CURIE: Evaluating LLMs on Multitask Scientific Long-Context Understanding and Reasoning

Hao Cui[1,⋆] **Zahra Shamsi**[1,⋆] **Gowoon Cheon**[1] **Xuejian Ma**[1] **Shutong Li**[1] **Maria Tikhanovskaya**[2,◇] **Peter Norgaard**[1] **Nayantara Mudur**[2,◇] **Martyna Plomecka**[3,◇] **Paul Raccuglia**[1] **Yasaman Bahri**[1] **Victor V. Albert**[4,5] **Pranesh Srinivasan**[1] **Haining Pan**[6] **Philippe Faist**[7] **Brian Rohr**[8] **Michael J. Statt**[8] **Dan Morris**[1] **Drew Purves**[1] **Elise Kleeman**[1] **Ruth Alcantara**[1] **Matthew Abraham**[1] **Muqthar Mohammad**[1] **Ean Phing VanLee**[1] **Chenfei Jiang**[1] **Elizabeth Dorfman**[1] **Eun-Ah Kim**[9] **Michael Brenner**[1,2] **Viren Jain**[1] **Sameera Ponda**[1] **Subhashini Venugopalan**[1,⋆,†]
[1]Google, [2]Harvard, [3]University of Zurich, [4]NIST, [5]UMD College Park,
[6]Rutgers, [7]FU Berlin, [8]Modelyst, [9]Cornell

Abstract

Scientific problem-solving involves synthesizing information while applying expert knowledge. We introduce CURIE, a scientific long-Context Understanding, Reasoning, and Information Extraction benchmark to measure the potential of Large Language Models (LLMs) in assisting scientists in realistic experimental and theoretical workflows. This benchmark introduces ten challenging tasks curated by experts in six disciplines: materials science, condensed matter physics, quantum computing, geospatial analysis, biodiversity, and proteins. We evaluate a range of closed and open LLMs on tasks in CURIE which requires domain expertise, comprehension of long in-context information, and multi-step reasoning. While Claude-3 shows consistent high comprehension across domains, the popular GPT-4o and command-R+ fail dramatically on protein sequencing tasks. Overall there is much room for improvement for all models. We hope this work can guide the future development of LLMs in sciences.

1 Introduction

The advancement of science relies on the ability to build upon the collective knowledge accumulated in scientific literature, requiring not only deep domain expertise and reasoning skills, but also the capacity to apply that knowledge within the context of a given problem. Recent benchmarks (e.g., MMLU [1]) have demonstrated proficiency in varied subjects. However, as LLMs transition from merely surfacing knowledge to actively solving problems, the capacity to understand and reason about long-form, context-rich information is paramount. Recent advances in model architecture have seen dramatic increases in context windows from 8k to 32k, 128k, and 1M+ tokens, reflecting a growing recognition of this need.

This has led to development of benchmarks testing capabilities of LLMs on long document understanding e.g. ZeroScrolls [2], Bamboo [3] on a variety of tasks including summarization [4], retrieval [5], multi-hop QA [6, 7], sorting sequences [8] and others [9]. However, current LLM benchmarks on science e.g., PubmedQA [10] or GPQA [11], focus primarily on short sequence questions, with answers often in multiple-choice form. To address this gap, we introduce the scientific long-Context Understanding, Reasoning, and Information Extraction benchmark (CURIE).

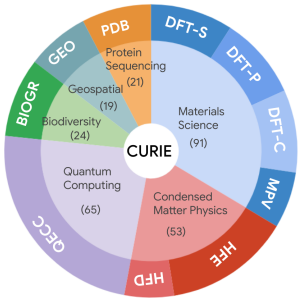
The CURIE benchmark encompasses 434 examples with human annotated ground truth across 10 tasks curated from 273 research papers in six diverse scientific disciplines (Fig. 1): materials science,

⋆equal technical contribution, ◇work done as a student researcher at Google Research.
†Lead and corresponding author (vsubhashini@google.com)

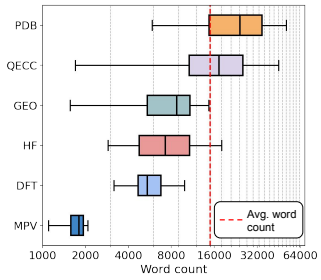
Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

Benchmark	Source Domain	Diverse Tasks	Long Context
ZeroSCROLLS	various	✓	✓
Bamboo	various	✓	✓
Ada-Leval	lit., code	QA,sort	✓
Ruler	synthetic	✓	✓
BLURB	biomed	✓	✗
QASA	ML lit.	QA	✓
Qasper	NLP lit.	QA	✓
GPQA	science	MCQ	✗
Curie	sci. lit.	✓	✓

(a) Comparison with existing datasets.



(b) Distribution of tasks.



(c) Input word lengths per domain.

Figure 1: **CURIE dataset.** (a) CURIE introduces diverse long context tasks on scientific literature (lit.). (b) Distribution of research documents in each discipline: 434 examples were curated from 273 research papers. (c) Length of input context in each domain (log scale).

theoretical condensed matter physics, quantum computing, geospatial analysis, biodiversity, and proteins – covering experimental and theoretical aspects of scientific research. These tasks not only require deep domain understanding but also challenge models on their capacity to comprehend full-length scientific papers for information extraction, concept tracking, aggregation, multimodal understanding, and cross-domain expertise (e.g., generating code for theoretical calculations).

We use the CURIE testbed to perform extensive evaluation and analysis of 8 state-of-the-art open and closed weight models (see Fig. 3) supporting context windows of 32k tokens or more. Among closed models, Claude-3 Opus performs consistently well across all disciplines, while Command-R+ does better amongst the open models. Surprisingly, while GPT-4o does well on most tasks, it fares dramatically poorly on the protein sequencing task, failing to stop generation and repeating subsequences to yield a very low score. This repetition in subsequence is not unique to GPT-4o, indicating that such tasks and data are not well represented in standard language datasets. Our materials tasks, which require models to exhaustively retrieve and aggregate information spread through the document, also prove to be exceptionally challenging for all models.

While the CURIE benchmark is aimed at facilitating evaluation of scientific reasoning over long contexts, we hope the rich human annotations can serve the community in advancing planning, instruction following, and evaluation of generated texts of mixed and heterogeneous formats including dates, locations, numerical values, units, descriptors, domain specific terms, equations and code.

2 CURIE dataset and tasks

The CURIE benchmark consists of a series of tasks that measure how well LLMs can assist in diverse scientific workflows, from synthesis of information towards final execution anchored on single scientific research documents. All tasks are (1) realistic and require scientific expertise, (2) require comprehension of substantial context, e.g., a research paper, and (3) can be evaluated by experts to highlight potential/limitations of models. We provide a brief motivation for each domain and the description of the tasks below with some examples in Fig. 2. Details of curation guidelines are in Appendix C.

Density Functional Theory Task (DFT). Density Functional Theory (DFT) is a robust framework for quantum mechanical modeling of materials, enabling first-principles predictions and validation of experimental findings. We define 3 tasks that measure the ability of LLMs to carry out DFT calculations: (1) extracting input material structures (DFT-S); (2) identifying DFT parameters associated with computation steps (DFT-P); and (3) translating computational steps essential for reproducing key results from the paper into functional code (DFT-C). Executing all these tasks successfully requires the LLM to comprehend domain-specific concepts, extract information dispersed across different sections of the publication, and generate scientific code. Our benchmark contains 75 papers with expert annotations.

Material Property Value Extraction (MPV). The published literature is an untapped resource with experimentally reported materials, properties, processing conditions and structure. Human curation is time intensive and expensive, and rule-based automation is limited in scope [12]. However, prompt-

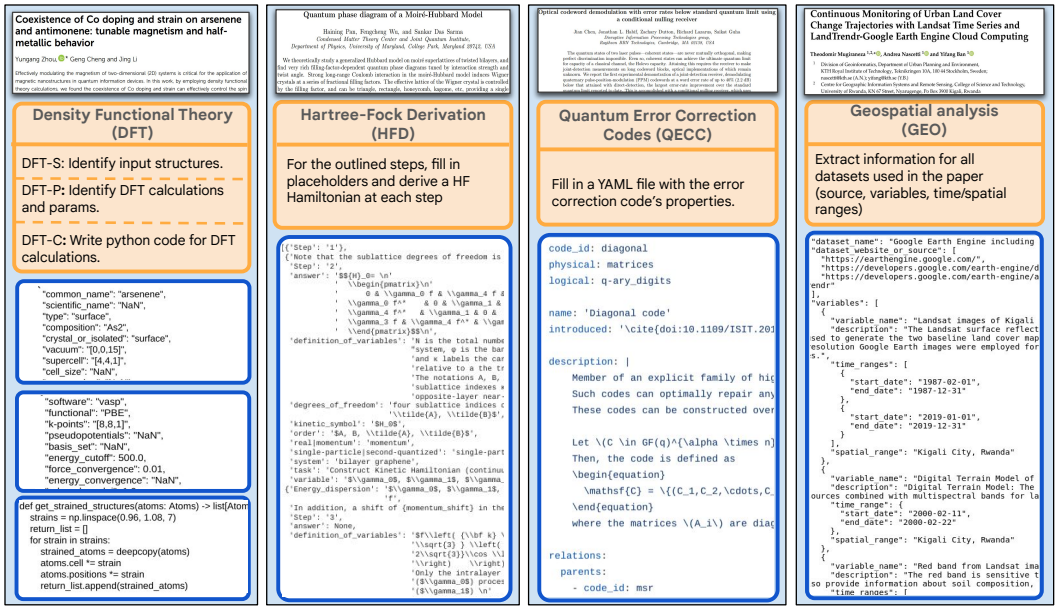


Figure 2: **Examples of tasks in the CURIE benchmark.** The DFT, HFD, QECC, and GEO tasks require the LLM to perform tasks on scientific papers (top blocks), as described in the prompt snippets (in orange), to extract, calculate, or aggregate information. Expected output (ground truth) snippets are shown in the blue blocks. (Only snippets of prompts /outputs are shown for illustrative purposes.)

based LLM extraction has shown promising early results [13]. Our benchmark contains 17 scientific papers for exhaustively extracting material properties. The main task is to identify all instances of material properties mentioned in the text, including material name, descriptor and particular property, along with the passage or table where the property is described.

Hartree-Fock Tasks (HFD, HFE). Hartree-Fock mean-field theory is a framework for simplifying mathematical descriptions of interacting quantum systems. We construct two tasks: derivation (HFD) and extraction (HFE). HFD measures the ability of LLM to derive the Hartree-Fock mean-field Hamiltonian for a quantum many-body system, motivated by prior work [14]. Deriving the correct answer requires 13-19 reasoning steps, making it extremely challenging without expert oversight. The second simpler task, HFE, evaluates an LLM's ability to identify and aggregate key equations from a research paper to extract the most general mean-field Hamiltonian. We have 53 papers (38 HFE, 15 HFD) with expert annotations including prompts for detailed reasoning.

Error Correction Zoo Task (QECC). The Error Correction Zoo [15] (EC Zoo) is an open source effort to build a Wikipedia-like repository collecting and categorizing error correcting codes from the literature. Creating an entry in the EC Zoo is a knowledge intensive process and requires listing the properties of a given EC code, which often include bespoke technical details, along with any relations to other codes in literature. We construct a benchmark of 65 papers that tests the ability of LLMs to curate the EC Zoo entries by taking a given paper and asking it to produce YAML file summaries.

Geospatial Dataset Extraction (GEO). Geospatial analysts integrate various datasets to answer complex questions. For example, a study of time-series snowmelt detection over Antarctica may combine satellite imagery, radar, weather station temperature data, topography information, etc. [16]. In this task, given a research paper, the LLM is required to identify all utilized datasets, including source websites, variable names, descriptions, time ranges and spatial ranges that may be scattered across the paper. Our benchmark includes 19 papers ranging across earth observation, economics, epidemiology and public health, along with detailed ground truth annotations necessary to reproduce each study.

Biodiversity Georeferencing Task (BIOGR). Critical geospatial information is often conveyed exclusively through maps. In this task, we investigate the core capability of georeferencing, where, given an image of a map and its associated caption, the task is to determine the latitude/longitude bounding box encompassing the region displayed. A domain expert would often use a multi-step process and specialized mapping tools (e.g., QGIS, ArcGIS), zooming in and switching between

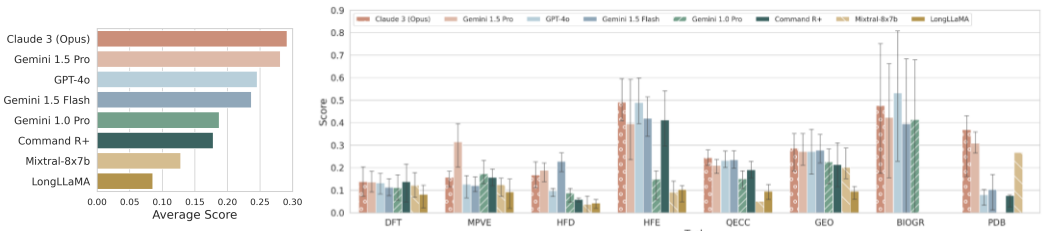


Figure 3: (a) Avg. performance of long-context LLMs across 10 tasks from six scientific domains in CURIE. (b) Per task normalized scores of various LLMs on CURIE (The 3 DFT tasks are averaged).

different imagery layers, etc. For this multimodal task, we assembled a dataset of 24 map images and captions from papers in ecology, of varying difficulty with ground truth labels for bounding boxes.

Protein Sequence Reconstruction (PDB). This final task tests the ability of an LLM to infer functions based on the structure of the protein. Specifically, given the 3D structural coordinates capturing the precise arrangement of atoms, we ask the LLM to reconstruct the protein's amino acid sequence. The data, 21 structures, curated from the Protein Data Bank (PDB) is stripped of any explicit functional annotation forces the LLM to rely on its understanding of structural patterns to deduce the underlying amino acid sequence.

3 Evaluation Setup and Results

Experimental Setup. We evaluate the CURIE benchmark tasks on several state-of-the-art LLMs supporting long-context windows, including five closed weight LLMs such as GPT-4o [17], Claude-3 Opus [18], Gemini 1.5 Pro [19], and three open-weight LLMs, Mixtral [20], Command-R-+ [21], and LongLLaMa-3B [22]. We follow a standard zeroshot prompt template across tasks, which describes the task, output format, and the full text of the paper. In the case of DFT, MPV, and GEO tasks, we provide the output format with a hand-crafted excerpt to clarify expectation of formats for each field. The BIOGR task is multimodal, and for this we provide just the image and caption as input, rather than the full paper. Performance is reported for each model on each task using a single run, except for BIOGR which is averaged over three runs due to observed variability.

Evaluation metrics. For the tasks requiring long text generation (8 of 10 tasks), we use ROUGE-L [23] and BERTScore F1 [24] metrics. BIOGR uses Intersection-over-Union (IoU), which when computed using latitude and longitude accounts for location and size while being scale-invariant. For the PDB task, we compare reconstructed sequences to ground truth [25], using pairwise sequence alignment scored using the number of identities. The raw scores are normalized by the alignment length to account for potential length discrepancies, yielding the identity ratio (ID_r) metric. Further, we compute **average** across all 10 tasks by normalizing ROUGE, IoU and ID_r to be in [0, 1] range. We also introduce two model based evaluations discussed in the supplement.

Main Results. Fig. 3(a) shows the performance of all models averaged across all tasks in the CURIE benchmark. Claude-3 Opus is the best performing with consistent high performance across all tasks. Fig. 3(b) shows task level performance of all models. The popular GPT-4o outperforms the others on GEO and BIOGR, but it's performance on PDB and HFD is surprisingly low. On closer inspection, we found that the GPT-4o model exhibited repetition in the outputs (in the PDB task), clipped responses (in the MPV task), and failed to follow formatting instructions on the HFD task, leading to lower performance. Overall though, most of the closed models had very close performance, and given the variability (e.g., 25%-75% error bars around the mean in Fig. 3), the difference between them is not significant. On several tasks there is considerable room for improvement, making CURIE an interesting benchmark for furthering model development. Appendix A includes more detailed results.

Appendix B includes model based evaluations, and Appendix A.2 has performance sliced based on difficulty of each example on each task, where difficulty is judged by the domain expert.

4 Discussion

Model responses lack robustness in instruction following. A common observation across tasks was that, there is variability across model runs, even though average performance remained fairly constant. The variability is usually higher on harder tasks. Instruction following remains a challenge: Models often had *pieces* of the right answer, but were unable to consistently format it despite examples

in the prompt. In rare cases, even though we explicitly asked models to provide answers when they weren't sure, they often refused to venture an educated guess. The model-based evaluation metrics were quite helpful in mitigating issues arising from lack of adherence to instructions.

Performance on retrieval. Fig. 3(b) and Table 1 report performance of the models on each task in each of the domains. Noticeably all models show high ROUGE-L scores on HFE which is a variation of the needle-in-a-haystack problem where the model needs to extract related equations that might be spread throughout the paper. On tasks requiring exhaustive retrieval of multiple values and aggregation, e.g., DFT, MPV, and GEO, the models have considerably lower performance than just single value retrieval tasks (e.g., HFE).

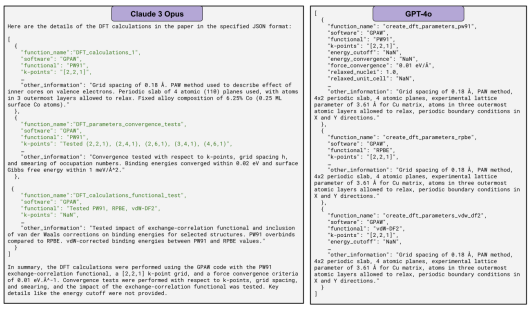


Figure 4: **Example of model outputs for the DFT-P** parameter identification task. Claude-3 Opus appears to understand the purpose of the calculations better than the other models and avoids unnecessary repetition. Claude-3 correctly (green) identifies that there is one set of DFT parameters used in the actual study as well as two more set of parameters which are used for convergence testing.

Concept tracking, aggregation, summarization. On tasks requiring concept aggregation and tracking, e.g. DFT-P, GEO and QECC, experts found responses from some models quite promising. With DFT-P, Claude-3 appeared to understand the purpose of DFT calculations better and grouped relevant parameters to appropriate functions. On QECC, the experts noted that the summaries generated by the LLM tended to be succinct while also including multitude of key informational "nuggets" and quantitative measurements. While not all of these were correct or important, experts noted that it would be easier to exclude the wrong bits (after examination) but harder to extract and comb out such details from the paper. On the GEO task, the closed models did well to extract some of the important datasets with the correct spatial and temporal ranges but performance degrades when multiple datasets are used to cover a larger spatial extent. Overall, carefully engineered prompts and agentic workflows could be effective on such tasks.

Closed vs Open models. One thing of note is that on QECC, DFT, and MPV extraction and aggregation tasks, the Command-R+ open weights model which uses retrieval-augmented approaches shows performance similar to the closed weight models. On PDB, Mixtral performed higher than GPT-4o which is quite surprising. Both LongLLaMA and Command-R+, failed to produce any sort of FASTA format on the PDB task. They either failed to fully understand the task, or missed one of the steps in aggregating the amino acid sequences. Also, the evals on BIOGR highlight that open models are yet to support both multimodal and long-context capabilities which can enable more scientific applications. Overall, across tasks, model performance has room for improvement.

5 Conclusion

In this work, we introduce the CURIE benchmark — a series of tasks designed to measure the ability of LLMs on understanding long-context scientific reasoning. Our main contributions are (i) A new benchmark of 434 examples from 273 research papers that can assess LLMs on comprehension of long-context information from across six scientific disciplines requiring deep expertise. (ii) 10 realistic tasks combining concept retrieval and extraction, and more to measure capability of models on different aspects of scientific workflows. (iii) We propose model-based evaluation metrics for mixed-format heterogeneous outputs and share guidelines for curation and annotation in the supplement. We hope the diverse tasks and rich annotations in the CURIE benchmark can serve the community in not only evaluating LLMs on their scientific problem-solving abilities but also advance research on scientific planning, instruction following, and evaluation of generated texts containing information of diverse types and formats.

Impact Statement and Limitations

Curation of expert annotated datasets on tasks from scientific domains is time intensive and expensive. While the CURIE benchmark is aimed at facilitating evaluation of scientific reasoning over long contexts, we hope the rich human annotations can serve the community in advancing planning, instruction following, and evaluation of generated texts of mixed and heterogeneous formats including dates, locations, numerical values, units, descriptors, domain specific terms, equations and code.

This work focused on a select set of domains and a narrow set of tasks with high quality annotations, thus limiting the scale. Increasing the scale of examples across tasks would provide a more robust evaluation benchmark. With the fast pace of language model advancements, evaluating the generated text responses on such complex tasks is challenging even with high quality human annotations. In particular, just based on instructions and output format provided in the prompt, existing automated evaluation metrics Rouge-L and BERTScore can be unforgiving resulting in low scores for responses that look different but might still be reasonable. While we propose model based evaluation metrics, these are still far from perfect and provides room for more creative strategies. Further, we primarily evaluate models in the zero-shot and two-shot settings, and we invite researchers to explore retrieval augmented generation and chained prompting strategies that evaluate the models on planning and task decomposition.

Disclaimer

V.V.A. participated only as a subject-matter expert for the QECC task. Certain equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[2] Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. Zeroscrolls: A zero-shot benchmark for long text understanding. *arXiv preprint arXiv:2305.14196*, 2023.

[3] Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv preprint arXiv:2309.13345*, 2023.

[4] Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*, 2021.

[5] Gregory Kamradt. Needle in a haystack - pressure testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack/tree/main, 2023.

[6] Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. Fanoutqa: Multi-hop, multi-document question answering for large language models. *arXiv preprint arXiv:2402.14116*, 2024.

[7] Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

[8] Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. Ada-leval: Evaluating long-context llms with length-adaptable benchmarks. *arXiv preprint arXiv:2404.06480*, 2024.

[9] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

[10] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *EMNLP*, pages 2567–2577, 2019.

[11] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

[12] Matthew C Swain and Jacqueline M Cole. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling*, 56(10):1894–1904, 2016.

[13] Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. Chatgpt chemistry assistant for text mining and prediction of mof synthesis. *arXiv preprint arXiv:2306.11296*, 2023.

[14] Haining Pan, Nayantara Mudur, Will Taranto, Maria Tikhanovskaya, Subhashini Venugopalan, Yasaman Bahri, Michael P. Brenner, and Eun-Ah Kim. Quantum many-body physics calculations with large language models. *arxiv 2403.03154*, 2024.

[15] Victor V. Albert and Philippe Faist, editors. *The Error Correction Zoo*. 2024.

[16] Dong Liang, Huadong Guo, Lu Zhang, Yun Cheng, Qi Zhu, and Xuting Liu. Time-series snowmelt detection over the antarctic using sentinel-1 sar images on google earth engine. *Remote Sensing of Environment*, 256:112318, 2021.

[17] OpenAI. Hello GPT-4o. Available online at: https://openai.com/index/hello-gpt-4o/.

[18] Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. Available online at: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

[19] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

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

[21] Cohere. Introducing Command R+: A Scalable LLM Built for Business. Available online at: https://cohere.com/blog/command-r-plus-microsoft-azure.

[22] Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. Focused transformer: Contrastive training for context scaling. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[24] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[25] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422, 2009.

[26] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

[27] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

[28] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

[29] Ehsan Kamalloo, Nouha Dziri, Charles LA Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. *arXiv preprint arXiv:2305.06984*, 2023.

[30] Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179, 2024.

[31] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[32] Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. *arXiv preprint arXiv:2401.06591*, 2024.

[33] David Donoho. Data science at the singularity. *Harvard Data Science Review*, 6(1), 2024.

[34] Qingyang Dong and Jacqueline M Cole. Auto-generated database of semiconductor band gaps using chemdataextractor. *Scientific Data*, 9(1):193, 2022.

[35] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.

[36] Juraj Mavracic, Callum J Court, Taketomo Isazawa, Stephen R Elliott, and Jacqueline M Cole. Chemdataextractor 2.0: Autopopulated ontologies for materials science. *Journal of Chemical Information and Modeling*, 61(9):4280–4289, 2021.

[37] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *EMNLP*, 2018.

[38] J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182, 2003.

[39] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access, 2018.

[40] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406, 2018.

[41] Arman Cohan, Waleed Ammar, Madeleine Van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. *arXiv preprint arXiv:1904.01608*, 2019.

[42] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *EMNLP*, 2019.

[43] Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. Chemprot-3.0: a global chemical biology diseases mapping. *Database*, 2016:bav123, 2016.

[44] Xiang Zhang, Zichun Zhou, Chen Ming, and Yi-Yang Sun. Gpt-assisted learning of structure-property relationships by graph neural networks: Application to rare-earth doped phosphors. *arXiv preprint arXiv:2306.14238*, 2023.

[45] Qingyang Dong and Jacqueline M Cole. Auto-generated database of semiconductor band gaps using chemdataextractor. *Scientific Data*, 9(1):193, 2022.

[46] Maciej Polak and Dane Morgan. *Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering -Example of ChatGPT*.

[47] Martin Krallinger, Anastasia Krithara, Anastasios Nentidis, Georgios Paliouras, and Marta Villegas. Bioasq at clef2020: Large-scale biomedical semantic indexing and question answering. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 550–556. Springer, 2020.

[48] Wai-Chung Kwan, Xingshan Zeng, Yufei Wang, Yusen Sun, Liangyou Li, Lifeng Shang, Qun Liu, and Kam-Fai Wong. M4le: A multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models. *arXiv preprint arXiv:2310.19240*, 2023.

[49] Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*, 2023.

[50] Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*, 2023.

[51] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

[52] Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. Qasa: advanced question answering on scientific articles. In *ICML*, pages 19036–19052. PMLR, 2023.

[53] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *NAACL*, pages 4599–4610, 2021.

A Detailed results

A.1 Quantitative comparisons.

Comparison with existing benchmarks and room for improvement. Fig. 5 (b) compares performance of models from two different generations, Gemini 1.0 pro (32k) and Gemini 1.5 pro (1M+ context window) on popular benchmarks evaluating linguistic capability (DROP) [26], breadth of knowledge [1], and expertise in science [11], alongside the performance on our benchmark evaluating expertise with long-context comprehension. We observe that there is considerable room for improvement on the types of realistic complex scientific tasks the CURIE benchmark provides.

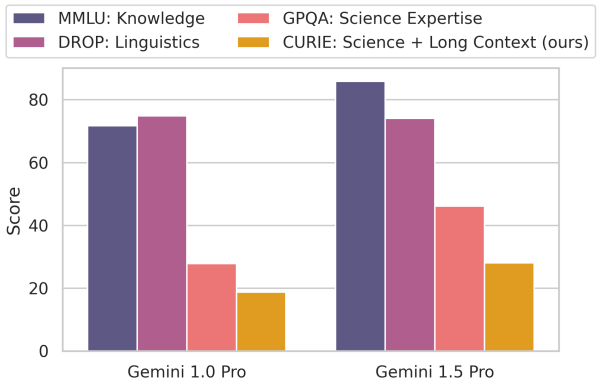


Figure 5: Comparing performance of different generation models supporting long-context windows on previous benchmarks testing Knowledge, Linguistic, and Science expertise, alongside our new scientific long-context understanding CURIE benchmark, highlighting current difficulty of the tasks and the role benchmarks play in advancing LLM capabilities.

Main Results Table. Table 1 show task level performance of all models on the ROUGE-L and BERTScore-F1 metrics. On scientific long-context tasks, Claude-3 Opus performas well across all tasks showing strong multitask multi-domain capabilities followed by Gemini 1.5 pro. The popular GPT-4o dramatically underperforms on the PDB task.

Method	**DFT**		**MPV**		**HFD**		**HFE**		**QECC**		**GEO**		**BIOGR**	**PDB**
	R-L	B-F1	R-L	B-F1	R-L	B-F1	R-L	B-F1	R-L	B-F1	R-L	B-F1	IoU	ID_r
Zero-shot Open Weight LLMs														
Mixtral	12.2	0.67	12.48	0.7	3.78	0.75	9.15	0.63	5.11	0.69	20.23	0.77	-	0.27
Command-R+	13.79	0.64	15.67	0.75	5.93	0.75	41.23	0.83	19.12	0.67	21.36	0.78	-	0.08
LongLLaMa	8.17	0.6	9.28	0.66	4.36	0.63	10.33	0.63	9.53	0.6	9.53	0.68	-	-
Zero-shot Closed Weight LLMs														
Gemini 1.0 Pro	11.22	0.63	17.37	0.75	8.72	0.66	14.91	0.69	15.08	0.63	22.56	0.77	0.41	-
GPT-4o	13.3	0.64	12.74	0.73	9.81	0.74	48.93	0.85	23.23	0.66	27.3	0.8	0.53	0.08
Gemini 1.5 Pro	13.66	0.65	31.54	0.79	18.86	0.79	39.56	0.84	21.04	0.63	27.24	0.78	0.42	0.31
Gemini 1.5 Flash	11.31	0.64	12.11	0.77	22.86	0.79	41.92	0.86	23.5	0.70	27.76	0.8	0.4	0.1
Claude 3 (Opus)	13.78	0.63	15.86	0.75	16.82	0.76	49.1	0.86	24.44	0.68	28.66	0.79	0.48	0.37

Table 1: **Results comparing performance of all models on all tasks based on automated metrics** R-L: Rouge-L, and B-F1:BertScore-F1. The avg. performance of all 3 DFT tasks are reported under DFT. All models support a context length of 32k or more. BIOGR has multimodal inputs which is unsupported by the chosen open models. Blue highlights the highest values across closed models.

A.2 Performance vs. Difficulty

For each of the tasks, for each input example, we requested experts to rate the difficulty of answering the example query as easy, medium or hard. Experts in all domains independently determined the rating scale. Surprisingly, they all chose a similar scale for determining difficulty of the example. For all tasks difficulty was measured based on how wide spread the requested information was

within the paper. If most of the requested information was available within a few paragraphs or a page, the example was rated as easy, if the information was spread over multiple parts of the paper the example was rated as medium difficulty, and if the information required knowledge of specific literature outside of the given context (e.g. based on referenced papers), the example was rated hard. Additionally for the DFT-C code generation task, the ratio of the number of implementable functions to the total number of functions mentioned in the paper was used to determine difficulty. For HFD, the number of reasoning steps was used in determining difficulty. Fig. 6 reports performance of each model sliced by difficulty. Overall, models perform substantially better on easy examples compared to the medium and hard examples. Models appear to perform about the same on examples marked medium or hard. Though, one thing of note is that there are usually many more medium examples than hard examples across all tasks. We will include the distribution and the ratings for each of the examples in the supplement.

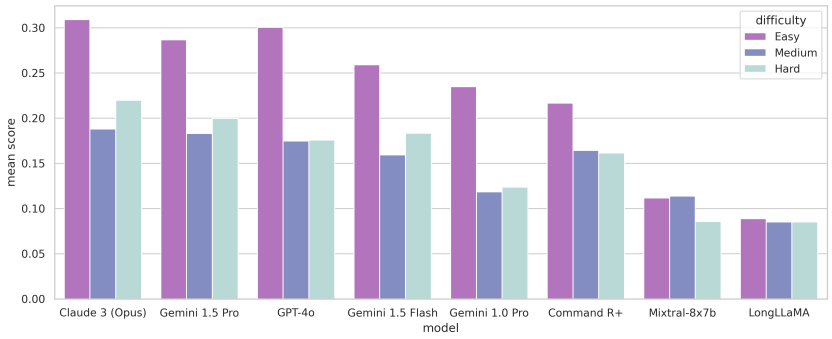


Figure 6: **Avg. performance of models sliced by difficulty of examples.** Consistent with expectations, all models perform substantially better on easy examples except in the case of Mixtral. Experts in each domain independently converged on measuring difficulty based on how spread-out the requested information was within the context of the full paper.

B Model-Based Evaluation of Mixed Long Form Responses

Tasks in CURIE are varied and have ground truth annotations in mixed and heterogenous outputs. Evaluating free-form generation is challenging because answers are often descriptive, and even when a format is specified as in most of our cases the response to each field can have differing forms e.g. in case of materials grid points may sometimes be specified as [p,q,r] and at other times as p×q×r. Hence most existing knowledge related benchmarks [1, 11] lean towards multiple choice format for answers. While these allow for clean evaluation, this doesn't allow us to evaluate the full expressiveness of the model. Inspired by the ability of LLMs to evaluate natural language [27, 28, 29], three recent approaches, LAVE [30], LIMA [31] and Prometheus-Vision [32], utilize the in-context capability of instruction-tuned LLMs to rate the candidate answers in 3-point, 6-point and 5-point Likert scales, respectively. While evaluation on Likert scales is suited for generated text, many of the outputs on our task have structured information that could benefit from more fine grained evaluation. So we propose two model-based evaluations (i) LMScore an overall weighted score on a 3-point scale obtained by asking the LLM if the predictions match ground truth, and (ii) LLMSim a more nuanced score for measuring similarity of elements in lists of dictionaries, which can then be used to compute precison and recall of elements.

(i) **LMScore** Given the ground truth and predicted responses, we ask the model to check if the predicted responses match the ground truth, and ask the model to output "good" (if the prediction has few minor errors), "okay" (if there are many minor errors), and "bad" if there are major errors. Instead of using the model generated response directly, we compute a score based on the model log-likelihood values. If x_t represents the tokens for the 3 categories we are interested in, $x_t \in \{\text{bad, ok, good}\}$, and w_t are the corresponding weights we want to assign to each category, $w_t \in \{0, 0.5, 1\}$, then

$$LMScore = \sum_{t=0}^{2} p(x_t) \times w_t \tag{1}$$

$p(x_t)$ is computed by renormalize the probabilities of the tokens by considering a $softmax()$ operation on the log-probabilities of the tokens: $([l_{bad}, l_{ok}, l_{good}])$.

(ii) **LLMSim** is used to compare similarity of dictionary elements to assist in comparison of sets of dictionaries. Our goal is to identify the number of ground truth dictionary items that have been retrieved correctly. So, we ask the LLM to examine all of the predicted dictionaries and match and identify the predicted dictionary most similar to the each of the ground truth dictionaries. We use a chain-of-thought (CoT) prompt that asks the LLM to identify the predicted dictionary indices that correctly match each field (key) of the ground truth, and then state the which predicted dictionary is most similar to the ground truth or output None. We will make the prompt and code available.

Concretely, suppose D_P is the set of predicted dictionaries, D_G the set of ground truth dictionaries, LLMSim helps find the optimal matching M between the predicted dictionaries and each ground truth $D_g \in D_G$:

$$\begin{aligned}\text{LLMSim} &= M(D_P, D_g) \\ &= \begin{cases} None, \text{if no match in values} \\ D_p \in D_P : \arg\max\ s(f_i, D_p, D_g) \end{cases}\end{aligned}$$

where f_i represents the i^{th} field (key) in the dictionary and $s(f_i, D_p, D_g)$ is the similarity of the value of each field of D_p with D_g. Given the matching, we can then compute precision, recall and F1

$$Pr = \frac{|(D_p, D_g) \in M|}{|D_P|}, Re = \frac{|(D_p, D_g) \in M|}{|D_G|}$$

B.1 Comparing human and model based evals.

Human vs Model-based eval on retrieval. On the information retrieval tasks: DFT-S and DFT-P tasks which requires LLMs to retrieve material structures and DFT parameters from a given paper; as well as the MPV tasks requiring models to retrieve material property and values, we use LLMSim to compare the dictionaries of extracted material properties. Table 2 reports precision, recall and F1 scores computed after on matching elements using LLMSim . We found the precision and recall to closely match those measured by human experts (on the Gemini 1.5 pro and GPT-4o models).

Model	DFT-S			DFT-P			MPV			MPV-non-trivial			MPV-specific		
	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1
Zero-shot Open Weight LLMs															
Mixtral	26.52	24.76	25.61	9.79	6.57	7.87	31.86	23.29	26.91	33.66	23.96	27.99	22.20	35.05	27.18
Command-R+	42.93	28.80	34.47	8.69	5.82	6.97	34.99	42.11	38.23	13.17	21.44	16.32	28.10	27.58	27.84
LongLLaMa	1.82	2.13	1.96	4.40	5.83	5.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Zero-shot Closed Weight LLMs															
Gemini 1.0 Pro	46.31	40.04	42.95	12.86	7.74	9.66	29.00	43.88	34.92	26.83	41.49	32.59	26.41	41.18	32.18
GPT-4o	37.51	29.94	33.30	29.96	21.09	24.76	39.22	24.14	29.88	48.04	24.94	32.83	34.38	23.13	27.65
Gemini 1.5 Pro	38.29	35.77	36.99	25.39	17.74	20.89	25.62	41.26	31.61	30.68	40.85	35.04	25.00	31.34	27.82
Gemini 1.5 Flash	39.95	38.75	39.34	27.45	17.49	21.36	18.22	45.33	25.99	22.70	40.07	28.98	14.77	32.90	20.39
Claude 3 (Opus)	40.45	32.89	36.28	28.22	17.78	21.81	46.71	38.22	42.04	52.83	49.83	51.29	32.18	47.06	38.23

Table 2: **Comparing performance using LLMSim.** On sub-tasks requiring exhaustive retrieval of information we use LLMSim based similarity to compute F1 scores for finer grained assessment on materials science. We also include 2 ablations for the MPV task where we ask the LLM to retrieve non-trivial or specific property values (refractive index and optical bandgap) for materials.

Human vs. Model-based 3-point evaluations We worked with experts in each domain to evaluate predictions generated by the models against ground truth responses on a 3-point scale identical to the proposed LMScore. For each example, the expert was asked to rate a response as "good" if it had few or no errors compared to the ground truth, "okay" if it had many minor errors, and "bad" if there were major errors. We use these human responses to compare and correlate the newly proposed LMScore which is reported in Fig. 7. While LMScore appears to be promising, it requires further analysis prior to wider usage.

C Data Collection and Examples

The CURIE benchmark consists of a series of tasks that measure how well LLMs can assist in diverse scientific workflows, from synthesis of information towards final execution anchored on

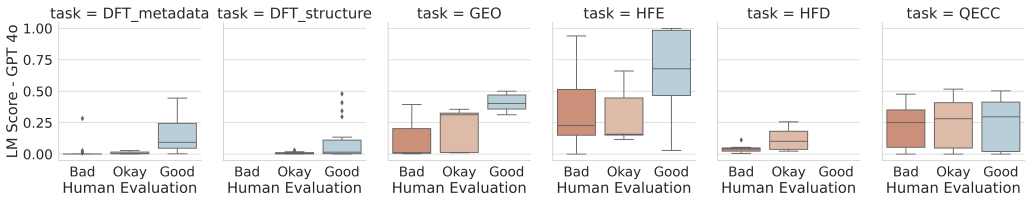


Figure 7: **Correlation of GPT-4o based LMScore metric with human evaluations,** Across tasks in domains where Rouge-L is the primary evaluation metric, LMScore appears to be a promising alternative to Rouge.

single scientific research documents. Each task in the benchmark: (1) is a realistic task performed by scientific experts on domains requiring years of study, (2) has information relevant to solve the given problem within the context provided (e.g. a full-length scientific paper, image/caption pair), and (3) ensures expert humans can evaluate task performance, providing metrics that highlight the potential limitations of current models.

Collection guidelines. We selected six domains requiring deep scientific expertise: materials science, theoretical condensed matter physics, quantum computing, geospatial analysis, biodiversity, and proteins. Within these, we worked with experts to define tasks representative of realistic scientific workflows, covering the following seven assessment categories: *concept extraction, concept tracking (co-reference resolution), aggregation, algebraic manipulation, summarization, visual comprehension, and integrating expertise across domains.* We focused on tasks that, if successful, could enable automation [33] of a time intensive critical component of a workflow e.g. extraction of experimentally reported values towards curating a database [34], or generate code or calculations to fully reproduce computational or theoretical analyses. We worked with domain experts on 3 critical aspects of the task preparation: (1) sourcing papers representative of the task and domain; (2) creating ground truth labels that were accurate, nuanced and comprehensive; and (3) creating metrics to evaluate model responses against ground truth answers that properly captured salient features of the task. Figure 1(b) shows the distribution and details of tasks in the CURIE benchmark.

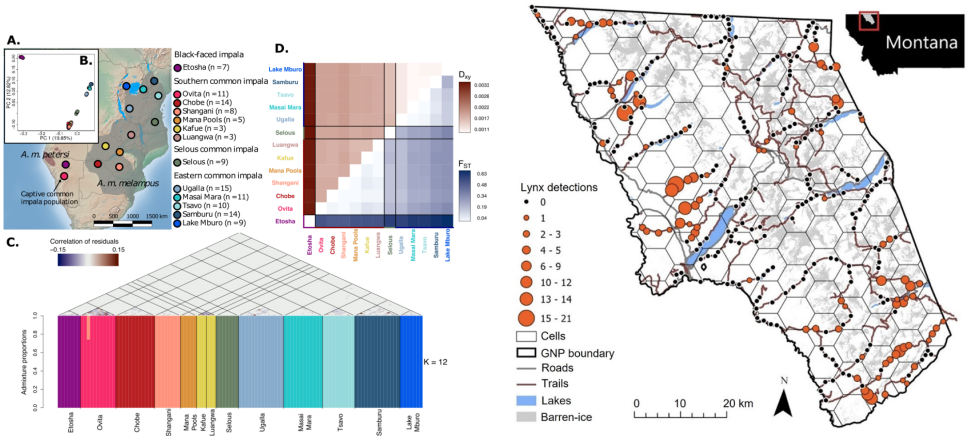


Figure 8: A sample of map images from the Biodiversity Georeferencing (BIOGR) Task.

D Related Work

Science NLP tasks. There have been numerous datasets created to perform core NLP tasks on scientific texts. This includes (i) *named entity recognition* to annotate entities such as disease names [35] or material properties [36], and relations such as disease-chemical interaction [37]; (ii) *dependency parsing* [38], (iii) participant-intervention-outcome *(PICO) annotation* [39], (iv) *text classification* such as citation intent classification [40, 41], paper domain classification [42] from titles; (v) *relation extraction* for chemical-protein-disease annotation [43] or material structure and

properties [44], (vi) *information extraction* e.g. material property values [45, 46], and (vii) *question answering* [10, 47]. However all of these focus on inputs of short length such as paper abstracts, sentences or text spans. They were curated for language models operating on short contexts, though the task in the actual scientific workflow requires application on full documents. Of the recent LLM benchmarks, GPQA [11] focuses on evaluating scientific domain expertise in biology, physics, and chemistry, while MMLU [1] covers a range of high school science. These too are limited to short questions and multiple choice answers.

Long context benchmarks. With the increase in context windows supported by the LLMs, there have been new benchmarks focusing on evaluating long context capabilties. ZeroScrolls [2] covers summarization, question answering, aggregation which are now present in many newer benchmarks: NIAH [5] includes retrieval, LongBench [9] includes bilingual tasks, Bamboo [3] has textual entailment tasks amongst others and M4LE [48] has tasks testing translation and classification, L-eval [49] includes a task on multi-document dialog, and Loogle [50] includes a computation task. However, while all of these benchmarks combine many existing datasets to cover a range of tasks none of them operate on data from the scientific domain. Further, for very long (100k+) contexts synthetically crafted data used, e.g. Ruler [51] proposes synthetically created length adaptable tasks and Ada-Leval [8] includes length adaptable sorting and QA tasks, where they add distractor texts to increase the context. These ignore, data from scientific literature that's naturally complex and requires processing long context.

Scientific literature. Of the tasks most relevant to scientific expertise, QASA [52] and QASPER [53] operate on a full scientific paper, Machine Learning (ML) and Natural Language Processing (NLP) papers respectively, however these focus solely on question answering, since it is expensive and labor-intensive to collect tasks requiring expert knowledge. In our work we introduce ten new tasks curated from six disciplines, all annotated by experts(with Ph.D. degrees) and requiring reasoning over long context information on average about 11k+ words ($\approx$15k tokens). See Fig. 1(a) for comparison.