2023.
- [41] Z. Lin, C. Deng, L. Zhou, T. Zhang, Y. Xu, Y. Xu, Z. He, Y. Shi, B. Dai, Y. Song, et al. Geogalactica: A scientific large language model in geoscience. *arXiv preprint arXiv:2401.00434*, 2023.
- [42] P. Liu, Y. Ren, J. Tao, and Z. Ren. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in biology and medicine*, 2024.
- [43] S. Liu, J. Wang, Y. Yang, C. Wang, L. Liu, H. Guo, and C. Xiao. Conversational drug editing using retrieval and domain feedback. In *ICLR*, 2024.
- [44] X. Liu, Y. Guo, H. Li, J. Liu, S. Huang, B. Ke, and J. Lv. DrugLLM: Open large language model for few-shot molecule generation. *arXiv preprint arXiv:2405.06690*, 2024.
- [45] Y. Liu, S. Ding, S. Zhou, W. Fan, and Q. Tan. Moleculargpt: Open large language model (llm) for few-shot molecular property prediction. *arXiv preprint arXiv:2406.12950*, 2024.
- [46] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024.
- [47] Z. Liu, W. Zhang, Y. Xia, L. Wu, S. Xie, T. Qin, M. Zhang, and T.-Y. Liu. Molxpt: Wrapping molecules with text for generative pre-training. *arXiv preprint arXiv:2305.10688*, 2023.
- [48] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.
- [49] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 2022.
- [50] Y. Luo, J. Zhang, S. Fan, K. Yang, Y. Wu, M. Qiao, and Z. Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.
- [51] R. K. Luu and M. J. Buehler. BioinspiredLLM: Conversational large language model for the mechanics of biological and bio-inspired materials. *Advanced Science*, 2024.
- [52] L. Lv, Z. Lin, H. Li, Y. Liu, J. Cui, C. Y.-C. Chen, L. Yuan, and Y. Tian. Prollama: A protein large language model for multi-task protein language processing. *IEEE Transactions on Artificial Intelligence*, 2025.
- [53] C. Ma, Z. Hua, A. Anderson-Frey, V. Iyer, X. Liu, and L. Qin. Weatherqa: Can multimodal language models reason about severe weather? *arXiv preprint arXiv:2406.11217*, 2024.

- [54] A. Madani, B. McCann, N. Naik, N. S. Keskar, N. Anand, R. R. Eguchi, P.-S. Huang, and R. Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.
- [55] A. McNaught. The iupac international chemical identifier. *Chem. Int*, 28(6):12–15, 2006.
- [56] G. Miolo, G. Mantoan, and C. Orsenigo. Electramed: a new pre-trained language representation model for biomedical nlp. *arXiv preprint arXiv:2104.09585*, 2021.
- [57] OpenAI. Gpt, 2023. URL <https://openai.com/gpt>. Large Language Model.
- [58] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [59] L. Phan, A. Gatti, Z. Han, N. Li, J. Hu, H. Zhang, C. B. C. Zhang, M. Shaaban, J. Ling, S. Shi, et al. Humanity’s last exam. *arXiv preprint arXiv:2501.14249*, 2025.
- [60] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.
- [61] T. Schmit, M. Gunshor, G. Fu, T. Rink, K. Bah, and W. Wolf. *GOES-R Advanced Baseline Imager (ABI) Algorithm Theoretical Basis Document for: Cloud and Moisture Imagery Product (CMIP)*. University of Wisconsin–Madison, 2010.
- [62] T. M. Smith, V. Lakshmanan, G. J. Stumpf, K. L. Ortega, K. Hondl, K. Cooper, K. M. Calhoun, D. M. Kingfield, K. L. Manross, R. Toomey, et al. Multi-radar multi-sensor (mrms) severe weather and aviation products: Initial operating capabilities. *Bulletin of the American Meteorological Society*, 2016.
- [63] L. Sun, Y. Han, Z. Zhao, D. Ma, Z. Shen, B. Chen, L. Chen, and K. Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *AAAI*, 2024.
- [64] R. Taconis, M. G. Ferguson-Hessler, and H. Broekkamp. Teaching science problem solving: An overview of experimental work. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 2001.
- [65] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [66] T. Truong Jr and T. Bepler. Poet: A generative model of protein families as sequences-of-sequences. In *NeurIPS*, 2023.
- [67] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 2019.
- [68] L. Wang, Y. Li, X. Xu, and F. Li. Characteristic analysis of dual-polarization weather radar echoes of convective precipitation and snowfall in the mount everest region. *Atmosphere*, 2021.
- [69] R. Wang, X. Fang, Y. Lu, and S. Wang. The pdbname database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004.
- [70] X. Wang, Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A. R. Loomba, S. Zhang, Y. Sun, and W. Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. In *ICML*, 2024.
- [71] Y. Wang, H. Zhao, S. Sciabola, and W. Wang. cMolGPT: a conditional generative pre-trained transformer for target-specific de novo molecular generation. *Molecules*, 2023.
- [72] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *NeurIPS*, 2024.
- [73] N. Webersinke, M. Kraus, J. A. Bingler, and M. Leippold. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, 2021.
- [74] J. W. Wilson, N. A. Crook, C. K. Mueller, J. Sun, and M. Dixon. Nowcasting thunderstorms: A status report. *Bulletin of the American Meteorological Society*, 1998.

- [75] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.
- [76] XAI. Grok, 2023. URL <https://x.ai/grok>. Large Language Model.
- [77] Y. Yan, H. Wen, S. Zhong, W. Chen, H. Chen, Q. Wen, R. Zimmermann, and Y. Liang. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *WWW*, 2024.
- [78] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.
- [79] X. Yue, T. Zheng, Y. Ni, Y. Wang, K. Zhang, S. Tong, Y. Sun, B. Yu, G. Zhang, H. Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.
- [80] Y. Yue, Z. Chen, R. Lu, A. Zhao, Z. Wang, S. Song, and G. Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.
- [81] D. Zhang, W. Liu, Q. Tan, J. Chen, H. Yan, Y. Yan, J. Li, W. Huang, X. Yue, W. Ouyang, et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024.
- [82] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- [83] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, and X. Mao. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [84] W. Zhang, M. Cai, Y. Ning, T. Zhang, Y. Zhuang, H. Chen, J. Li, and X. Mao. Earthgpt-x: Enabling mllms to flexibly and comprehensively understand multi-source remote sensing imagery. *arXiv preprint arXiv:2504.12795*, 2025.
- [85] Y. Zhang, M. Long, K. Chen, L. Xing, R. Jin, M. I. Jordan, and J. Wang. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 2023.
- [86] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, Y. Duan, H. Tian, W. Su, J. Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.
- [87] L. Zhuo, Z. Chi, M. Xu, H. Huang, H. Zheng, C. He, X.-L. Mao, and W. Zhang. Protllm: An interleaved protein-language llm with protein-as-word pre-training. *arXiv preprint arXiv:2403.07920*, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We clearly state our contributions and scope in the both Abstract and Introduction sections.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitations of the work in the Conclusion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: There is no assumptions and proof in this work.

Guidelines:

- The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We fully disclose our experimental settings in the Experiments and Evaluations section. Additionally, all data and code are publicly released for reproduction.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: All code and data are released on HuggingFace. Please check the link in the Abstract section.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: Our experiments are based on lmms-eval. All evaluation details are detailedly described.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[No\]](#)

Justification: Each primary experiment was conducted a single time to obtain the final results. This approach was adopted due to cost considerations, as the API calls involve a substantial financial expenditure.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The majority of the experiments in this work were conducted using APIs provided by various service providers.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: This paper conform NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: This paper discuss positive and negative societal impacts in the Appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [\[Yes\]](#)

Justification: All data are manually inspected to ensure no risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: Creators and original owners of assets used in the paper are all properly credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: New assets introduced in the paper are well documented in the paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing or research with human subjects is involved in this paper. All data are labeled and verified by authors.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No research with human subjects is involved in this paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The LLM usage is declared in the experimental setups.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A PrismaX Team

We provide the full list of PrismaX team members as follows:

- Yuhao Zhou^{3*} (Shanghai Artificial Intelligence Laboratory & Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education & Sichuan University)
- Yiheng Wang^{3*} (Shanghai Artificial Intelligence Laboratory & Shanghai Jiao Tong University)
- Xuming He^{3*} (Shanghai Artificial Intelligence Laboratory & Zhejiang University)
- Ao Shen⁴ (Shanghai Artificial Intelligence Laboratory & Fudan University)
- Ruoyao Xiao (Chinese Academy of Meteorological Sciences)
- Zhiwei Li (Shanghai Artificial Intelligence Laboratory)
- Qiantai Feng (Shanghai Artificial Intelligence Laboratory)
- Zijie Guo (Shanghai Artificial Intelligence Laboratory & Fudan University)
- Yuejin Yang (Shanghai Artificial Intelligence Laboratory & Research Institute of Intelligent Complex Systems, Fudan University)
- Hao Wu (Shanghai Artificial Intelligence Laboratory)
- Wenxuan Huang (Shanghai Artificial Intelligence Laboratory)
- Jiaqi Wei (Shanghai Artificial Intelligence Laboratory & Zhejiang University)
- Dan Si (Sichuan University)
- Xiuqi Yao (Shanghai Artificial Intelligence Laboratory)
- Jia Bu (East China Normal University)
- Haiwen Huang (East China Normal University)
- Tianfan Fu (Shanghai Artificial Intelligence Laboratory & Nanjing University)
- Shixiang Tang (Shanghai Artificial Intelligence Laboratory)
- Ben Fei (Shanghai Artificial Intelligence Laboratory)
- Dongzhan Zhou (Shanghai Artificial Intelligence Laboratory)
- Fenghua Ling (Shanghai Artificial Intelligence Laboratory)
- Yan Lu (Shanghai Artificial Intelligence Laboratory)
- Siqi Sun (Fudan University & Shanghai Artificial Intelligence Laboratory)
- Chenhui Li (East China Normal University)
- Guanjie Zheng (Shanghai Jiao Tong University)
- Jiancheng Lv (Sichuan University)
- Wenlong Zhang[†] (Shanghai Artificial Intelligence Laboratory)
- Lei Bai[†] (Shanghai Artificial Intelligence Laboratory)

[†] Corresponding author: zhangwenlong@pjlab.org.cn, bailei@pjlab.org.cn

³This work was primarily conducted during the authors internship at the Shanghai Artificial Intelligence Laboratory.

⁴Ao Shen is served as an external advisor only.

B Tasks Overview

Table 6: Details of tasks in SFE.

Discipline	Task	TaskID	Subtask	Question Type
Astronomy (8)	Star Perception and Comparison	A001	Galaxy morphology classification	MCQ
		A011	Transient detection	MCQ
	Light Curve Analysis Spectral Analysis	A003	Light curve classification	MCQ
		A006	Surface temperature estimation	Exact Match
		A007	Gravitational constant estimation	Exact Match
		A008	Metallicity estimation	Exact Match
		A009	Alpha-element abundance estimation	Exact Match
		A010	Radial velocity estimation	Exact Match
Chemistry (19)	Structure Analysis	C001	Elemental composition recognition	Exact Match
		C003	IUPAC name recognition	Exact Match
	Property Prediction	C004	Molecular description generation	Open Question
		C005	Lipinski drug-likeness estimation	Exact Match
		C007	Topological polar surface area calculation	Exact Match
		C023	Absorption property prediction	Exact Match
		C024	Distribution property prediction	Exact Match
		C025	Metabolism property prediction	Exact Match
		C026	Excretion property prediction	Exact Match
		C027	Toxicity property prediction	MCQ
		C028	QM9 quantum chemical property prediction	Exact Match
		C029	Protein-ligand binding affinity prediction	Exact Match
	Reaction Prediction	C009	Product SMILES prediction	Exact Match
		C010	Reaction classification	MCQ
		C012	Reactant molecular recognition	Exact Match
		C013	Reaction condition and catalyst prediction	MCQ
	Molecular Design and Optimization	C006	Synthetic accessibility estimation	Exact Match
	Drug screening	C018	Molecular property optimization	MCQ
		C030	Virtual screening	MCQ
Earth (14)	Circulation Understanding	E004	perception of extreme precipitation distribution	MCQ
		E006	Moisture source attribution	MCQ
		E007	subtropical high ridge control region	MCQ
		E008	thermocline depth recognition	MCQ
		E017	Outlier analysis	MCQ

Continued on next page

Continued

Discipline	Task	Subtask ID	Subtask	Question Type
	Multivariate Understanding	E005	Precipitation event analysis	Open Question
		E021	Convective weather types identification	MCQ
		E022	Convective influence regions identification	MCQ
	Remote Sensing Perception	E011	SAR image grounding	Exact Match
		E012	Infrared image grounding	Exact Match
	Ground Data Comparison Reasoning	E014	temperature sequence comparison	MCQ
		E016	Vertical profile comparison	MCQ
		E018 E001	differential prediction comparison Satellite-radar matching	Open Question MCQ
Life Science (14)	Proteomic Sequencing	L001	Fragment ion peaks count	Exact Match
		L003	De novo peptide sequence	Exact Match
	Metabolomic Sequencing	L005	Atom count inference	Exact Match
		L006	Molecular composition inference	Exact Match
	Protein Structure Analysis	L007	Spectrum Matching	MCQ
		L008	Protein chain count	Exact Match
		L009	Small molecule count	Exact Match
		L010	Specified protein detection	Exact Match
	RNA Structure Analysis	L019	Protein structure feature analysis	Open Question
		L014	Structural domains identification	MCQ
		L015	Structural domain count	Exact Match
		L016	RNA type identification	Exact Match
		L017	RNA secondary structure inverse folding	Exact Match
		L020	Structural motifs and positions description	Open Question
Materials Science (11)	Crystal Structure Analysis	M001	Atomic composition description	MCQ
		M002	Crystal group identification	Exact Match
		M003	Crystal formula determination	Exact Match
		M004	Elemental valence state prediction	Exact Match
		M006	Stability estimation	MCQ
	Electronic Property Analysis	M008	Energy band and DOS interpretation	Open Question
		M010	Band gap classification	Open Question
		M018	Valence state and electronic orbital analysis	Open Question
	XRD Pattern Inference	M020	Phase identification	Open Question

Continued on next page

				Continued
Discipline	Task	Subtask ID	Subtask	Question Type
		M021	Lattice constant estimation	Open Question
		M022	Crystal grain size estimation	Open Question

C Tasks Description

Astronomy. In the field of astronomy, raw astronomical observations span multiple modalities, including images, time series, spectra, and hyperspectral observations [1, 32, 2]. Specifically, data collected for celestial objects often covers multiple bands and includes rich attribute information. Therefore, inspired by [3], we divide astronomical tasks into three categories based on data modalities, including *Celestial Image Understanding and Comparison*, *Time-series Light Curve Analysis*, and *Spectral Analysis*, each of which is detailed below.

- *Star Perception and Comparison.* The morphology and brightness of celestial objects provide crucial insights into galaxy formation, which is essential for studying the evolutionary history of stars. For instance, in the *transient detection* sub-task, given the pre-transient and post-transient images, as well as the difference between the two, MLLMs are required to determine whether a transient event has occurred during this process.
- *Light Curve Analysis.* Photometric light curves record the brightness of celestial objects over time, and are crucial for estimating the intrinsic properties and predicting the dynamic behavior of stars. For example, in the *Light Curve Classification* sub-task, MLLMs are provided with time-series photometric data and are required to classify the type of celestial object.
- *Spectral Analysis* Spectral data enables the inference of physical properties such as surface temperature and chemical composition. Therefore, we have designed 5 sub-tasks based on different physical properties. For example, in the *Surface Temperature Estimation* task, MLLMs are required to estimate the surface temperature of a target celestial object based on its spectral profile.

Chemistry. To systematically evaluate MLLMs performance in the field of chemistry, we organize our benchmark around **four major capability categories**: *Structure Analysis*, *Molecular Property Prediction*, *Chemical Reaction Prediction*, and *Molecular Design and Optimization*. These capabilities reflect fundamental scientific questions central to chemical research and application.

- *Structure Analysis.* Structure Analysis tests a models grasp of symbolic chemistry, such as identifying elemental composition, describing molecules using natural language, and generating IUPAC (International Union of Pure and Applied Chemistry) names (*e.g.*, methane, ethane) [55]. It examines the foundational ability to ensure that the model correctly recognizes the chemical structure and elements of the molecules.
- *Property Prediction.* Molecular Property Prediction aims to assess a models ability to infer chemical and biological properties from molecular structures. It focuses on the properties of chemical molecules that scientists are most interested in, such as drug molecules’ ADMET (*i.e.*, absorption, distribution, metabolism, excretion, toxicity) [19], synthetic accessibility, and quantum mechanics energy, etc.
- *Reaction Prediction.* Chemical Reaction Prediction assesses the MLLM’s understanding of chemical reactions, including the products of the reaction, reaction centers (the region within a molecule where bond breaking and forming occur), and required reaction conditions (*e.g.*, catalyst).
- *Molecular Design and Optimization.* Molecular Design and Optimization evaluates the MLLMs capacity to generate molecules with desirable chemical properties (*e.g.*, pharmaceutical properties). These tasks focus on identifying novel and diverse molecules with desirable molecular properties (*e.g.*, druglikeness, solubility, binding affinities to the target proteins), which are fundamental tasks in drug and material discovery [18].
- *Drug Screening.* Drug Screening evaluates the MLLMs ability to identify biologically relevant candidate molecules based on protein sequences and molecular structure information [30].

Earth. The Earth science domain includes various data modalities. First, variables can be either temporal or static. Meteorological time series are often used to analyze strong convective weather events [85, 21, 74]. Static attributes, such as long-term average temperature maps, are typically used to capture climate patterns. Second, strong correlations often exist between variables. For example, the position of the western Pacific subtropical high (*i.e.*, the 5880 hPa line) is positively correlated with the coverage range of precipitation belts. Third, variables may originate from diverse sources, including ground-based weather radar [62], multi-spectral satellite imagers [61], and various remote sensing instruments. To cater to such diverse scenarios, we establish a comprehensive Earth science task paradigm that focuses on 4 tasks, each comprising multiple sub-tasks.

- *Task1: Circulation Understanding.* In atmospheric physics, large-scale circulation patterns play a central role in modulating regional weather phenomena, such as precipitation anomalies. For example, in the *moisture source attribution* sub-task, MLLMs are required to reason the origin of moisture contributing to the regional rainfall by integrating information from moisture fluxes and wind fields.
- *Task2: Multivariate Understanding.* Complex weather phenomena often arise from the interplay among multiple meteorological variables. For example, in the *precipitation event analysis* sub-task, MLLMs are required to consider precipitation, temperature, wind direction, moisture flux, and geopotential height collectively to infer the underlying causes of rainfall events.
- *Task3: Remote Sensing Perception.* Remote sensing imagery provides multispectral observations of Earth’s surface, enabling comprehensive monitoring of environmental and urban conditions. To assess the capability of MLLMs in identifying object categories and localizing their spatial positions from satellite data, we design two sub-tasks using different image modalities. MLLMs are required to first count the number of objects belonging to a specified category, and then identify the spatial location of each instance using the bounding box.
- *Task4: Ground Data Comparison Reasoning.* Ground observational data have been widely used in scenarios such as global climatological analysis and extreme precipitation monitoring [68, 9]. To evaluate the capability of MLLMs in comparing and reasoning meteorological terminology and data distributions, we define 4 tasks based on different variables. For example, in the *temperature sequence comparison task*, MLLMs need to compare annual temperature series from two different time periods and describe differences of statistical characteristics

Life. To access the capabilities of MLLMs in the domain of Life Science, we define **three core categories**: *Biomolecular Profiling*, *Sequence Reasoning*, and *Structure Interpretation*. These capabilities reflect core scientific practices in the Life Science, where understanding arises from linking molecular measurements, symbolic sequences, and structural representations to functional biological meaning.

- *Biomolecular Profiling.* Biomolecular Profiling involves extracting quantitative and compositional features of biological molecules. These profiling skills are fundamental in high-throughput biological workflows such as mass spectrometry and molecular diagnostics, where precise characterization of biomolecular species is critical for downstream analysis.
- *Sequence Reasoning.* Sequence Reasoning addresses the functional interpretation of biological sequences, including proteins and RNAs. Such reasoning requires models to understand sequence-function relationships, modularity, and constraints that are central to biological information flow and molecular design.
- *Structure Interpretation.* Structure Interpretation targets the analysis of spatial and symbolic representations of biological macromolecules. These tasks emphasize spatial abstraction and multi-view understanding, as biological structures are often represented through hybrid formats that combine 2D, symbolic, and textual information.

Materials Science. Materials science presents unique challenges grounded in structural periodicity, quantum-level electronic properties, and empirical characterization signals. To reflect these, we design tasks around **three core capabilities**: *Crystal Structure Analysis*, *Electronic Property Analysis*, and *XRD Pattern Inference*.

- *Crystal Structure Analysis.* Crystal Structure Analysis involves symbolic, compositional, and descriptive reasoning over crystalline systems. These tasks reflect how materials scientists represent, classify, and formally describe the structural foundations of solids, where a deep

understanding of symmetry operations, periodicity, and the thermodynamic and structural stability conditions is required.

- *Electronic Property Analysis.* Electronic Property Analysis aims at interpreting the electronic behavior of a material through visual and symbolic representations. These tasks require models to bridge diagrammatic understanding with functional inference, emulating expert workflows in electronic materials analysis and property prediction.
- *XRD Pattern Inference.* XRD Pattern Inference focuses on scientific reasoning grounded in experimental diffraction signals. These tasks require models to connect peak pattern distributions with structural parameters through symbolic and numerical reasoning.

D Question Format

For multiple-choice questions, the question is formatted as:

Multiple-choice questions format example

You are an expert in discipline and need to solve the following question. The question is a multiple-choice question. Answer with the option letter from the given choices.

{task prompt}

{question}

{options}

For exact match questions, the question is formatted as:

Exact match questions format example

You are an expert in discipline and need to solve the following question. The question is an exact match question. Answer the question using a single word or phrase.

{task prompt}

{question}

For open-ended questions, the question is formatted as:

Exact match questions format example

You are an expert in discipline and need to solve the following question. The question is an open-ended question. Answer the question using a phrase.

{task prompt}

{question}

Finally, for bounding boxes extraction questions, the question is formatted as:

bounding boxes extraction questions format example

You are an expert in discipline and need to solve the following question. The question is an open-ended question. Answer the question using a phrase.

Each bounding box is represented by four numbers, corresponding to the positions x_{min} , y_{min} , x_{max} , and y_{max} . The coordinate origin is located at the top-left corner of the image, and the bottom-right corner has coordinates (1, 1). Therefore, both x and y range from 0 to 1. In the final output, two bounding boxes are separated by a semicolon, and all bounding boxes are enclosed in square brackets. Here is an example output: "[$x1_{min}$, $y1_{min}$, $x1_{max}$, $y1_{max}$; $x2_{min}$, $y2_{min}$, $x2_{max}$, $y2_{max}$]" (no quotation marks).

{task prompt}

{question}

E Responses Evaluation

Given the MLLM's responses, the ground truth answer and the problem, we ask GPT-4o to judge the correctness with the following prompt.

LLM-as-a-Judge prompt

You are a strict evaluator assessing answer correctness. You must score the model's prediction on a scale from 0 to 10, where 0 represents an entirely incorrect answer and 10 indicates a highly correct answer.

Input

Question:

{question}

Ground Truth Answer:

{answer}

Model Prediction:

{prediction}

Evaluation Rules

- The model prediction may contain the reasoning process, you should spot the final answer from it.
- For multiple-choice questions: Assign a higher score if the predicted answer matches the ground truth, either by option letters or content. Include partial credit for answers that are close in content.
- For exact match and open-ended questions:
 - * Assign a high score if the prediction matches the answer semantically, considering variations in format.
 - * Deduct points for partially correct answers or those with incorrect additional information.
- Ignore minor differences in formatting, capitalization, or spacing since the model may explain in a different way.
- Treat numerical answers as correct if they match within reasonable precision
- For questions requiring units, both value and unit must be correct

Scoring Guide

Provide a single integer from 0 to 10 to reflect your judgment of the answer's correctness.

Strict Output format example

4