OPENSCHOLAR: SYNTHESIZING SCIENTIFIC LITERATURE WITH RETRIEVAL-AUGMENTED LMS

Akari Asai[1,5] **Jacqueline He**[1]* **Rulin Shao**[1,5]* **Weijia Shi**[1,2]
Amanpreet Singh[2] **Joseph Chee Chang**[2] **Kyle Lo**[2] **Luca Soldaini**[2]
Sergey Feldman[2] **Mike D'arcy**[2] **David Wadden**[2] **Matt Latzke**[2]
Minyang Tian[3] **Pan Ji**[6] **Shengyan Liu**[3] **Hao Tong**[3] **Bohao Wu**[3] **Yanyu Xiong**[7]
Luke Zettlemoyer[1,5] **Graham Neubig**[4] **Dan Weld**[1,2] **Doug Downey**[2]
Wen-tau Yih[5] **Pang Wei Koh**[1,2] **Hannaneh Hajishirzi**[1,2]
[1]University of Washington [2]Allen Institute for AI [3]University of Illinois, Urbana-Champaign
[4]Carnegie Mellon University [5]Meta [6]University of North Carolina, Chapel Hill [7]Stanford University
{akari, pangwei, hannaneh}@cs.washington.edu

ABSTRACT

Scientific progress depends on researchers' ability to synthesize the growing body of literature. Can large language models (LMs) assist scientists in this task? We introduce OPENSCHOLAR, a specialized retrieval-augmented LM that answers scientific queries by identifying relevant passages from 40 million open-access papers and synthesizing citation-backed responses. To evaluate OPENSCHOLAR, we develop SCHOLARQABENCH, the first large-scale multi-domain benchmark for literature search, comprising 2,967 expert-written queries and 208 long-form answers across computer science, physics, neuroscience, and biomedicine. On SCHOLARQABENCH, OPENSCHOLAR-8B outperforms GPT-4o by 5% and PaperQA2 by 7% in correctness, despite being a smaller, open model. While GPT4o hallucinates citations 78–90% of the time, OPENSCHOLAR achieves citation accuracy on par with human experts. OPENSCHOLAR's datastore, retriever, and self-feedback inference loop also improves off-the-shelf LMs: for instance, OPENSCHOLAR-GPT4o improves GPT-4o's correctness by 12%. In human evaluations, experts preferred OPENSCHOLAR-8B and OPENSCHOLAR-GPT4o responses over expert-written ones 51% and 70% of the time, respectively, compared to GPT4o's 32%. We open-source all of our code, models, datastore, data and a public demo.

Demo	⚙	openscholar.allen.ai/
Blog	⚙	https://allenai.org/blog/openscholar
OpenScholar code	⚙	github.com/AkariAsai/OpenScholar
ScholarBench code	⚙	github.com/AkariAsai/ScholarBench
Checkpoints, Data, Index	🤗	OpenScholar/openscholar-v1
Expert Evaluation	⚙	AkariAsai/OpenScholar_ExpertEval

1 INTRODUCTION

Synthesizing knowledge from scientific literature is essential for uncovering new research directions, refining methodologies, and supporting evidence-based decisions. However, the vast volume of papers published annually makes it increasingly difficult for researchers to stay informed. Effective synthesis requires precise retrieval, accurate attribution, and real-time access to current literature. While large language models (LLMs) show promise in assisting researchers, they face significant challenges, including hallucinations (Mallen et al., 2023; Mishra et al., 2024), reliance on outdated pre-training data (Kasai et al., 2022), and a lack of transparent attribution. For instance, when tasked with citing up-to-date literature, GPT-4 fabricated citations in 78-90% of cases across fields like computer science and biomedicine in our experiments.

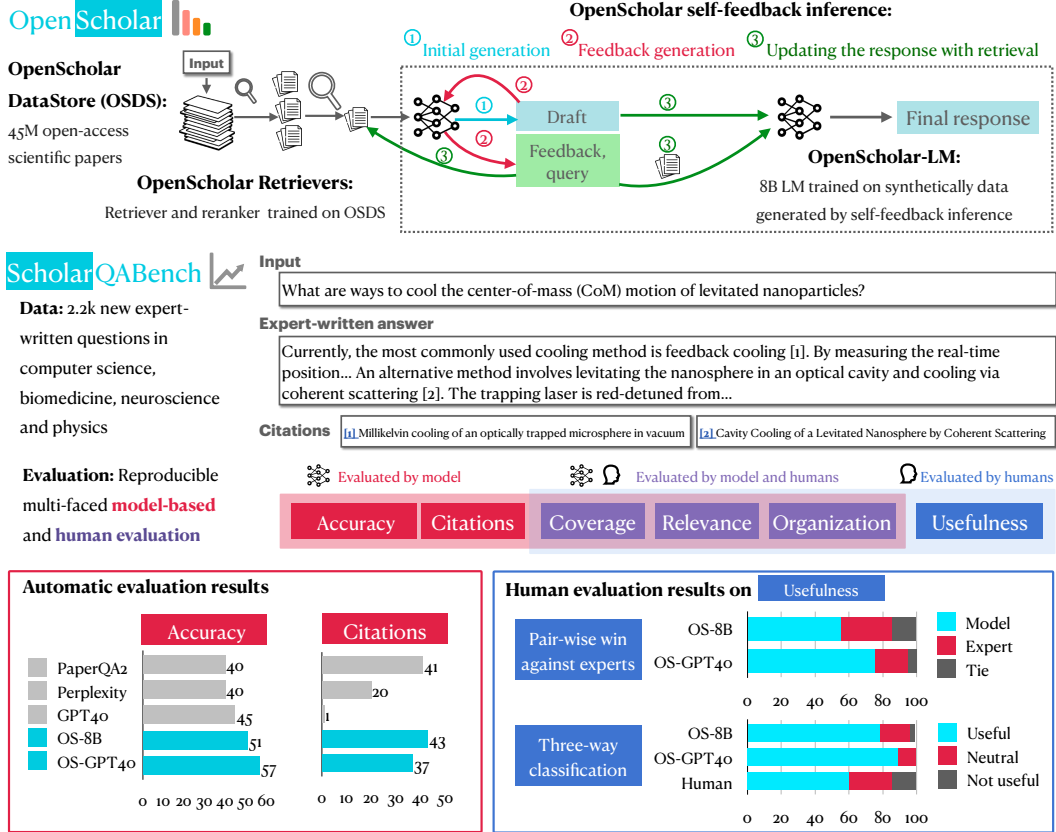*Contributed equally (alphabetical order). All authors' contributions are detailed in the Contribution section.

Figure 1: **(Top) Overview of OPENSCHOLAR:** OPENSCHOLAR consists of a specialized datastore, retrievers and LMs and iteratively improves responses using self-feedback inference with retrieval. **(Middle) Overview of SCHOLARQABENCH:** SCHOLARQABENCH consists of 2.2k expert-written questions across multiple scientific disciplines, and we introduce automatic and human evaluation protocols for SCHOLARQABENCH. **(Bottom) Automatic and Human Evaluation Results:** Experimental results show the effectiveness of SCHOLARQABENCH, and that OPEN-SCHOLAR with our trained 8B or GPT4o significantly outperforms other systems, and is preferred over experts over 50% of the time in human evaluations.

Retrieval-augmented LMs (Lewis et al., 2020; Guu et al., 2020), on the other hand, can mitigate many of these issues by integrating retrieved external knowledge sources at inference-time, driving the development of systems for literature search and synthesis (Agarwal et al., 2024; Zheng et al., 2024; Skarlinski et al., 2024). However, many such systems rely on black-box APIs or general-purpose LLMs that are neither optimized for literature synthesis nor paired with open, domain-specific retrieval datastores (i.e., a processed corpus and corresponding retrieval index) that are specifically suited for scientific domains. Moreover, evaluations for scientific literature synthesis have been limited, using single-discipline and small-scale human evaluations (Agarwal et al., 2024; Zheng et al., 2024) or simplified tasks such as multiple-choice question answering (Skarlinski et al., 2024).

To address these gaps, we present OPENSCHOLAR (Figure 1, top), a state-of-the-art retrieval-augmented LM with a specialized paper datastore and retrievers trained on scientific literature. At inference time, OPENSCHOLAR retrieves relevant passages and uses iterative self-feedback generation to refine its own output. We further train a new, efficient 8B LM. To evaluate the effectiveness of OPENSCHOLAR, we introduce SCHOLARQABENCH (Figure 1, middle), a benchmark specifically designed to enable a realistic and reproducible evaluation of open-ended scientific question answering.

OPENSCHOLAR (Section 2) uses our new OPENSCHOLAR-DATASTORE (OSDS), which contains 40 million open-access papers from Semantic Scholar, along with 237 million corresponding passage embeddings. To the best of our knowledge, this is the largest open-sourced datastore of scientific domains. OPENSCHOLAR first retrieves passages from OSDS using a retriever and reranker.

Subsequently, an LM synthesizes the retrieved passages to generate responses with citations. OPEN-SCHOLAR iteratively refines its outputs through natural language feedback, which improves quality and adaptively incorporates supplementary information. This pipeline is also used to create large-scale, high-quality training data for smaller, more efficient models. We generate synthetic queries and instructions from sampled datastore passages, feed them into OPENSCHOLAR, and use intermediate and final output to train open 8B model, OPENSCHOLAR-8B and retrieval models.

SCHOLARQABENCH (Section 3) is a benchmark designed to evaluate the ability of models to understand and synthesize existing research. Unlike previous benchmarks that assume that answers can be found in a single paper (e.g., scientific fact-checking; Wadden et al., 2020; Skarlinski et al., 2024), many real-world scenarios require identifying multiple relevant papers and generating long-form output with accurate citations. To address these challenges, we curated a dataset of 2,967 literature synthesis questions, along with 208 long-form responses that are written by experts and span four scientific disciplines, namely computer science, physics, biomedicine, and neuroscience. These responses were crafted by Ph.D. students and postdoctoral researchers with more than three years of experience and relevant publications in the field. On average, each response required approximately one hour to compose. We also introduce a multifaceted evaluation protocol that combines automated metrics and human assessments to measure citation accuracy, factual correctness, content coverage, coherence, and overall quality. This multifaceted approach ensures robust and reproducible evaluations, both automatic and human-driven.

We evaluated proprietary and open models (e.g., GPT4o, Llama 3.1 8B, 70B) with and without retrieval capabilities, as well as specialized systems like PaperQA2 (Skarlinski et al., 2024), on SCHOLARQABENCH (Section 4). While GPT4o demonstrated strong general performance, it struggled with citation accuracy and coverage, often producing inaccurate or non-existent citations. OPEN-SCHOLAR outperformed both LM-only and retrieval-augmented pipelines, surpassing proprietary and open-source systems. Notably, using fully open-source checkpoints, OPENSCHOLAR outperformed PaperQA2 (Skarlinski et al., 2024), built on proprietary LMs, and production systems like Perplexity Pro, achieving 6% and 10% improvements, respectively. Additionally, OPENSCHOLAR's use of smaller, efficient retrievers significantly reduced costs. Combining OPENSCHOLAR with GPT4o also improved correctness by 12% over GPT4o alone. The OPENSCHOLAR pipeline can also enhance off-the-shelf LMs. For example, when using GPT-4o as the underlying model, OPENSCHOLAR-GPT4o achieves a 12% improvement in correctness compared to GPT-4o alone.

In addition to automatic evaluations on SCHOLARQABENCH, we conducted detailed expert assessments with 16 scientists from fields such as computer science, physics, and biomedicine (Section 5). These experts performed pairwise and fine-grained evaluations of OPENSCHOLAR's outputs against 108 expert-written responses to literature synthesis queries in SCHOLARQABENCH. OPENSCHOLAR, when paired with GPT-4o and our trained 8B model, consistently outperformed expert-written responses, with win rates of 70% and 51%, respectively. In contrast, GPT-4o without retrieval struggled with information coverage and was rated as less helpful than human experts, achieving only a 31% win rate against human responses. This highlights that OPENSCHOLAR-generated outputs are more comprehensive, well-organized, and useful for synthesizing literature. These findings demonstrate that OPENSCHOLAR produces high-quality outputs that are not only competitive with expert-written answers but, in some cases, exceed them, particularly in terms of coverage and organization.

OPENSCHOLAR-8B is an open retrieval-augmented LM that avoids reliance on proprietary LMs or retrieval systems, leveraging one of the largest datastores in scientific literature domains. We release the full OPENSCHOLAR ecosystem, including code, trained retrievers, the LM checkpoint, the datastore, the SCHOLARQABENCH benchmark, expert evaluation tools, and a public demo.

2 OPENSCHOLAR: OPEN RETRIEVAL-AUGMENTED LM TO SYNTHESIZING SCIENTIFIC LITERATURE

OPENSCHOLAR (detailed in Figure 2) is a new retrieval-augmented LM designed to ensure reliable, high-quality responses to a range of information-seeking queries about scientific literature.

Task formulation. Given a scientific query x, the task is to identify relevant papers, synthesize their findings, and generate a response y that effectively addresses the query. This response should be accompanied by a set of citations, $\mathbf{C} = c_1, c_2, \ldots, c_K$, wherein each citation c_i corresponds to an
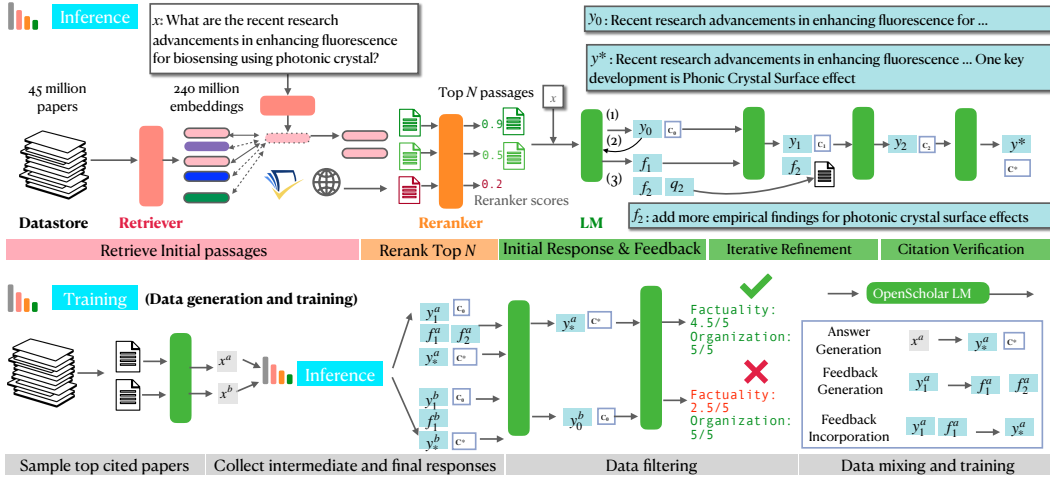
Figure 2: **Detailed overview of OPENSCHOLAR inference (top) and training (bottom).** At inference time, given an input x, OPENSCHOLAR first uses a retriever to identify relevant papers from a specialized datastore (OPENSCHOLAR-Datastore), and then uses a reranker to refine and identify the top N retrieved documents. The retrieved output is then passed to the LM, which generates both an (1) initial response y_0 and (2) self-feedback f_1. By incorporating its own feedback, the LM iteratively refines its output a pre-defined number of times. Subsequently, an LM (1) generates initial response y_0, (2) generates self-feedback on the initial output, and (3) incorporate feedback (f_i) to generates an updated response y_1. The LM repeats the process until all feedback is incorporated. To train a smaller yet competitive 8B LM, we generate high-quality training data using this inference-time pipeline followed by data filtering and mixing.

existing scientific paper. Each c_i in **C** corresponds to specific passages from scientific literature, and should be provided as an in-line citation, linked to the relevant spans of text in y, following standard practice in scientific writing. These citations allow researchers to trace the output back to the original literature, ensuring transparency and verifiability.

Overview of OPENSCHOLAR. To ensure the retrieval of relevant papers and generate high-quality outputs, SCHOLARQABENCH consists of three key components: a datastore **D**, a retriever $\mathcal{R}$, and a generator LM $\mathcal{G}$. In standard retrieval-augmented inference pipelines, the process begins with $\mathcal{R}$, which retrieves a set of passages $\mathbf{P} = \{p_1, p_2, \ldots, p_N\}$ from **D**—a large-scale corpus of previously published scientific papers—based on semantic relevance to the input query x. These passages serve as context for the next step. The generator LM $\mathcal{G}$ then takes both the retrieved passages **P** and the input query x to produce the output y along with corresponding citations **C**. Formally, this process can be represented as:

$$y, \mathbf{C} = \mathcal{G}(x, \mathcal{R}(x, \mathbf{D})),$$

where each c_i in **C** corresponds to a specific passage from **P**. In OPENSCHOLAR (Figure 1), we leverage a suite of specialized components designed for scientific domains: the OPENSCHOLAR-DATASTORE **D**, a OPENSCHOLAR-RETRIEVER/-RERANKER, and an LM, enabling flexible use of either off-the-shelf LMs (e.g., GPT4o) or our newly trained OPENSCHOLAR-LM. We develop self-feedback retrieval-augmented inference to improve reliability and citation accuracy.

OPENSCHOLAR-DATASTORE (OSDS) is a database of 45 million scientific papers, for which we build embeddings. We train OPENSCHOLAR-RETRIEVER and OPENSCHOLAR-RERANKER on scientific data, which passes the top N passages to the generator $\mathcal{G}$ (Section 2.1). Subsequently, we use iterative self-feedback inference with retrieval: the LM first generates an initial draft y_0 with $\mathcal{G}$, then iteratively enhances its output through retrieval-augmented self-feedback (Section 2.2). We use this pipeline to generate high-quality training data (Section 2.3), enabling the training of specialized LMs that produce higher-quality output and more accurate citations.

2.1 OPENSCHOLAR RETRIEVAL PIPELINE

Figure 2 (top left) shows our retrieval pipeline, consisting of a datastore $\mathbf{D}$, a bi-encoder retriever θ_{bi}, and a cross-encoder reranker θ_{cross}. We first select initial candidate paragraphs using $\mathbf{D}$ and θ_{bi}, as well as external APIs, and then refine and identify the top N relevant paragraphs using θ_{cross}.

Collect scientific papers to construct datastore. While prior work often uses a small subset of papers, such as arXiv papers from 2023-2024 (Zheng et al., 2024), it is important to have a diverse set of papers to improve the quality and coverage of model generation (Shao et al., 2024). To this end, we use peS2o (Soldaini et al., 2024) as our retrieval source, which consists of open-access academic papers from S2ORC (Lo et al., 2020). We built our datastore using peS2o v2, which includes 40 million papers up until January 2023.[1] Following prior work (Shao et al., 2024), we split the main text of each paper into discrete, 250-word text blocks (as determined by white space) and concatenate the paper title to each block to formulate passages in $\mathbf{D}$. Our datastore consists of 182 million passages.[2] To our knowledge, this is the largest open-sourced retrieval datastore for scientific literature.

Retrieve initial paragraphs. We retrieve passages from three sources: (1) the peS2o datastore using our trained retriever, (2) publicly available abstract from papers returned via the Semantic Scholar API (Kinney et al., 2023) based on search keywords, and (3) publicly available texts from papers retrieved through a web search engine using the original query x.[3] For (1), we first generate embeddings of each passage in $\mathbf{D}$ using the passage bi-encoder θ_{bi}, which processes text chunks (e.g., queries or passages) into dense vectors (Karpukhin et al., 2020) offline. Off-the-shelf retrieval models often struggle in out-of-domain scenarios (Thakur et al., 2021). To overcome this limitations, we develop θ_{bi} by continually pre-training Contriever (Izacard et al., 2022) on the peS2o datastore in an unsupervised fashion to improve domain-specific retrieval performance (see Appendix C.1 for details). During inference, we encode the query using θ_{bi} and retrieve the top 100 passages through a nearest neighbor search (Karpukhin et al., 2020). For (2), we first generate keywords from the query x using a generator LM. These keywords are then used to retrieve the top 10 papers for each, as ranked by citation count, via the Semantic Scholar Search API. This approach addresses a limitation of the Semantic Scholar API, which cannot effectively handle long, question-like search queries. For (3), we obtain the top 10 search results using the You.com retrieval API,[4] restricting the search to academic platforms such as ArXiv and PubMed. If the papers are open-access, we extract and add their full texts to the candidate pool; otherwise, we include only their abstracts.

Rerank and finalize top N paragraphs. After the initial stage, we have gathered over 100, or even a thousand of relevant passages per query. However, passages retrieved by the bi-encoder may include unhelpful context due to deep interactions between a query and passages, as they are encoded separately (Asai et al., 2023). Feeding a large number of documents that might including irrelevant content to LLMs can cause efficiency and performance issues, even with state-of-the-art models (Liu et al., 2023a; Xu et al., 2023a). To overcome these challenges, we use a cross-encoder reranker (Nogueira & Cho, 2019; Xiao et al., 2023), denoted as θ_{cross}. For each candidate paragraph, the cross-encoder reranker jointly encodes and computes the relevance score between the input query and each of the passages. We then use the relevance score to rank the passages accordingly. To train θ_{cross} for scientific domains, we fine-tune a BGE-reranker (Xiao et al., 2023) using synthetic data generated by Llama 3 70B Instruct. Specifically, we randomly generate queries based on abstracts from peS2o and retrieve the top 10 passages. Llama 3 70B Instruct then assigns relevance scores from 1 to 5 for these passages, where we consider scores of 4 or 5 as positive, and scores of 1 or 2 as negative. Passages with a score of 3 are discarded. More details of θ_{cross} training are in Appendix C.2. During reranking and finalization of top N passages, we also implement additional meta-filtering,

---

[1] https://huggingface.co/datasets/allenai/peS2o#pes2o-v2-latest. We further reran the peS2o pipeline and expanded it with the newly collected peS2o v3, which consists of 45 million papers up until October 2024. By default, we use peS2o v2 for evaluation.

[2] peS2o includes 8 million full-text papers and 30 million abstracts, with an average of 19 passages per full-text paper.

[3] We enhance the retrieval candidates by incorporating results from additional sources, such as the Semantic Scholar Search API and a commercial web search API, which can return papers relevant to short search queries. Since peS2o v2 lacks papers published after its knowledge cutoff (January 2023), these APIs help keep OPENSCHOLAR up-to-date.

[4] https://api.you.com/

which includes: (1) limiting the number of passages per paper to three passages, and (2) incorporating normalized citation counts into relevance scores predicted by the cross-encoder.

2.2 INFERENCE: ITERATIVE GENERATION WITH RETRIEVAL-AUGMENTED SELF-FEEDBACK

In standard retrieval-augmented generation (RAG; Lewis et al. 2020; Ram et al. 2023), a generator LM takes in the original input x and top N retrieved passages $\mathbf{P}$ and generates the output y_0. Although effective for tasks such as question answering (Mallen et al., 2023), this one-step generation can lead to unsupported claims (Liu et al., 2023b) or incomplete output due to missing information (Asai et al., 2024; Jiang et al., 2023). To address these challenges, in OPENSCHOLAR, we introduce an iterative generation approach with self-feedback, which involves three steps: (1) **initial response and feedback generation** to output the initial draft y_0 and a set of feedback on y_0; (2) **iterative refinement with additional retrieval** to improve y_0 using the feedback, and (3) **citation verification**. Full details are in the Appendix.

Initial response and feedback generation. Given the input x and retrieved passages $\mathbf{P}$, the generator LM first produces an initial response y_0 with citation markers tied to the corresponding passages in $\mathbf{P}$. After generating y_0, the LM generates a set of feedback on y_0, $\mathbf{F} = f_1, f_2, \ldots, f_T$, that is aimed at improving the initial response, wherein each feedback f_t is a natural language sentence that describes potential improvements. Although the model can generate an arbitrary number of feedback (T), we set a maximum limit of three feedback sentences for efficient inference. Unlike prior work that relies on a predefined set of feedback signals (Asai et al., 2024), our approach allows the LM to generate flexible natural language feedback on various aspects of the response, such as organization, completeness, or additional needed information. If the feedback sequence identifies missing content (e.g., "The answer only includes empirical results on QA tasks. Add results from other task types."), the LM also generates a retrieval query for additional retrieval using the pipeline in Section 2.1.

Iterative refinement. We then iterate over the feedback $\mathbf{F}$ to incrementally refine the output. If f_k indicates that further retrieval is needed, the query q_k is used to retrieve additional passages, which are appended to $\mathbf{P}$ before producing y_k.[5] The LM uses the previous output y_{k-1}, the retrieved passages $\mathbf{P}$, and newly retrieved passages if any, to generate an updated output y_k. This process is repeated until all feedback has been addressed, resulting in a final output y_T by timestep T.

Citation verification. Finally, we instruct the generator LM to verify the citations in y_t. Specifically, the generator ensures that all citation-worthy statements—scientific claims requiring justification—are adequately supported by references from the retrieved passages. If any claims lack proper citations, the LM performs a post hoc insertion to ensure that citation-worthy statements are supported by passages. In our pipeline, we do not remove sentences that lack citation-worthy statements.

2.3 TRAINING: HIGH-QUALITY SYNTHETIC DATA GENERATION WITH INFERENCE PIPELINE

Building powerful LMs that can effectively synthesize scientific literature is challenging due to the lack of training data for this problem. While there are some resources to train scientific LMs (Wadden et al., 2024), most tasks do not require open-retrieval settings and are single-paper tasks. As a result, most prior work in this area (Skarlinski et al., 2024) rely on proprietary LMs, which poses challenges for reproducibility and inference costs.

We leverage our inference-time pipeline to synthetically generate high-quality training data through self-feedback, so that the resulting model can get better at generating higher-quality output without going through the self-feedback process (Figure 2 bottom).

Question and response generations. Our data generation process involves three steps: first, selecting the top-cited papers from $\mathbf{D}$; second, generating information-seeking queries based on their abstracts; and third, using the OPENSCHOLAR inference-time pipeline to produce high-quality responses. We generate data using LLama 3.1 70B (Dubey et al., 2024). Specifically, we begin by sampling 1 million paper abstracts from the peS2o dataset and the retrieve papers' meta information such as publication year or citations. We then randomly select 10,000 papers that published later than 2017, and then

---

[5]Although we could iteratively regenerate the output each time feedback is provided, doing so introduces additional latency. Empirically, we found that feedback is often diverse, addressing different aspects of generation. As a result, sequentially incorporating feedback from the initial output remains effective.

prompt an LM to generate literature review questions or information-seeking queries based on each abstract, that may require multiple papers to answer. Next, we employ our OPENSCHOLAR pipeline to produce the final output y_T, along with intermediate generations such as feedback $\mathbf{F}$ and initial outputs.

Data filtering. Despite its effectiveness and scalability, synthetic data may also contain issues such as hallucinations, repetitive writing, or limited instruction-following (Li et al., 2024c). To address this, we introduce a two-step data filtering process: pairwise-filtering and rubric-filtering, leveraging the same LM used for data generation. In pair-wise filtering, we compare the quality of model outputs y_T (output at the final step) and y_0 (initial output), and retain the output that is judged to be higher quality. We find that y_0 is preferred over y_T around 20% of the time, due to over-editing or increased redundancy after multiple iteration steps. We then evaluate the quality of the chosen response on a five-point scale across two aspects: **organization** and **factual precision and citation accuracy**. A valid model output must achieve a score of 4.5 or higher in both categories, and we discard instances whose outputs do not meet this requirement. More details are provided in the Appendix.

Data mixing and training. From this synthetic pipeline, we generate three types of training data: answer generation ($x \rightarrow y$), feedback generation ($y_0 \rightarrow \mathbf{F}$), and feedback incorporation ($y_{t-1}, f_t \rightarrow y_t$). We found that incorporating both final and intermediate outputs during training helps smaller LMs learn to generate more effective feedback. We further blend this synthetic training data with existing general-domain instruction-tuning data (Ivison et al., 2023) and scientific instruction-tuning data (Wadden et al., 2024), ensuring that 50% of the training data comes from scientific domains, while the remaining 50% is sourced from general-domain data. We also generate synthetic fact verification and boolean QA data based on sampled abstract data from peS2o. For this, we sort the papers based on citation count and select the top 100,000 papers. More details of data mixing and training are available at Appendix C.3. After data mixing, we train generator LMs on our large-scale synthetic training data. We train Llama 3.1 8B Instruct on the generated training data.

3 SCHOLARQABENCH: REALISTIC LITERATURE REVIEW EVALUATION BENCHMARK ANNOTATED BY PH.D. EXPERTS

Challenges and overview. Prior studies on building LLMs to synthesize scientific literature employ either small-scale, single-domain human evaluation (Agarwal et al., 2024; Zheng et al., 2024) or over-simplified multiple-choice QA setups (Skarlinski et al., 2024). Building high-quality benchmarks for literature review has two major challenges. First, creating such datasets is resource-intensive, as it requires Ph.D.-level domain expertise and research experience, particularly when annotating realistic questions and high-quality answers. Second, even when high-quality data is available, reliably evaluating long-form natural language responses presents a significant challenge, especially in expert domains (Xu et al., 2023b; 2024). This contrasts with benchmarks for other scientific processes, such as automated experimental code generation, for which clearer evaluation criteria, such as Pass@1, are more readily available (Si et al., 2024).

To address these gaps, we introduce SCHOLARQABENCH, a benchmark that supports diverse formats of scientific literature synthesis tasks, including closed-form classification, multiple-choice, and long-form generation, as shown in Table 1. We adopt three existing single-paper datasets, and then construct a suite of high-quality, expert annotated datasets for computer science, biomedicine, physics, and neuroscience (Section 3.1). We also build a reliable automatic evaluation pipeline (Section 3.2). Table 1 provides a list of tasks in SCHOLARQABENCH, and Figure 3 shows an example and an overview of the evaluation pipeline.

3.1 DATA CURATION

SCHOLARQABENCH is designed to evaluate model capabilities in automating scientific literature review. The curation process is guided by three key factors: **Diversity of tasks**: SCHOLARQABENCH includes tasks with a range of input-output formats; **Diversity of disciplines**: Unlike previous analyses that often focus on a single discipline such as computer science, SCHOLARQABENCH spans four scientific disciplines; **Inclusion of multi-paper tasks**: Unlike prior work that focuses on understanding single, pre-selected papers, all tasks require retrieving from the entire open-access

Dataset	Task Format	Discipline	Size	Evaluation	Multi-paper
SciFact (Wadden et al. 2020)	Claim → Label (True or False)	Biomedicine	208	`Corr`, `Cite`	
PubMed QA (Jin et al. 2019)	Question → Answer (Yes, No)	Biomedicine	843	`Corr`, `Cite`	
QASA (Lee et al. 2023)	Question → Answer (Long-form)	Computer Science	1,375	`Corr`, `Cite`	
SCHOLAR-CS	Question → Answer[†] (Long-form)	Computer Science	100	`Corr`, `Cite`	✓
SCHOLAR-BIO	Question → Answer[*] (Long-form)	Biomedicine	1,451	`Cite`	✓
SCHOLAR-NEURO	Question → Answer[*] (Long-form)	Neuroscience	1,308	`Cite`	✓
SCHOLAR-MULTI	Question → Answer (Long-form)	Computer Science, Physics, Biomedicine	108	`Cite`, `LLM`, `Expert`	✓

Table 1: **Overview of SCHOLARQABENCH**. The top three rows show single-paper datasets adopted from prior datasets. The bottom four rows are new datasets, which we constructed by recruiting Ph.D.-level experts. Answer[*] indicates that the dataset comes with questions only, and answer[†] indicates that answer will be evaluated based on human-annotated rubrics. The evaluation columns corresponds to the multi-faceted evaluations in Section 3.2. The "Multi-paper" column indicates whether the task requires multiple papers to answer. `LLM` and `Expert` indicate the fine-grained evaluations by evaluator LLMs (i.e., Prometheus; Kim et al. 2024a) and expert humans, respectively.

collection of full text of papers (Section 3.1.1) and four datasets specifically require reasoning over multiple retrieved papers (Section 3.1.2).

3.1.1 SINGLE-PAPER TASKS

For single-paper tasks, we curate and adapt existing widely-used single-paper datasets. Figure16 shows examples of single-paper tasks; more details are in Appendix B.2.

SciFact. SciFact (Wadden et al., 2020) is a dataset of 1.4K expert-written scientific claims in the biomedical domain, paired with gold evidence from existing PubMed paper abstracts annotated with labels and rationales. We include validation set queries labeled as either `supports` (true) or `contradicts` (false), discarding the original gold evidence, and reformulate the task as binary open-retrieval, wherein a system needs to identify relevant papers from a large collection of papers.

PubMedQA. PubMedQA (Jin et al., 2019) has expert-annotated (yes/no/maybe) QA data on PubMed paper abstracts. Similarly to SciFact, we only keep instances with yes or no labels, and discard the original abstract passage to formulate the task as an open-retrieval setup.

QASA. QASA (Lee et al., 2023) is a single paper QA dataset that consists of question answering pairs, requiring reasoning over scientific articles in AI and ML. We evaluate the model's ability to sufficiently answer a detailed question about the target paper. While the original dataset provides three subtasks (answer selection, rationale generation and answer compositions) as well as end-to-end QA, we evaluate the models' performance based on an end-to-end QA setup.

3.1.2 MULTI-PAPER TASKS

Single-paper, closed-set tasks may provide reliable evaluations. However, they may not be reflective of realistic scenarios, in which complex, open-ended questions are asked independently from existing papers, and require multi-paper retrieval and reasoning. Few datasets (Xu et al., 2024; Malaviya et al., 2023) explore multi-paper setups with realistic queries, and most lack a reliable evaluation pipeline or human-written references. We address this gap by curating three new long-form QA datasets, annotated by experts, for these challenging settings (details in Appendix B.2). Furthermore, our multi-paper tasks include four scientific disciplines.
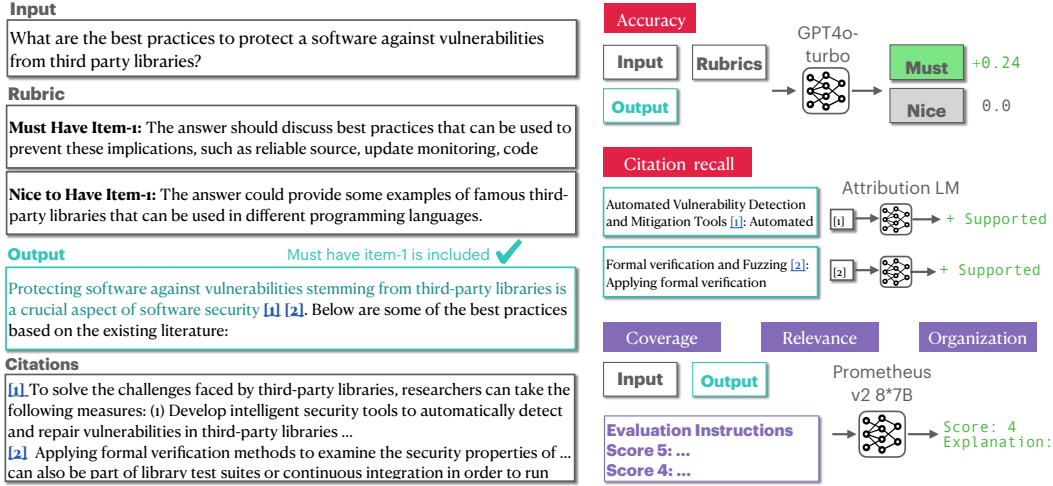
Input

What are the best practices to protect a software against vulnerabilities from third party libraries?

Rubric

Must Have Item-1: The answer should discuss best practices that can be used to prevent these implications, such as reliable source, update monitoring, code

Nice to Have Item-1: The answer could provide some examples of famous third-party libraries that can be used in different programming languages.

Output   Must have item-1 is included ✔

Protecting software against vulnerabilities stemming from third-party libraries is a crucial aspect of software security [1] [2]. Below are some of the best practices based on the existing literature:

Citations

[1] To solve the challenges faced by third-party libraries, researchers can take the following measures: (1) Develop intelligent security tools to automatically detect and repair vulnerabilities in third-party libraries ...

[2] Applying formal verification methods to examine the security properties of ... can also be part of library test suites or continuous integration in order to run

Figure 3: **An example from SCHOLAR-CS and an overview of evaluation metrics.** SCHOLAR-CS consists of 100 questions and an average of 4.4 expert-written rubrics to be satisfied. Our SCHOLARQABENCH evaluation pipeline evaluates aspects like correctness and citation accuracy.

SCHOLAR-CS. We collected 100 questions along with detailed answer rubrics for each question across various computer science disciplines by recruiting expert annotators holding Ph.D.s in the field (professors, postdoctoral researchers, and research scientists). Annotators were tasked with writing literature review questions that were expected to require multiple research papers to answer. The question topics span areas such as networks, algorithms, the Internet of Things, artificial intelligence, and human-computer interaction. Then, for each question, two other annotators searched the Web to produce a rubric listing the key ingredients for a correct answer, categorized by importance ("must-have" and "nice-to-have"), along with supporting quotes from sources for each ingredient. The annotators were instructed not to use any LLM services for this initial part of the task. After the initial web search, the annotators were shown corresponding responses from four LLM services (Claude 3.5 Sonnet, GPT-4o, Perplexity Pro and an unpublished RAG prototype based on Claude 3.5) in a randomized order in case they wanted to revise their rubrics. On average, each question is annotated with 4.4 key ingredients, each supported by 4.4 quotes.

To measure agreement, we had both annotators produce rubrics for a subset of 10 randomly sampled questions. We then compute the scores for responses from the four LLM services the annotators were exposed to using our automated approach, once for each set of annotator rubrics. Finally, we compute Pearson's correlation coefficient among the scores for each question and take the mean. Since the rubric annotation task is subjective, we compute this agreement both with and without the general criterion as part of the scores, which comes out to be 79.3 and 59.5, respectively. Figure 3 shows one example, and more examples and details are available in Appendix F.2.

SCHOLAR-BIO, SCHOLAR-NEURO. We further collected 2,759 expert-written literature review questions in biomedicine and neuroscience, recruiting six experts who have a Ph.D. in relevant areas and are currently research scientists and engineers. The annotators were asked to choose papers from their area of expertise, and generate complex scientific questions that biomedical scientists might reasonably ask about the scientific literature based upon their parsing of those papers. We collected questions from different areas, such as bioimaging, genetics, microbiology, and neuromodulation for each. Due to the cost of annotation, we focused solely on curating the questions. Full instructions and examples are available in Appendix 6 and F.3.

SCHOLAR-MULTI. Lastly, we collected 108 literature review questions and expert-written answers with citations in four domains: computer science (AI/ML, HCI), Biomedicine (Bioimaging, Genetics), Physics (Astrophysics, Photonics, Bio Physics). All annotations are conducted by Ph.D. students or post-Ph.D. scientists, who have more than three years of research experience in the corresponding areas and have multiple first-author publications. We asked them to come up with questions that are related to most recent literature, and to compose answers to the questions using relevant papers that they found via search. Our annotators were instructed not to use any LLM-based systems such as

ChatGPT, and told to only use general search (e.g., Google Search) or paper search systems (e.g., Semantic Scholar). Statistics of the collected questions and answers are available in Table 14 and the distribution of subjects are in Figure 6a, along with the average annotation time per subject. We show several examples in Appendix F.4. On average, each annotator spent 56 minutes per instance.

3.2 METRICS AND EVALUATION PROTOCOLS

We developed a multifaceted automatic evaluation pipeline to facilitate reproducible and efficient evaluations, complementing expert assessments. An overview of our evaluations is in Figure 3.

Correctness (Corr). Correctness evaluates the degree of overlap or matching of a model-generated answer and human-annotated reference. This metric is only applied to tasks with human-annotated reference answers. For short-form generation tasks given a fixed set of answer classes, namely SciFact and PubMedQA, we use accuracy as the correctness metric. For QASA, we use ROUGE-L as an evaluation metric, following Lee et al. (2023). For SCHOLAR-CS, we develop a new long-form evaluation pipeline, which employs expert-annotated rubrics. Each rubric has two criteria: general (accounting for 40% of the score) and annotation-driven (60%). General criteria cover the evaluation of length, expertise, citations, and excerpts, while annotation-driven criteria involve assessing the presence of specific key ingredients identified by annotators. GPT4o-turbo assigns scores for each criterion, and a weighted sum is computed to obtain a final score. More details are in appendix B.3.1.

Citation accuracy (Cite). Evaluating long-form responses to literature review questions requires citation accuracy: LMs should correctly attribute relevant evidence for all citation-worthy statements. In SCHOLARQABENCH, all systems generate outputs with reference numbers (e.g., [1], [2]) linked to passages provided during inference. Following prior work (Gao et al., 2023; Liu et al., 2023b), we measure citation precision and recall. Specifically, we check if each citation-worthy statement has appropriate citations and if the citations support the statement (**Citation Recall**, Cite-r). For each citation, we then verify its relevance and necessity—specifically, whether the citation supports the statement and if its removal impacts the integrity of remaining citations (**Citation Precision**, Cite-p). Finally, we compute **Citation F1** (Cite-F1) as well, and use it as a primarily metric for citation accuracy. Citation accuracy does not require gold reference answers or rubrics, so we apply this evaluation across all tasks.

Content quality and organization (LLM, Expert). We further define key aspects to evaluate the generated answers beyond Corr or Cite alone. Specifically, we assess **Relevance** (Rel) to the question, **Coverage** (Cov) of topics (e.g., diversity of discussed papers) and depth (e.g., sufficiency of details), and **Organization and Writing Flow** (Org). These aspects are challenging to capture with standard metrics. Since LMs can effectively follow detailed evaluation rubrics (Zheng et al., 2023a; Kim et al., 2024a), we use Prometheus v2 (Kim et al., 2024a) to assign five-scale scores based on defined rubrics and use the same schema for human evaluations. For human evaluation, we further evaluate **Overall Usefulness** (Use). Full instructions for this evaluation are provided in the Appendix B.3. As prior studies show that LLM is less reliable when gold reference answers are not available (Kim et al., 2024b), this evaluation is only applied to a task with human-annotated reference answer, namely SCHOLAR-MULTI. We analyzed the agreement between human and model assessments on fine-grained aspects (Appendix E.3) and found that the model's evaluations often align with human rankings, showing higher correlation especially in organization and coverage.

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENTAL DETAILS

Models.

First, we evaluate both open-weight and proprietary LMs, including Llama 3.1 (8B, 70B) and GPT-4o (gpt-4o-2024-05-13). In this setup, each LM generates an answer independently, without external retrieval, and provides a list of referenced paper titles. For evaluation, we verify whether the generated paper titles exist. If they do, we retrieve their corresponding abstracts to use as citations. For multi-paper tasks, we further evaluate other proprietary systems: Perplexity

Pro,[6] and PaperQA2 (Skarlinski et al., 2024), a concurrent literature review agent system that uses GPT4o for reranking, summarization, and answer generation.[7] Then, we evaluate models using our OPENSCHOLAR-DATASTORE (+OSDS), where we retrieve the top N passages, and concatenate and feed them together with the original input. Lastly, we evaluate our proposed OPENSCHOLAR, leveraging our custom inference-time pipeline using trained 8B model models (**OS-8B**), as well as Llama 3.1 70B and GPT4o (**OS-70B**, **OS-GPT4o**).

Details of OPENSCHOLAR. We use peS2o v2 as our default datastore **D**. We analyze the effect of different datastores in Appendix E.2. For θ_{bi} and θ_{cross} in OPENSCHOLAR, we use our trained bi-encoder and cross-encoder models, which consist of 110 million and 340 million parameters, respectively. We set the maximum number of papers from web search and Semantic Scholar to 10. For the generator LMs, we set the temperature to 0.7 and limit the maximum token count to 3,000 for response generation and 1,000 for feedback generation, and use the `vllm` package for faster inference. We train Llama 3.1 8B for two epochs on 130k training instances for two epochs, using `torchtune`. Further details are in Appendix C. For all models, we set the number of passages input into the generator LM to five for single-paper tasks and ten for multi-paper tasks. No few-shot demonstrations are provided, except for SciFact and PubMed, where we include one-shot demonstrations.

4.2 RESULTS

Table 2 show scores for multiple aspects of the main baselines. In summary, OPENSCHOLAR achieves state-of-the performance, significantly outperforming GPT4o and their standard RAG version, as well as specialized literature review systems such as PaperQA2 (Skarlinski et al., 2024) by a large margin.

Single-paper tasks. On single-paper tasks, OPENSCHOLAR consistently outperforms other models. OS-8B and OS-70B outperforms Llama 3.1 8B and 70B with and without retrieval augmentation in terms of final `Corr` and `Cite` in Table 2. OS-70B even matches or outperforms GPT4o on PubMedQA and QASA.

Multi-paper tasks. OPENSCHOLAR-8B, 70B, and GPT4o (OS-8B, OS-70B and OS-GPT4o) demonstrate strong performance in multi-paper tasks. Specifically, OS-GPT4o provides a 12.7 point improvement over GPT4o alone in SCHOLAR-CS `Corr` and a 5.3 improvement over standard RAG. When combined with trained OS-8B, OPENSCHOLAR significantly outperforms the pipeline that uses off-the-shelf Llama 3.1 8B, showcasing the benefits of domain-specific training. Furthermore, this OPENSCHOLAR-8B outperforms proprietary systems such as GPT4o, Perplexity Pro, or PaperQA2, which uses GPT4o models for passage reranking, summarization and answer generation, by a substantial margin. Notably, by leveraging efficient retrieval pipelines with lightweight bi-encoders, cross-encoders, and in-house models, OPENSCHOLAR-8B and OPENSCHOLAR-GPT4o achieve significantly lower costs—orders of magnitude cheaper than PaperQA2—while maintaining high performance.

Limitations of parametric LMs. On both single-paper and multi-paper tasks, we observe that non-retrieval augmented baselines struggle and retrieval is almost always conducive to achieving better performance, and models without any retrieval often struggle to generate correct citations and show limited coverage on multi-paper tasks. As shown in Table 3, the proportion of cited papers that actually exist is strikingly low. In particular, while models such as GPT4o and Llama can generate plausible reference lists, we find that 78-98% of the cited papers are fabricated, and the issue is exacerbated in biomedical domains. Even when citations refer to real papers, the majority are not substantiated by the corresponding abstracts, resulting in near-zero citation accuracy.

We also observe that such models also generate responses with limited coverage. On SCHOLAR-Multi, non-retrieval models (Llama 3.1 8B, 70B, and GPT4o) consistently exhibit significantly lower average scores compared to retrieval-augmented models. This discrepancy is largely driven by much lower `Cov` scores; for instance, Llama 3.1 8B achieves a `Cov` score of 3.45, while Llama 3.1 8B + OSDS

---

[6] `https://www.perplexity.ai/`. We used the paid subscription version for the experiments. Note that Perplexity Search does not have an API, so we use the `selenium` toolkit and save their final prediction results from the interface. Due to this, we could not retrieve their citation information.

[7] We use their official codebase. As PaperQA2 does not release their retrieval corpus and requires downloading PDF files offline, we downloaded PDF files of papers suggested by our retrieval pipelines and the Semantic Scholar API. Unlike PaperQA2, we do not have access to private or license-protected papers, which may limit the effectiveness of our replication to some extent.

| | Single-paper performance | | | | | | Multi-paper performance | | | | | | Cost |
| | Pub | | Sci | | QASA | | CS | | Multi | | Bio | Neu | CS |
Model	Corr	Cite	Corr	Cite	Corr	Cite	Corr	Cite	LLM	Cite	Cite	Cite	USD / q
Llama3-8B	61.5	0.0	66.8	0.0	14.3	0.0	41.9	0.0	3.79	0.0	0.0	0.0	0.0001
+OSDS	75.2	63.9	75.5	36.2	18.6	47.2	46.7	26.1	**4.22**	25.3	38.0	36.8	0.0001
OS-8B	**76.4**	**68.9**	**76.0**	**43.6**	**23.0**	**56.3**	**51.1**	**47.9**	4.12	**42.8**	**50.8**	**56.8**	0.003
Llama3-70B	69.5	0.0	76.9	0.0	13.7	0.0	44.9	0.0	3.82	0.0	0.0	0.0	0.0004
+OSDS	77.4	71.1	78.2	42.5	22.7	63.6	48.5	24.5	4.24	41.4	53.8	58.1	0.0004
OS-70B	**79.6**	**74.0**	**82.1**	**47.5**	**23.4**	**64.2**	**52.5**	**45.9**	4.03	**54.7**	**55.9**	**63.1**	0.01
GPT4o	65.8	0.0	77.8	0.0	**21.2**	0.0	45.0	0.1	4.01	0.7	0.2	0.1	0.006
+OSDS	**75.1**	73.7	79.3	47.9	18.3	53.6	52.4	31.1	4.03	31.5	36.3	21.9	0.01
OS-GPT4o	74.8	**77.1**	**81.3**	**56.5**	18.7	**60.4**	**57.7**	**39.5**	**4.51**	37.5	51.5	43.5	0.05
PaperQA2	–	–	–	–	–	–	45.6	48.0	3.82	47.2	56.7	56.0	0.3~2.3
Perplexity	–	–	–	–	–	–	40.0	–	4.15	–	–	–	0.002**

Table 2: **Results of SCHOLARQABENCH**. CS, Multi, Bio and Neu indicate SCHOLAR-CS, SCHOLAR-MULTI, SCHOLAR-BIO and SCHOLAR-NEURO, respectively. Corr indicates correctness metrics (accuracy for PubMedQA and SciFact, ROUGE-L for QASA and Overall scores for SCHOLAR-CS) and Cite indicates citation F1. LLM indicates the average score of Org (organization), Rel (relevance), Cov (coverage) as predicted by Prometheus (Kim et al., 2024a). *PaperQA2 is based on GPT4o, and its pricing is dependent on the number of PDF files used during inference. For the 8B and 70B model costs, while evaluations were conducted on our local machines, we estimated costs based on Together.ai pricing. **We used Perplexity Pro (which requires a monthly subscription at $20 USD) and divided this cost by 9,000, which is the maximum number of queries allowed under the Pro subscription. Since the Perplexity UI does not provide snippets for each citation, we were unable to evaluate its citation accuracy.

| | Computer Science | | | Biomedicine | | |
Model	Total #	# of Hallucinated (↓)	Ratio (↓)	Total #	# of Hallucinated (↓)	Ratio (↓)
OS-8B	9.65	0.0	0.0	6.25	0.0	0.0
Llama 3.1 8B	5.20	4.79	92.1%	5.58	5.46	97.6%
Llama 3.1 70B	6.14	4.78	78.1%	6.98	6.74	96.6%
GPT4o	5.74	4.52	78.7%	5.24	4.97	94.8%

Table 3: **Statistics of hallucinated papers in computer science and biomedicine domains**. Our analysis revealed a significant number of non-existent cited papers in predictions made by LLMs without retrieval, which is a problem not observed in OPENSCHOLAR.

(a standard RAG baseline) improves it to 4.01. These results suggest that solely relying on models' parametric knowledge alone is particularly difficult in scientific domains, especially for smaller LMs.
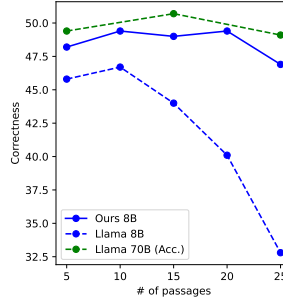
4.3 ANALYSIS

Ablation studies. We conduct ablations to assess the effectiveness of individual components of OPENSCHOLAR (inference and training). Specifically, we remove each of the inference-time procedures, reranking, feedback and attribution, and for OS-8B, we ablate the training, where we use Llama3-8B without any further training.

As shown in Figure 4 (a), removing these components significantly impacts both the overall correctness and citation accuracy of the model outputs. Notably, removing the reranker led to substantial performance drops on both models. The pronounced decline in performance after removing feedback loops in GPT4o suggests that more powerful models greatly benefit from a self-feedback cycle, consistent with Madaan et al. (2023), while it gives limited performance drops in our trained 8B. Additionally, the removal of post-hoc attribution assessments negatively affected both citation accuracy and final output correctness, highlighting the importance of ensuring that models verify their outputs. The significant performance gap between trained versus vanilla OS-8B suggests that further training on high-quality, domain-specific data is key to building efficient, task-specialized LMs. In the next analysis, we demonstrate that training has a significant impact on an LM's ability to effectively utilize more context, while maintaining citation accuracy.

	SCHOLAR-CS	
	Corr	Cite
OS-8B	51.3	47.9
- training	49.4	42.3
- reranking	49.6	28.2
- feedback	51.1	50.2
- attribution	49.3	44.0
OS-GPT4O	57.7	39.5
- reranking	52.4	22.9
- feedback	55.1	31.0
- attribution	55.6	30.6

(a) Ablation of different components of OPENSCHOLAR.

(b) Top N Ablations (Corr) on SCHOLAR-CS.
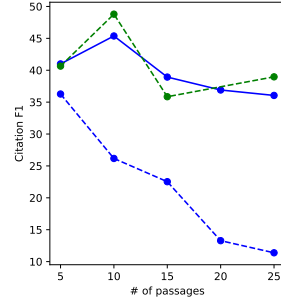
(c) Top N Ablations (Cite-F1) on SCHOLAR-CS.

Figure 4: **Analysis on OPENSCHOLAR:** (a) **Ablation studies** for key components of OPENSCHOLAR training and inference based on different underlying LMs. (b) **Top N docs:** Analysis of the effect of varying the number of context chunks for final downstream tasks. We evaluate final model performance based on citation accuracy and correctness on multi-doc QA tasks, using OPENSCHOLAR 8B and Llama 3.1 8B.

Number of context passages. We analyzed how varying the number of context passages (top N) impacts model performance. Specifically, we experimented with Standard `RAG` and OPENSCHOLAR using our trained 8B model and Llama 3.1 8B, and evaluated both generation accuracy and citation accuracy on SCHOLAR-CS. Figures 4 (b)(c) show the results. Although Llama 3.1 is trained to handle and accept a context length of up to 128K tokens, we found that its performance deteriorates after a certain context size. While increasing the top N context window from 5 to 10 does improve the model's correctness score, expanding further actually worsens both correctness and citation accuracy. This suggests that even though LMs can process large numbers of passages, they may struggle to effectively use them without specialized training, particularly for smaller models.

In contrast, our trained 8B model maintains strong performance for up to $N = 20$ passages. We also found larger models such as Llama 3.1 70B to be more robust to increased context length. In terms of citation accuracy, as shown in Figures 4 (c), Llama 3.1 8B observes quick decline and citation F1 gets as low as 10, while our 8B LM and Llama 70B both maintain around 40 citation F1, although they also see some performance decline.

5 EXPERT EVALUATION

To complement our automatic evaluations and better understand the effectiveness and limitations of OPENSCHOLAR, we conducted human evaluations. This study involved over 100 literature review questions and more than 15 participants, including Ph.D. students, research scientists, and university professors with expertise in the relevant fields. In total, we curated more than 400 fine-grained expert evaluations on human and model answers.

5.1 HUMAN EVALUATION DESIGN

Evaluations against human experts. For human evaluations, we use 108 question-answer pairs from SCHOLAR-MULTI, written by experts. We run three models on these questions to generate answers with citations: GPT4o (without external retrieval), OPENSCHOLAR with GPT4o as the generator (OS-GPT4o), and OPENSCHOLAR with our trained 8B model (OS-8B). Expert annotators are then asked to evaluate the model-generated answers against human-written answers.

Each evaluation involves presenting a question, a model-generated answer, and a human-written answer. Expert annotators then conduct fine-grained assessments of each answer and provide pairwise preference judgments between the two. For fine-grained evaluations, we use the five-scale evaluation criteria described in Section 3 (Cov, Org, Rel), with annotators scoring both model and human answers using the same rubrics. For usefulness (Use), annotators assign scores on a scale from 1-5,

	Fine-grained (1-5, Avg.)			Overall Usefulness	Relative to Human (%)		
	Org	**Cov**	**Rel**	**Use** (%)	Win	Tie	Lose
GPT4o	4.63 **(+0.4)**	4.06 **(-0.2)**	4.50 **(-0.1)**	69.7 **(-13.9)**	31.9	13.8	**54.2**
OS-8B	3.82 **(-0.3)**	4.30 **(+0.7)**	4.00 **(-0.4)**	72.1 **(+8.7)**	**50.8**	12.3	36.9
OS-GPT4o	4.47 **(+0.8)**	4.38 **(+0.9)**	4.30 **(0.0)**	80.0 **(+22.5)**	**70.0**	6.8	23.2

Table 4: **Human evaluation results**. Fine-grained aspect evaluations are conducted on a five-point scale across four aspects using our detailed instructions and rubrics. Values in parentheses represent the relative performance difference; **(+)** indicates that the model shows higher performance, and **(-)** indicates that the human shows higher performance.
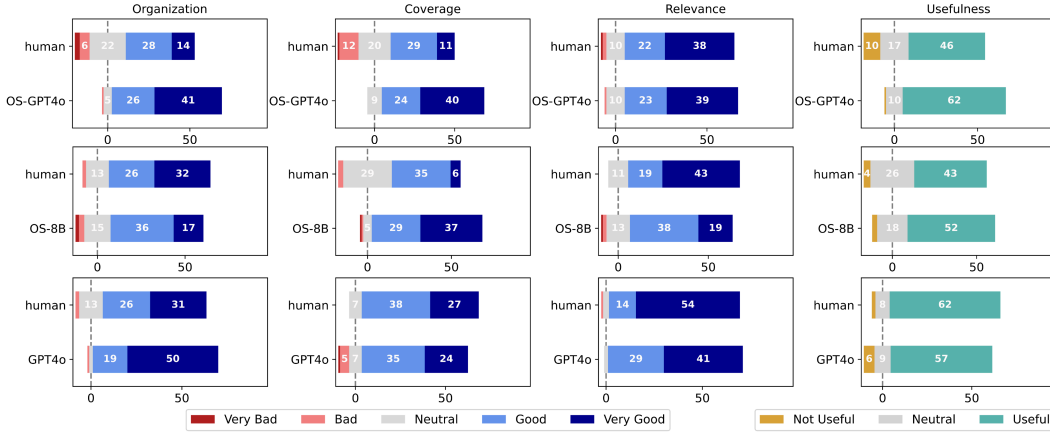
Figure 5: **Fine-grained evaluation results.** Score distributions between 1) GPT4o (top), OPEN-SCHOLAR with 8B (middle), OPENSCHOLAR with GPT4o with Human (bottom).

which we convert into three classes: Not Useful (1-2), Neutral (3), and Useful (4-5). We then calculate the percentage of answers that fall into the Useful category. For pairwise preference, annotators either choose one of the answers or mark a "tie" if they judge both answers to be of equal quality. Optionally, experts provide explanations on why one answer is better than the other.

Expert annotators for answer writing. Our expert annotators for question and answer writing are 12 Ph.D. students and post-doctoral researchers from research institutions across the United States, all of whom have at least three years of research experience and published multiple papers in journals or conferences from their fields. Our annotators' expert areas span computer science (Natural Language Processing, computer Vision, Human-Computer Interaction), physics (Astrophysics and Photonics / Optics), and biomedical (Neuroscience, Bioimaging) domains, and we assign our expert annotators to questions in their expertise. On average, we paid 35-40 USD per person.

Expert annotators for evaluations. A total of 16 expert annotators from the three fields contributed to our evaluations, with 12 of them also participating in answer generation. All expert annotators met the same qualifications as those who composed the answers. To minimize potential biases, we ensured that annotators did not evaluate responses to their own questions by assigning evaluation tasks to different groups of experts. Each instance was reviewed by 1 to 3 expert annotators, depending on availability. The inter-annotator agreement was 0.68 using pairwise comparison with ties, and 0.70 using a relaxed approach, wherein ties were merged. On average, each expert spent five minutes per instance on evaluation, and received compensation ranging from 25–35 USD.

5.2 HUMAN EVALUATION RESULTS

Results of human evaluations. Table 4 presents the average scores for each evaluation aspect, alongside the relative win rates against human responses. Figure 5 illustrates the score distributions for Human, GPT4o, and OPENSCHOLAR with Llama 3 8B and GPT4o. Notably, both OS-GPT4o and our OS-8B versions outperform human answers in over 50% of cases, with their advantage

primarily attributed to their ability to provide a greater breadth and depth of information (coverage; $\boxed{\text{Cov}}$). In contrast, GPT4o, which lacks retrieval capabilities, demonstrates significantly limited coverage and wins in fewer than 35% of cases, with its overall usefulness rated much lower than responses from humans and the other two models. These results highlight that even for state-of-the-art models, synthesizing and answering scientific literature review questions remains a challenging task, consistent with our findings on SCHOLARQABENCH. Overall, OPENSCHOLAR-GPT4o and OPENSCHOLAR-8B are rated as Useful in 80% and 72% of the queries, respectively.

While OPENSCHOLAR with a smaller open 8B LM already suppresses human experts, the 8B model's output is judged to be less organized or fluent than the current state-of-the-art private LM-based OPENSCHOLAR. We found that GPT4o incorporates feedback more effectively, and generally generates longer and more fluent outputs, leading to significantly higher organization scores compared to 8B based OPENSCHOLAR as well as humans.

Effects of length control on model responses. While we found model outputs to often be preferred over human outputs, one potential confounding factor is the significant difference in their output length–OPENSCHOLAR-GPTo and OPENSCHOLAR-8B are 2.4 times and 2.0 times longer than human-written answers, respectively, which may affect human judgment Dubois et al. (2024). To understand the effect of output length, we conducted a controlled experiment: for randomly sampled 50 questions, we generate shorter version of responses for OPENSCHOLAR-GPT4o, by prompting GPT4o to create summaries of the responses that fall under 300 words. As a result, we collected OPENSCHOLAR answers that average around 333 words, which is close to the average human answer length. We then conduct the same human evaluation and evaluate fine-grained and overall responses. On average, the shortened GPT4o scores 4.5 for organization, 4.6 for coverage, and 4.6 for relevance. The shortened OPENSCHOLAR-GPT4o responses are preferred or tied with expert answers in 75% of the queries. The experimental results show that the model's superior performance is not merely due to the increased length of the OPENSCHOLAR answers. Moreover, human annotators' explanations often mention that both shortened OPENSCHOLAR and human answers could be improved by incorporating more details, implying that a 300-word restriction may limit the utility of answers.

Analyses on human explanations for pair-wise explanations. We randomly sampled 59 instances with free-form explanations of pairwise preferences and conducted a manual analysis to identify factors that influence overall preferences. Specifically, we examined whether the explanations referenced one or more of the following four categories: organization, relevance, coverage, and citations. While the first three categories align with the fine-grained human evaluation criteria, the citation category also considers the quality of the cited papers (e.g., whether the system includes key representative papers in the field). Our analysis revealed that 12%, 23%, 29%, and 9% of the explanations cited organization, relevance, coverage, and citations, respectively, as key factors in pairwise decisions. This suggests that coverage plays a crucial role in how humans assess the quality of responses, with annotators largely favoring model-generated answers for their greater coverage and depth of information. However, annotators also noted that the citations provided by models could be improved, pointing out that the suggested papers were occasionally outdated or less relevant compared to more representative work. Appendix 15 shows example explanations.

6 RELATED WORK

Scientific LMs. Scientific LMs have spanned various domains, including biomedical (Phan et al., 2021; Yuan et al., 2022; Li et al., 2023; Luo et al., 2022), medical (Singhal et al., 2022; Gu et al., 2024; Tan et al., 2023; Singhal et al., 2023b), biomedical (Zhang et al., 2023; Fang et al., 2024; Li et al., 2024a), geoscience (Feng et al., 2023), astronomy (Lu et al., 2023) and multidisciplinary science (Thiagarajan et al., 2023; Rajan et al., 2024; Shaikh et al., 2023), with some models such as SciGLM (Wu et al., 2024) and UniSmart (Chi et al., 2024) that aim to cover diverse scientific domains in a single model. Recently, several works show that powerful general-purpose LLMs can also show strong capabilities in scientific tasks, such as medical question answering (AI4Science & Quantum, 2023; Singhal et al., 2022), chemistry experimentation (Zheng et al., 2023b) and applied mechanics (Brodnik et al., 2023). However, the language model's reliance on information memorized within its parameters leads to frequent hallucinations in its output (Li et al., 2024b).

LMs to assist scientists. Recent studies have also examined LLMs' capabilities to assist scientists in performing a range of scientific procedures, including generating novel research ideas (Baek et al.,

2024; Yang et al., 2023) and automating experimental code generation (Huang et al., 2023; Tian et al., 2024). Our work, however, focuses specifically on benchmarking and developing methods for automating literature reviews and addressing questions related to up-to-date research—tasks that are crucial to, and particularly challenging, for scientific inquiry. Several concurrent studies have attempted to build retrieval-augmented pipelines using proprietary LLMs and external APIs (e.g., Semantic Scholar API) for scientific literature review agents (Agarwal et al., 2024; Skarlinski et al., 2024; Wang et al., 2024). While these studies and our research all explore the potential of retrieval-augmented LMs in automating literature synthesis, prior works often relied on proprietary, black-box systems and limited evaluations, which commonly entail small-scale human evaluation or simplified setups such as multiple-choice QA. In contrast, our work introduces a comprehensive benchmark with automated metrics, involves user studies with experts across three scientific disciplines, and develops new methodologies to train specialized open models. OPENSCHOLAR significantly outperforms previously introduced systems and shows superiority over human experts in five domains.

Benchmarks for scientific literature understanding. Several works have developed benchmarks to evaluate models' abilities to understand scientific literature. Prior datasets, such as SciFact (Wadden et al., 2020), QASPER (Dasigi et al., 2021), and QASA (Lee et al., 2023), largely focus on single-paper settings, where the necessary information to answer queries is contained within a single pre-selected paper. However, in real-world scenarios, experts often need to synthesize information from multiple papers to answer questions. To address this gap, SCHOLARQABENCH introduces newly annotated tasks that require reasoning across multiple papers. There are also scientific summarization tasks, such as Multi-XScience (Lu et al., 2020), where models are provided with multiple papers and asked to generate summaries, typically based on the related work sections of those papers. However, in this work, we focus on scenarios where the relevant papers are not specified in advance, making the task more challenging. Recently, Xu et al. (2024) introduced KIWI, a dataset containing 200 questions and human-verified or edited answers generated by state-of-the-art LLMs, with a focus on the NLP domain. KIWI also provides a set of relevant papers that models must consider. While both KIWI and SCHOLARQABENCH feature multi-paper, information-seeking tasks, SCHOLARQABENCH includes both human-written answers and automatic evaluation pipelines. In contrast, KIWI focuses more on human evaluations, and its reference answers are primarily model-generated.

7 CONCLUSION

In order to further research on LM-based systems that can assist scientific progress, we introduce OPENSCHOLAR and SCHOLARQABENCH, which can help navigate the complex, ever-growing task of scientific literature review. OPENSCHOLAR, a retrieval-augmented system, leverages open-checkpoint LLMs and trained retrieval models to iteratively refine scientific output, addressing challenges such as hallucinations and citation accuracy. SCHOLARQABENCH, a novel large-scale benchmark, provides a standardized way to evaluate literature review automation across multiple scientific domains. In evaluations using SCHOLARQABENCH, OPENSCHOLAR demonstrates substantial improvements, outperforming existing systems, including GPT-4o and the concurrent proprietary system PaperQA2. Our expert evaluation across three scientific disciplines reveals that SCHOLARQABENCH, when paired with fully open-checkpoint models and open-access data stores, generates answers that are more helpful than those produced by expert annotators, who required an hour per annotation. This approach also significantly increases coverage. OPENSCHOLAR using our trained 8B and GPT4o achieves a 51% and 70% win rate against human-generated answers. We open-source the OPENSCHOLAR code, data, model checkpoints, datastores, and SCHOLARQABENCH, along with a public demo, to support and accelerate future research efforts.

LIMITATIONS

We highlight several limitations of our work in this section. It is important to note that we do not claim that LM-based systems can fully automate scientific literature synthesis. To further advance research in this area, we are releasing both SCHOLARQABENCH and OPENSCHOLAR to the community.

Limitations of SCHOLARQABENCH. There are several limitations to SCHOLARQABENCH. First, due to the cost and time required to engage expert annotators—individuals with either a Ph.D. or are currently pursuing one in relevant fields—the evaluation dataset with human-written answers is

relatively small (e.g., 110 for CS-LFQA and 108 for expert-written answers). This limited dataset may introduce statistical variance and potential biases stemming from the specific expertise of the annotators. To support future research in expanding the size and scope of SCHOLARQABENCH, we open-source our data and annotation pipelines.

Second, our automatic evaluation pipelines may not always perfectly capture the quality of generated content. For example, in SCHOLAR-CS, we combine various components (e.g., length, excerpts, rubric items) using heuristically determined weight terms. Further, we discovered that often annotators asked for specific kinds of ancillary information in their rubrics—background, elaborations and challenges—even though these aspects might not be strictly required to answer the question. In our experiments, we found that LLMs are proficient at generating the background aspects, which can give them an advantage over systems that directly answer a query but do not satisfy all the constraints of the rubrics. Moreover, future systems could potentially take advantage of the stylistic biases in the rubrics and be prompted to address more rubric elements in a way that does not improve answer quality. Although we carefully analyzed the correlation between final scores and human expert evaluations, there is still room for improvement in refining which aspects should be emphasized and how these scores should be aggregated. Additionally, our evaluations of citation precision and recall are conducted at the sentence level, but we found that some sentences without direct citations are often supported by citations in adjacent sentences. As a result, our precision and recall metrics might be overly strict, potentially underestimating the true citation accuracy. We also note that our annotations were captured at particular times (July 2024 for SCHOLAR-CS and September 2024 for SCHOLAR-MULTI), and may not reflect subsequent scientific developments. Researchers who use our evaluation benchmark should ignore papers published after these dates for a fair comparison.

Lastly, SCHOLARQABENCH primarily focuses on computer science, biomedicine, and physics, with no instances from social sciences or other engineering and scientific disciplines. We recognize that our findings may not fully generalize to other domains, particularly those with more restricted access to paper data.

Limitations of OPENSCHOLAR. While OPENSCHOLAR demonstrates strong performance on SCHOLARQABENCH and in human evaluations, as discussed in the relevant sections, our expert annotators identified several limitations. Despite these issues, we believe OPENSCHOLAR remains a valuable tool for supporting human experts.

First, as highlighted by our expert annotators, OPENSCHOLAR does not consistently retrieve the most representative or relevant papers for certain queries. Enhancing retrieval methodologies by incorporating additional information, such as citation networks or metadata like publication recency, could significantly improve its performance. OPENSCHOLAR outputs may contain factual inaccuracies or unsupported information, particularly in versions based on our 8B model, which has limited capacity for instruction-following and scientific knowledge. Future work can explore training that further improve OPENSCHOLAR-8B. In parallel, although it is competitive, OPENSCHOLAR-GPT4o relies on the proprietary GPT4o API, which may evolve over time, making exact reproduction of results challenging. OPENSCHOLAR does not use license-protected papers during inference time. There are ongoing discussions on how to ensure fair data use in retrieval-augmented LMs, and we leave the exploration of properly incorporating copyright-protected content to future work.

We encourage future research to address these limitations and continue improving LM-based systems for scientific literature review.

Limitations of our human evaluation process. In our human evaluations, annotators performed fine-grained assessments on aspects such as coverage, relevance, organization, and usefulness, while other factors such as citation precision and recall were separately evaluated. As a result, when assessing usefulness or pairwise preferences, annotators may have focused more on the overall quality of writing instead of carefully evaluating factual correctness or citation accuracy. We leave more detailed human analysis on citation accuracy, validity, and factuality for future work.

Our evaluations were conducted by 16 Ph.D. students and postdoctoral professionals, and we made an effort to align their expertise with the topics being evaluated. However, since research often demands deep domain knowledge, the annotators may not have captured more nuanced differences for questions outside their immediate areas of expertise. Additionally, these evaluations were based on 108 questions spanning three scientific disciplines, meaning that findings may not fully generalize to other fields or domains.

AUTHOR CONTRIBUTION

The author contributions are summarized below:

- **Project Lead**: Akari Asai
- **Project Conception**: Akari Asai, Wen-tau Yih, Pang Wei Koh, Hannaneh Hajishirzi
- **OPENSCHOLAR Development**: Akari Asai, Weijia Shi, Rulin Shao, Jacqueline He
- **OPENSCHOLAR Public Demo Development**: Amanpreet Singh, Joseph Chee Cheng, Akari Asai, Rulin Shao, Doug Downey, Matt Latzke
- **peS2o Construction**: Luca Soldaini, Kyle Lo
- **Datastore (peS2o Index) Construction**: Rulin Shao, Jacqueline He, Akari Asai
- **Legal Discussions on Paper Licensing**: Kyle Lo, Luca Soldaini, Doug Downey, Pang Wei Koh, Amanpreet Singh, Akari Asai
- **OPENSCHOLAR-LM Training**: Akari Asai, Weijia Shi
- **OPENSCHOLAR-Retrievers Training and Evaluation**: Akari Asai, Jacqueline He, Rulin Shao
- **SCHOLARQABENCH Design and Conception**: Akari Asai, Pang Wei Koh, David Wadden, Doug Downey, Kyle Lo, Weijia Shi, Amanpreet Singh, Sergey Feldman, Dan Weld
- **SCHOLARQABENCH Collections (Single-paper Tasks)**: Akari Asai
- **SCHOLARQABENCH Evaluation Pipeline Design and Development**: Akari Asai
- **SCHOLARQA-CS Collection and Evaluation**: Doug Downey, Amanpreet Singh, Sergey Feldman, Dan Weld, Mike D'arcy
- **SCHOLARQA-Multi Collection**: Akari Asai, Minyang Tian, Rulin Shao, Jacqueline He, Weijia Shi, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong
- **SCHOLARQA-Neuro, Bio Collection**: Doug Downey
- **Results and Codebases**: Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh
- **Human Evaluation Design**: Akari Asai, Pang Wei Koh, Graham Neubig
- **Human Evaluation Interface Development and Supervision**: Akari Asai, Minyang Tian
- **Manuscript Writing**: Akari Asai, Jacqueline He, Doug Downey, Amanpreet Singh, Kyle Lo, Pang Wei Koh
- **OPENSCHOLAR Public Demo Testing**: Everyone
- **Manuscript Editing**: Everyone
- **Advisory**: Pang Wei Koh, Hannaneh Hajishirzi, Doug Downey, Wen-tau Yih, Graham Neubig, Dan Weld, Luke Zettlemoyer

REFERENCES

Shubham Agarwal, Issam H Laradji, Laurent Charlin, and Christopher Pal. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*, 2024. URL `https://arxiv.org/abs/2402.01788`.

Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*, 2023.

Akari Asai and Eunsol Choi. Challenges in information-seeking QA: Unanswerable questions and paragraph retrieval. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1492–1504, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.118. URL `https://aclanthology.org/2021.acl-long.118`.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics*, 2023. URL `https://aclanthology.org/2023.findings-acl.225`.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=hSyW5go0v8`.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen Marcus McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=4WnqRR915j`.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

Neal R. Brodnik, Samuel Carton, Caelin Muir, Satanu Ghosh, Doug Downey, McLean P. Echlin, Tresa M. Pollock, and Samantha Daly. Perspective: Large Language Models in Applied Mechanics. *Journal of Applied Mechanics*, 90(10):101008, 07 2023. ISSN 0021-8936. doi: 10.1115/1.4062773. URL `https://doi.org/10.1115/1.4062773`.

Chenglei Chi, Qiaozi Cheng, Zheng Wen, Rongzhe Lin, Chunyang Wen, Zhaowei Wang, Cuiling Gao, Jian Zhang, Xu Jiang, Jian Yin, et al. Unismart: A unified framework for solving scientific problems with language models. *arXiv preprint arXiv:2403.10301*, 2024.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2174–2184, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1241. URL `https://aclanthology.org/D18-1241`.

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 Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.365. URL `https://aclanthology.org/2021.naacl-main.365`.

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

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Ziheng Fang, Guangxu Wang, Jinsung Xu, Yifeng Cai, Jingya Wang, Qicheng Qiu, Ruixuan Zhang, Xiaofeng Chen, Jinna Wang, Jiayi Dong, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and knowledge reasoning tasks. *arXiv preprint arXiv:2403.18421*, 2024.

Qiang Feng, Yuxi Li, Jintao Zou, Zhiwei Li, Zhiqiang Ding, Chao Zhang, Qinyan Zhang, Xueqi Hu, Weihao Peng, Xiangyu Meng, et al. K2: A foundation language model for geoscience knowledge understanding and generation. *arXiv preprint arXiv:2306.05064*, 2023.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*, 2023. URL `https://arxiv.org/abs/2305.14627`.

Xiaodan Gu, Zhen Wang, Zhengliang Shi, Hongyan Li, Xiaoye Chen, and Dehong Cheng. Mellama: Foundation model for medical language understanding and generation. *arXiv preprint arXiv:2402.12749*, 2024.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, 2020. URL `https://dl.acm.org/doi/pdf/10.5555/3524938.3525306`.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Mlagentbench: Evaluating language agents on machine learning experimentation. In *International Conference on Machine Learning*, 2023. URL `https://api.semanticscholar.org/CorpusID:263671541`.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*, 2023.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. URL `https://openreview.net/forum?id=jKN1pXi7b0`.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*, 2023. URL `https://arxiv.org/abs/2305.06983`.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL `https://aclanthology.org/D19-1259`.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. Realtime qa: What's the answer right now? *arXiv preprint arXiv:2207.13332*, 2022. URL `https://arxiv.org/abs/2207.13332`.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL `https://openreview.net/forum?id=8euJaTveKw`.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, et al. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models. *arXiv preprint arXiv:2406.05761*, 2024b. URL `https://arxiv.org/abs/2406.05761`.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita Rao, Shaurya Rohatgi, Paul L Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. The semantic scholar open data platform. *ArXiv*, abs/2301.10140, 2023.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moon-tae Lee. Qasa: advanced question answering on scientific articles. In *International Conference on Machine Learning*, pp. 19036–19052. PMLR, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020. URL `https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf`.

Junfeng Li, Junjie Gao, Siru Zhang, Yiwen Wang, Xinhang Yan, Hongyan Liu, Shiping Yang, Jie Qiao, and Qian Zhan. Biomistral: A collection of open-source pretrained large language models for biomedicine. *arXiv preprint arXiv:2402.10373*, 2024a.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10879–10899, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. URL `https://aclanthology.org/2024.acl-long.586`.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14255–14273, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. URL `https://aclanthology.org/2024.acl-long.769`.

Yuwei Li, Yingbo Bao, Zhihong Zhang, Yun Yao, and Jianyi Zhang. Bioinspired large language models for protein design. *Advanced Science*, 10(36):2306724, 2023.

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. corr abs/2307.03172 (2023). *arXiv preprint arXiv:2307.03172*, 10, 2023a.

Nelson F Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*, 2023b. URL `https://arxiv.org/abs/2304.09848`.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL `https://aclanthology.org/2020.acl-main.447`.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Weizhe Lu, Chao Xu, Julien Becker, and Jie Liu. Astrollama: Towards specialized foundation models in astronomy. *arXiv preprint arXiv:2309.06126*, 2023.

Yao Lu, Yue Dong, and Laurent Charlin. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. *arXiv preprint arXiv:2010.14235*, 2020.

Renqian Luo, Liai Sun, Yingce Xie, Zhiting Jiang, Yangbin Gu, Kun Shi, Dejia Xiong, Sheng He, Zhen Xu, and Tao Qin. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *arXiv preprint arXiv:2210.10341*, 2022.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self. *Feedback*, 2023. URL https://arxiv.org/abs/2303.17651.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. Expertqa: Expert-curated questions and attributed answers. *arXiv preprint arXiv:2309.07852*, 2023. URL https://arxiv.org/abs/2309.07852.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023. URL https://aclanthology.org/2023.acl-long.546.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. Fine-grained hallucination detection and editing for language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=dJMTn3QOWO.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. Scifive: a text-to-text model for biomedical literature. *arXiv preprint arXiv:2106.03598*, 2021.

Karthik Rajan, Hatim T Ali, Supraj Ghimire, Sarvesh Bhooshan, Yejin Chen, and Zhangyang Dong. Indus: A scientific foundation model family for domain-specific and general-purpose scientific tasks. *arXiv preprint arXiv:2405.10725*, 2024.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 2023. URL https://arxiv.org/abs/2302.00083.

Shishir G Shaikh, Jaideep Ramachandran, Varnith Nanda, Benjamin Lunt, Miltiadis Allamanis, Daman Sharma, Sebastien Bubeck, and Prateek Jain. Darwin: Data analytics and reasoning with large language models for science. *arXiv preprint arXiv:2308.13565*, 2023.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei Koh. Scaling retrieval-based language models with a trillion-token datastore. *arXiv preprint arXiv:2407.12854*, 2024. URL https://arxiv.org/abs/2407.12854.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023a.

Zeming Singhal, Charles Sutton, Adam Mottram, Owain Lavelle, Iz Beltagy, Leonardo Neves, Kyle Lo, Stephanie Hyland, Michael Wainwright, Alexander Wettig, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023b.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. Language agents achieve superhuman synthesis of scientific knowledge. *preprint*, 2024. URL https://paper.wikicrow.ai.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, 2024. URL https://arxiv.org/abs/2402.00159.

Cheng Tan, Miao Huang, Xianxin Huang, Qian Fu, and Bo Wu. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2023.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.

Jayaraman J Thiagarajan, Rushil Anirudh, Peer-Timo Bremer, Jim A Gaffney, Bhavya Kailkhura, Ravi Madduri, Ankur Mondal, Siddharth Nemani, Tirthankar Rastogi, Rushil Sen, et al. Forge: Foundation models for science and engineering. Technical report, Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), 2023.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, et al. Scicode: A research coding benchmark curated by scientists. *arXiv preprint arXiv:2407.13168*, 2024. URL https://arxiv.org/abs/2407.13168.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609.

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

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, et al. Autosurvey: Large language models can automatically write surveys. *arXiv preprint arXiv:2406.10252*, 2024.

Fangyuan Wu, Hongbin Zhang, Lei Zhang, Wenyue Zhu, Xu Liu, Yuqi Qin, Linzheng Wang, Xinqi Zhou, Bin Liang, Dawei Zheng, et al. Sciglm: Training scientific language models with self-reflective in-context learning. *arXiv preprint arXiv:2401.07950*, 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with compression and selective augmentation, 2023a. URL https://arxiv.org/abs/2310.04408.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering. *arXiv preprint arXiv:2305.18201*, 2023b.

Fangyuan Xu, Kyle Lo, Luca Soldaini, Bailey Kuehl, Eunsol Choi, and David Wadden. KIWI: A dataset of knowledge-intensive writing instructions for answering research questions. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 12969–12990, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.findings-acl.770`.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and E. Cambria. Large language models for automated open-domain scientific hypotheses discovery. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL `https://api.semanticscholar.org/CorpusID:261557055`.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. Biobart: Pretraining and evaluation of a biomedical generative language model. *arXiv preprint arXiv:2204.03905*, 2022.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*, 2023. URL `https://arxiv.org/abs/2305.06311`.

Yuqi Zhang, Zihao Zhao, Lanqing Hu, Shuai Wang, Penghui Jiao, Min Leng, Yuzhi Liu, Guotong Li, Chengming Xu, Chenhui Lin, et al. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023a. URL `https://openreview.net/forum?id=uccHPGDlao`.

Yuxiang Zheng, Shichao Sun, Lin Qiu, Dongyu Ru, Cheng Jiayang, Xuefeng Li, Jifan Lin, Binjie Wang, Yun Luo, Renjie Pan, et al. Openresearcher: Unleashing ai for accelerated scientific research. *arXiv preprint arXiv:2408.06941*, 2024. URL `https://arxiv.org/abs/2408.06941`.

Zhiling Zheng, Zichao Rong, Nakul Rampal, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. A gpt-4 reticular chemist for guiding mof discovery. *Angewandte Chemie International Edition*, 62(46):e202311983, 2023b.

APPENDIX

A RELEASED ARTIFACTS

We release a set of artifacts to facilitate future research:

Demo	⊙	`openscholar.allen.ai/`
OpenScholar	⊙	`github.com/AkariAsai/OpenScholar`
ScholarBenchQA	⊙	`github.com/AkariAsai/ScholarBench`
OpenScholar-8B LM	🤗	`OpenScholar/OpenScholar_Llama-3.1-8B`
OpenScholar-Retriever	🤗	`OpenScholar/OpenScholar_Retriever`
OpenScholar-Reranker	🤗	`OpenScholar/OpenScholar_Reranker`
OpenScholar-DataStore-V2	🤗	`OpenScholar/OpenScholar-DataStore-V2`
OpenScholar-DataStore-V3	🤗	`OpenScholar/OpenScholar-DataStore-V3`
ScholarBench (Data)	🤗	`OpenScholar/ScholarBench`
OpenScholar (Training Data)	🤗	`OpenScholar/OS_Train_Data`
Expert Evaluation	⊙	`AkariAsai/OpenScholar_ExpertEval`

B MORE DETAILS ON SCHOLARQABENCH

B.1 GOAL OF SCHOLARQABENCH

There are two key principles to SCHOLARQABENCH: it should serve as a realistic benchmark for literature review, and as a reproducible, mutifaceted evaluation pipeline.

Realistic benchmark for literature review (Section 3.1). SCHOLARQABENCH integrates tasks from two key sources: (i) curated and adapted existing datasets related to literature review tasks and annotated by scientists, and (ii) three new datasets, annotated by Ph.D. experts from three scientific domains that reflect realistic literature review scenarios, such as information synthesis from multiple papers. The tasks in SCHOLARQABENCH require different output formats and disciplines.

For multi-paper tasks, we instruct our expert annotators to formulate information-seeking questions—questions they are genuinely interested in finding answers to, rather than questions they already know the answers to or that could be answered using small text chunks from a single paper (Asai & Choi, 2021; Choi et al., 2018). We found this approach crucial for curating realistic questions that real-world scientists might ask. These questions are typically more detailed, contextualized (e.g., "I am planning to generate synthetic training data using GPT4o and filter noisy data using GPT4o, but I'm concerned that GPT4o is not filtering models in this case, as it may favor its own generation."), and require nuanced, long-form answers rather than simple yes/no or multiple-choice responses. We collected human-written answers or rubrics to ensure reliable evaluation, rather than relying on answers generated by state-of-the-art proprietary LMs, as done in Xu et al. (2024); Malaviya et al. (2023). While these proprietary models are powerful, they still exhibit limitations, such as hallucinations from lacking domain knowledge, biases, and rapid information changes, making them unsuitable for consistent evaluation with newer models. Additionally, using model-generated answers as references can unfairly favor models from the same family, introducing possible evaluation biases. To avoid these issues, we curated expert-written answers for SCHOLARQA-CS and SCHOLARQA-MULTI.

Reproducible multi-face evaluation pipelines (Section 3.2). Due to low correlation between conventional similarity-based metrics such as ROUGE (Xu et al., 2023b; Malaviya et al., 2023) and human judgments, evaluations of long-form generation tasks in expert domains have typically relied on small- to medium-scale expert annotations (Zheng et al., 2024; Singhal et al., 2023a; Si et al., 2024). While expert human evaluation is valuable (as we detail in Section 5), it requires significant costs for hiring annotators and is hard to reproduce. To address these limitations, we introduce automated evaluation pipelines that comprehensively assess the quality of long-form generation outputs from important aspects such as citation correctness or coverage.

(a) Subject distributions of expert-annotated answers

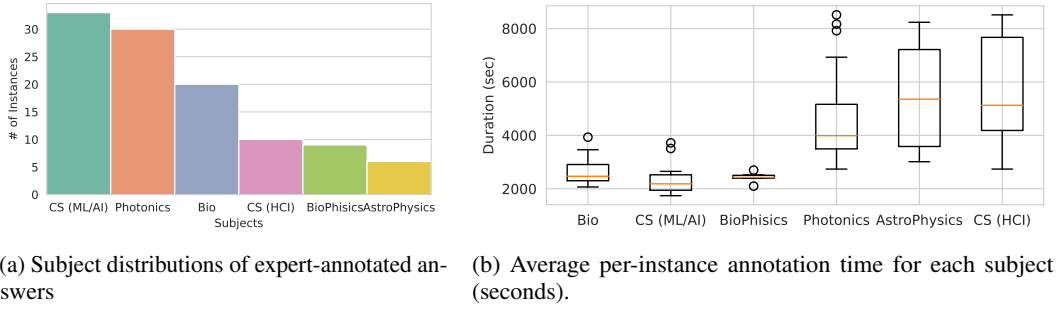(b) Average per-instance annotation time for each subject (seconds).

Figure 6: **Analysis of human-written answers:** (a) shows the distribution of instances per subject, and (b) shows the average annotation time per instance per subject.

B.2   DATA CURATION DETAILS

B.2.1   DETAILS OF MODIFICATION OF SINGLE-PAPER TASKS

SciFact. SCIFACT (Wadden et al., 2020) is a dataset of 1.4K expert-written scientific claims in the biomedical domain, paired with evidence-based abstracts annotated with labels and rationales. The original task involves three subtasks—paragraph selection, sentence selection, and label prediction—based on a collection of 5,000 abstracts. However, we reformulate this as an open-retrieval label prediction task, where the model is given only a query and must predict the label from a larger corpus of 40 million passages. We exclude queries labeled as `not enough information` and retain only instances labeled as either `supports` (true) or `contradicts` (false).

PubMedQA. We leverage PUBMEDQA (Jin et al., 2019), which has 1k expert-annotated (yes/no/maybe) QA data on PubMed paper abstracts. Similarly to SCIFACT, we keep instances with yes or no labels, and discard the original abstract passage to formulate the task as an open setup.

QASA. QASA (Lee et al., 2023) is a single paper QA dataset consisting of 1,798 novel question answering pairs that require reasoning over scientific articles in AI and ML. We evaluate the model's ability to sufficiently answer a detailed question about the target paper. While they provide three subtasks (answer selection, rational generation and answer compositions) and the end-to-end full-stack QA, we evaluate the models' performance based on full-stack QA.

B.2.2   DETAILS OF DATA COLLECTIONS OF MULTI-PAPER TASKS

Recruiting annotators. For data curation, we recruit expert annotators through UpWork and inter-institution channels, ensuring that they meet the following criteria: (1) they hold a Ph.D. or are enrolled in a relevant Ph.D. program, (2) they have over three years of research experience in the field, and (3) they have published papers in the target areas. In total, we recruited over 20 annotators, including Ph.D. students, postdoctoral fellows, professors, and research scientists, across various multi-paper subsets in the target domains.

Annotation instructions. Table 5 outlines our annotations instructions to collect rubrics for SCHOLARQA-CS, and Table 6 shows annotation instructions given to the annotators for SCHOLARQA-BIO and SCHOLARQA-NEURO. Table 7 shows the instructions given to annotators for SCHOLARQA-MULTI.

Statistics of human-written questions and answers in SCHOLARQA-MULTI. Figure 6a shows the subject distribution, and Figure 6b depicts the average time in seconds that experts spend on annotating each answer across different subjects. Annotators on average spend at least 30 minutes per answer, but those from some subjects, such as Physics, took over an hour (e.g., approximately 5,000 seconds) to complete a single annotation. This demonstrates that the task is highly time-consuming and challenging even for domain experts.

Instructions

This project is aimed at building a dataset that can be used to evaluate how well large language models can answer scientific questions. Specifically, for a set of complex scientific questions that require multiple documents to answer, you're asked to provide the points that are important to include in an answer, along with supporting text from the scientific literature and other sources. We will release your output to the community in the form of a dataset that allows researchers to assess how well AI systems can answer complex scientific questions.

For each question, you are given two documents -

- A key ingredients document that you will fill-in with components that you feel are necessary to include in a correct answer to the question. Key ingredients are short (1-4 sentences) statements of an important item to include in a correct answer, such as "near the beginning, the answer should should briefly define state space models and transformers and detail their technical differences."
- (do not read right away) A sources document that you will use for reference, giving the output for the question from different online services

Perform the following steps:

1. First, BEFORE reading the existing text in the sources document, based on your knowledge of the question and about 40 minutes of literature review using Google Scholar or Semantic Scholar (but NOT any online large language model services like ChatGPT or similar), compose a set of key ingredients of the answer and supporting quotes from the documents. Starting from the corresponding key ingredients file for the question, add your key ingredient to the list along with zero or more snippets of text from the documents under "supporting quotes." Each snippet should be less than 150 words. After each quote, paste a link from the source (if you quote a document multiple times, paste a link after each quote). Your key ingredients should typically have supporting snippets, but may not always if you wish to include an important fact that you know but cannot find support for.

2. Then, in a total of about 20-50 additional minutes, read through the sources document for the question (you may have to skim over portions, as some source documents are long), and complete the set of key ingredients in the answer based on the combination of what you'd listed in step 1, and what you learn from the sources. You should delete or change ingredients you added after step one, if appropriate.

   Note: you can copy and paste text directly from the sources document as Supporting Quotes, and for those quotes you do NOT need to include a link, just paste the relevant portion in quotation marks.

In your key ingredients, you do not need to cover general stylistic points (e.g., that the answer should be well-organized and include citations), those characteristics are already captured in a general rubric we've written. Instead, focus on the specific content needed in the answer, organized as bullet points. When capturing the ingredients, follow the template of the document to separate them into **Most important** and **nice-to-have**. The "most important" items should be ones that, if they were missing, would lead the answer to be misleading or incomplete. The "nice to have" should include other helpful elements, but ones that could potentially be omitted in a short answer.

To help you understand the task, an example of a completed ingredients file, and the corresponding sources file is provided.

Table 5: **SCHOLARQA-CS Annotation Instructions**: Instructions given to our expert annotators for producing key ingredients for SCHOLARQA-CS questions.

Instructions

In this task, you are asked to generate complex scientific questions that biomedical scholars might reasonably ask about scientific literature. These should be broad literature review questions that require multiple papers to answer, e.g.: "What are the various molecular mechanisms through which chronic inflammation contributes to the development of colon cancer?" This question requires multiple papers to answer, as there are a variety of different studies on such molecular mechanisms. By contrast, an example of a question we are NOT looking for would be "What was the efficacy of Pembrolizumab in reducing tumor size in patients with advanced melanoma based on the S1801 clinical trial?", since a single paper (the results of the clinical trial) holds the answer.

To help ensure we obtain a wide variety of questions, we'd like you to use existing papers as inspiration for your questions. You can think of it as trying to guess what kinds of literature review questions the authors of the paper may have had while they were performing the research in the paper. Specifically, follow the following steps:

1. Pick a paper of interest in your domain. This can be any paper, but review or survey articles can be an especially useful source of inspiration as they cover broad topics encompassing many previous papers. Also, previous work or future work sections of papers can be good sources.

2. Write up to 5-10 literature review questions that are relevant to the paper. These should be: (a) Questions that generally require multiple documents to answer; (b) Questions about the published literature—the kinds of questions that a scientist could attempt to answer by searching and reading papers (i.e., they don't necessarily require new experiments)

We have provided a Google Sheet with three columns, "question," "inspiring paper," and "estimated number of papers needed to answer (1/10/100/1000+)," for you to list each of your questions and roughly how many papers you think a person would need to examine in order to answer it accurately.

While it is important to choose papers on topics you are familiar with, please also strive to choose a diverse set of papers on a variety of subtopics, and do not write more than 10 questions about any single paper. Paste each of your questions and the Semantic Scholar link for the inspiring paper in the spreadsheet (in the unlikely event that the link cannot be found, please paste the bibliographic information for the paper in APA format).

As a concrete example, consider this paper:

- -

You do not need to make any attempt to determine the answers to your questions, or even try to verify whether such answers exist. It is okay if some of your questions turn out to be unanswerable (for example, for Q1 above, it might be the case that there are no known biomarkers; or for Q2, perhaps no such studies exist). We want you to choose the most realistic literature review questions you can, and sometimes realistic questions will not have answers in the current literature.

Table 6: **SCHOLARQA-BIO, NEURO Annotation Instructions**: Instructions given to our expert annotators for authoring questions in biomedicine. The example questions in this passage were composed with assistance from GPT.

B.3 EVALUATION METRICS DETAILS

B.3.1 SCHOLARQA-CS CORR EVALUATION

For SCHOLARQA-CS, we employ expert-annotated rubrics to evaluate the generated answers. Each rubric has two criteria—general (accounting for 40% of the score) and annotation-driven (60%). The general criterion covers the evaluation of length, expertise, citations, and excerpts, while the annotation criterion involves assessing the extent to which each specific key ingredient (and associated quotes) identified by annotators is present in the answer, on a scale of 0 to 1. Each must-have ingredient is considered twice as important as a nice-to-have. However, there is no distinction

> **Instructions**
> You will be given a question and your task is to annotate your answers with appropriate citations. You are allowed to use tools like Google Search, Google Scholar, or Semantic Scholar to find sources. However, you may not use any language model-powered systems such as ChatGPT, GPT-4, Claude, Perplexity, or you.com.
> **Steps to Follow:**
> 1. Start the Stopwatch: Before you begin reading the question, start a stopwatch to track how long it takes you to complete the task.
> 2. Answer the Question: After reading the question, write your answer. Ensure that every sentence that requires a citation is followed by a citation number (e.g., [1]).
> 3. Track Time: Once you have finished writing your answer, stop the stopwatch and note the exact duration it took to complete the task.
> **Answer Format:**
> **Output**: Write your answer under `Output`. Ensure that all citation-worthy sentences include at least one citation, indicated as [citation number] (e.g., [1]).
> **Citations**: For each citation you use (denoted as [i] in your answer), you need to fill out the following information:
> `citation_i_title`: Provide the title of the paper.
> `citation_i_corpus_id`: Semantic Scholar Corpus ID for the citation.
> `citation_i_text`: Copy the relevant paragraph from the paper that supports your answer. In this context, a "paragraph" refers to a set of sentences separated by a new line in the paper. Below, we show an example of paragraphs. If there's no new line and one paragraph gets really long e.g., over 0.5 pages, then please copy and paste the core sentences including the useful information.
> `citation_i_url`: Provide the URL link to the paper (e.g., arXiv abstract page, Semantic Scholar page, PubMed abstract page, etc.).

Table 7: **SCHOLARQAMULTI Annotation Instructions**: Instructions given to our expert annotators to annotate question and answers.

Aspect	Definition
Relevance	Does the response stay on topic and maintain a clear focus to provide a useful response to the question?
Coverage	Evaluate if the output provides sufficient coverage and amount of information.
Organization	Evaluate if the output is well-organized and logically structured.
Overall Usefulness	Evaluate if the output is useful overall.

Table 8: **Evaluation protocols for writing quality**.

between individual ingredients of the same type. Using LLM-as-a-judge (Zheng et al., 2023a), a score is assigned for each criterion, and a weighted sum is computed across the criterion to obtain a final score. This score emphasizes the presence of citations and excerpts, such that systems lacking either tend to underperform even when they score relatively high on the remaining criteria, as seen in our experiments.

B.3.2 CONTENT QUALITY AND ORGANIZATION

An overview of the evaluation aspects is in Table 8. We use Relevance, Coverage and Organization for automatic evaluation. We found that existing evaluator LMs struggle with evaluating overall usefulness, and tend to be over-optimistic.

Evaluation instructions and rubrics. We show annotator rubrics for organization, coverage, relevance and overall usefulness in Tables 10,11, 12, and 13, respectively.

Prometheus configuration. For evaluations, we combine Prometheus BGB (Kim et al., 2024b) (`prometheus-eval/prometheus-bgb-8x7b-v2.0`) and Prometheus v2 (Kim et al.,

2024a) (`https://huggingface.co/prometheus-eval/prometheus-8x7b-v2.0`).
We found that Prometheus BGB generally works well, while on $\boxed{\text{Rel}}$ it sometimes gives scores that are much higher than human assessments, especially for GPT4o. Consequently, we use Prometheus BGB for $\boxed{\text{Org}}$ and $\boxed{\text{Cov}}$, and Prometheus v2 for relevance. We use human-written answers as gold references, which are shown to improve Prometheus' correlation with human evaluation. We use these models with `vllm`. Following the default setup, we set the maximum number of new tokens to 512, top-p to 0.95, and temperature to 0.01.

B.3.3 CITATION ACCURACY

To evaluate citation accuracy, we use `osunlp/attrscore-flan-t5-xl` (Yue et al., 2023), which is FLAN-T5-XL trained on a mixture of attribution tasks. We follow citation precision and recall formulation from Gao et al. (2023) and compute citation precision and recall at the sentence level. We discard sentences under 50 characters, as these sentences are often paragraph or subsection headers that do not require citations. We use the original citation evaluation instructions from Yue et al. (2023), as shown in Table 9.

> As an Attribution Validator, your task is to verify whether a given reference can support the given claim. A claim can be either a plain sentence or a question followed by its answer. Specifically, your response should clearly indicate the relationship: Attributable, Contradictory or Extrapolatory. A contradictory error occurs when you can infer that the answer contradicts the fact presented in the context, while an extrapolatory error means that you cannot infer the correctness of the answer based on the information provided in the context.
> **Claim:**
> **Reference:**

Table 9: Evaluation instruction for citation evaluation (adapted from Yue et al. (2023)).

C MORE DETAILS ON OPENSCHOLAR

C.1 TRAINING A SCIENTIFIC BI-ENCODER θ_{bi}

For θ_{bi}, we follow the unsupervised training methodology from Izacard et al. (2022), and continually pre-train the Contriever bi-encoder on a mixture of peS2o version 2, CCNews, and Proofpile2 (Azerbayev et al., 2024) data for 500k steps, using a batch size of 4,096 and a learning rate of 0.00005. We initialize the model checkpoint from Contriever.

C.2 TRAINING A SCIENTIFIC CROSS-ENCODER θ_{cross}

Paragraphs that score above 3 are labeled as positive, while those that score below 3 are labeled negative. Finally, we select the top N passages based on this process, passing the top paragraphs **P** to the generator LM. For θ_{cross}, we fine-tune the BGE-large reranker for 5 epochs on our newly created training data for five epochs, using a learning rate of 6e-5.

C.3 TRAINING DETAILS OF GENERATORS $\mathcal{G}$

Training data statistics. Figure 7 shows the training data distribution. "Tulu" indicates general-domain instruction-tuning data from Ivison et al. (2023) and SciRIFF Wadden et al. (2024) indicates task-specific data from SciRIFF. For both datasets, we ensure that we do not include any
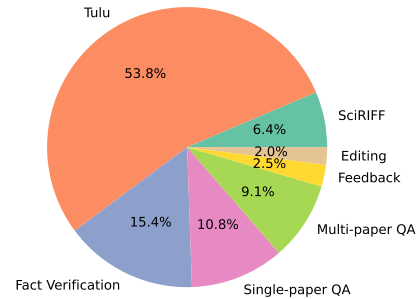

Figure 7: Generator training data distribution.

data used for evaluation, namely PubMedQA, SciFact and QASA. In total, this leads to 130,135 training instances.

Training hyperparameters. For OS-8B and 70B, we initialize model checkpoints from Llama 3.1 8B and 70B Instruct (Dubey et al., 2024). We use `torchtune` [8] for training.

We train both models for two epochs using learning rates of 5e-6 and 1e-4, respectively, with a maximum context length of 10k and batch size of 1 and gradient accumulation step of 2 with bf16. We use AdamW (Loshchilov & Hutter, 2019) as an optimizer. For the 70B model, due to its high memory usage requirements, we conduct LORA tuning. We use 8 A100 GPUs with 80 GB memory for training.

Instructions
Evaluate if the answer is well-organized and logically structured. An acceptable response should: 1. be clearly structured, grouping related points together for a logical flow., and 2. be coherent, without any contradictions or unnecessary repetition.

- -

Score 1: Poor Organization
- The response is disorganized, with no clear structure.
- Points are scattered without any logical order, making it difficult to follow.
- The text lacks coherence, and there are contradictions or irrelevant repetitions throughout.

Score 2: Basic Organization
- The response has some organization, but the structure is inconsistent. - There are occasional lapses in coherence, with minor contradictions or repetitive statements that disrupt the overall clarity.

Score 3: Moderate Organization
- The response is generally well-organized, with a clear structure that is mostly maintained. - Points are grouped logically, though there may be some minor lapses in flow or coherence. The answers have several clear paragraphs, each of which clearly discuss some core points. - The text is mostly clear, with only occasional repetition or slight contradictions.

Score 4: Strong Organization
- The response is well-organized, with a clear and logical structure that is followed consistently. - Points are effectively grouped, and the flow is smooth. - There may be minor lapses in coherence, but overall, the response is clear and easy to follow, with minimal repetition or contradictions.

Score 5: Exceptional Organization
- The response is exceptionally well-organized, with a flawless logical structure. - Points are grouped perfectly, and the flow between them is seamless. - The text is coherent throughout, with no contradictions or unnecessary repetition, making the argument clear and compelling.

Table 10: Evaluation rubrics for organization.

D MORE RESULTS

D.1 FULL RESULTS ON SCHOLARQABENCH

E MORE ANALYSIS

E.1 COMPARISON OF PES2O V2 AND V3

Figure 10 shows the distributions of the top retrieved papers using our retrieval systems on different datastores (peS2o v2 and v3).

---

[8] https://github.com/pytorch/torchtune

> **Instructions**
> Evaluate if the output provides sufficient coverage and amount of information. The output should: 1. (coverage) Offer a comprehensive review of the area of interest, citing a diverse range of representative papers and discussing a variety of sources, not just a few (1-2 papers) 2. (depth) Provide enough relevant information to understand each discussion point.
>
> -
>
> **Score 1:**
> - **Severely Lacking Coverage**: The output lacks coverage of several core lines of research or focuses predominantly on a single line of work, missing a holistic view of the area. - **Greatly Limited Depth of Information**: The output either lacks essential details needed to fully understand the topic (e.g., definitions of methods, relationships between methods).
> **Score 2:**
> - **Partial Coverage**: The output covers some key aspects of the area but misses significant lines of research or focuses too narrowly on a few sources. It lacks a well-rounded view and fails to adequately represent the diversity of work in the field.
> - **Limited Amount of Information**: The response provides some relevant information but leaves out important details that would help in understanding the topic fully.
> **Score 3:**
> - **Acceptable Coverage / Overview**: The output discusses several representative works and provides a satisfactory overview of the area. However, including more papers or discussion points could enhance the answer significantly. It addresses the core aspects of the question but may miss some details.
> - **Acceptable Amount of Relevant Information**: The output provides a reasonable amount of relevant information, though it may lack some helpful details.
> **Score 4:**
> - **Good Coverage**: The output offers good coverage of the area, discussing a variety of representative papers and sources. While it provides a broad overview, it may miss a few minor areas or additional papers that could enhance the comprehensiveness.
> - **Mostly Sufficient Amount of Information**: The response includes most of the necessary and relevant information to understand the topic. It avoids excessive irrelevant details, but a few points might benefit from deeper exploration or more specific examples.
> **Score 5:**
> - **Comprehensive Coverage**: The answer covers a diverse range of papers and viewpoints, offering a thorough overview of the area. It includes additional important discussion points not explicitly mentioned in the original question.
> - **Necessary and Sufficient Amount of Information**: The response provides all necessary and sufficient information.

Table 11: Evaluation rubrics for coverage.

E.2   DATASTORE ANALYSIS

In this section, we conduct a brief analysis on datastores used in OPENSCHOLAR. Figure 10 shows the distributions of top 20 retrieved papers from the two datastores, peS2o v2 and v3 for the same SCHOLARQA-CS queries. Note that our retrieval models are trained on peS2o v2 data. Although our model is not directly trained on the new version 3 datastore, we found that it can constantly retrieve relevant papers from the newer datastore at test time, resulting in many papers from 2023 - 2024.

E.3   HUMAN EVALUATION

Automatic evaluations of human and model-generated answers. We ran our automatic evaluation pipelines, including citation precision, recall, and writing quality assessments, on human- and model-generated answers. Table 14 provides basic statistics for human and OPENSCHOLAR-generated responses, detailing the average length, the number of cited articles, and the scores for each evaluation metric. We observed that human answers are generally shorter and reference fewer papers compared to model-generated outputs. While a concurrent study normalizes human-written answers using an

Instructions
Evaluate if the response stays on topic and maintain a clear focus to provide a useful response to the question. Specifically, the output should: 1. fully address the core points of the original question and satisfy your information needs if they are factual 2. not contain a lot of minor information that is not related to the original question.

- -

Score 1: Off-topic
- The content significantly deviates from the original question, making it difficult to discern its relevance or distract the user.
Score 2: Frequently off-topic with limited focus
- The response addresses the question to some extent but often strays off-topic. - There are several sections with irrelevant information or tangential points that do not contribute to answering the main question. - These distractions make it difficult to maintain a clear focus and reduce the overall usefulness of the response.
Score: 3: Somewhat on topic but with several digressions or irrelevant information
- While the response still centers around the original question, there are frequent deviations that distract from the main question or redundant information.
Score 4: Mostly on-topic with minor deviations
- The response stays largely on-topic, with a clear focus on addressing the question. However, there may be a few minor digressions or slightly irrelevant details that momentarily detract from the main focus. - These deviations are infrequent and do not significantly undermine the overall clarity or usefulness of the response. .
Score 5: Focused and entirely on topic
- The response remains tightly centered on the subject matter with enough depth and coverage of the core information, with every piece of information contributing directly to a comprehensive understanding of the topic.

Table 12: Evaluation rubrics for relevance.

LM to control for confounding factors (Si et al., 2024), we retain the original human and model answers for our human evaluation and conduct extensive analysis to understand how these differences influence human preferences.

Human evaluation overview. Figure 11 shows the overview of human evaluation.

Qualitative analyses on experts' explanations of pair-wise preference. Table 15 shows experts' explanations on pair-wise preferences.

Agreements between humans and LLM-as-a-judge.

We also examine the alignment between human assessments and LLM assessments for evaluating the fine-grained quality of responses. Specifically, we assess both the accuracy and the mean absolute error of the evaluator LLM (i.e., `prometheus-eval/prometheus-bgb-8x7b-v2.0`), using human annotations as the gold standard. Table 16 presents the results.

Figures 13, 14, and 15 show the confusion matrices between human and Prometheus predictions. Confusions often happen between adjacent classes (e.g., 4 v.s. 5), and the evaluator LMs rarely confuse negative classes (lower than 3) and positive classes. Overall, for organization and coverage, Prometheus shows a weak correlation and about 50% accuracy, with disagreements that often occur between adjacent classes (e.g., scoring a 4 versus a 5). When we compare the differences between the average scores of each aspect, as predicted by human and evaluator LMs, on relevance, coverage, and overall usefulness, the absolute difference is less than -0.2, and the ranking between models remains the same.

Qualitative examples of model and human answers. Table 17

> **Instructions**
> Do you think the provided answers are overall helpful and assist your literature review? In particular:
>
> -
>
> **Score 1: Unhelpful**
> - The response does not answer the question or provide rather confusing information. - It does not serve as a useful starting point for learning or writing about the area or understanding the literature.
> **Score 2: Better than searching by myself from scratch but limited utility**
> - The response provides at least one helpful paper that I can take a closer look at. - However, overall the discussion in the answer is somewhat irrelevant and does not provide helpful information.
> **Score 3: Provides some useful discussions and papers, although I need to read individual papers by myself**
> - The response is generally helpful and provides at least 2-3 helpful papers that I wasn't aware of. I can start looking at some of the suggested papers. - It provides some good overview of the areas and multiple viewpoints, which I can dive into. - Yet, I may need to verify some details or consult additional core research papers to ensure completeness.
> **Score 4: Useful. I may try to verify some details, but overall the output gives me a great summary of the area of my interest.**
> - The answer gives me a great set of papers highly related to my question. - The answer provides a great overview and detailed information. - I may need to check some minor details from 1-2 specific papers included in the answers, but overall the provided answer is useful with little need of additional editing.
> **Score 5: Super Useful. I can fully trust the answer**
> - The response provides a comprehensive overview of the area, and sufficiently answers my question. - I don't think I need to search additional papers or check details by myself.

Table 13: Evaluation rubrics for overall helpfulness.

	Length	Inf. time (min)	# of Citations	Prec.	Rec.	Org.	Cov.	Rel.
Human	289.5	56.8	6.0	44.4	41.5	–	–	–
OS-GPT4o	716.3	1.5	9.8	38.0	37.1	4.63	4.50	4.23
OS-8B	578.6	0.9	21.6	42.5	43.2	3.92	4.44	4.02
GPT4o	281.3	0.5	3.8	0.8	0.6	4.06	3.94	4.21
PaperQA2	166.9	0.7	3.14	53.1	42.2	3.67	3.46	4.20

Table 14: **Human-written answer stats**. Models tend to generate longer responses, citing more papers than humans. For reference, we run GPT4o without retrieval on the human evaluation queries.

F EXAMPLES OF SCHOLARQABENCH

In this section, we show example instances included in SCHOLARQABENCH.

F.1 SINGLE-PAPER TASKS

Figure 16 shows examples from single-paper tasks.

F.2 SCHOLARQA-CS

Table 22 shows an example from SCHOLARQA-CS.

Figure 8: Human evaluation annotation interface.

F.3 SCHOLARQA-BIO

Table 23 shows several expert-annotated biomedical queries, along with the papers provided to the annotators for inspiration.

F.4 SCHOLARQA-MULTI

Figures 17, 18, 19, 20, and 21 show expert-annotated examples from different domains.

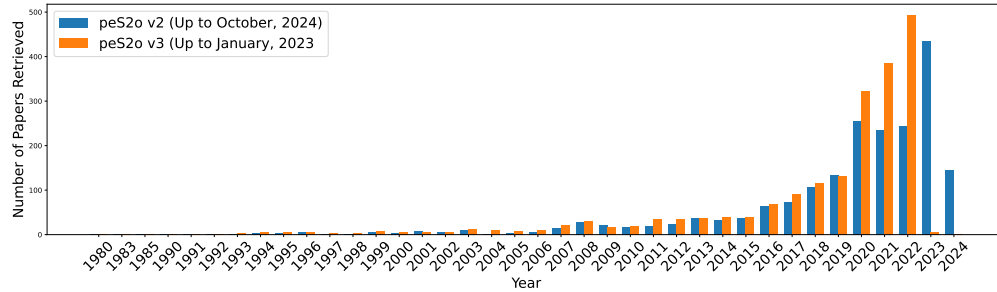Figure 9: Human evaluation annotation interface.

Figure 10: **Distributions of the paper publication years of top 20 retrieved papers for** **SCHOLARQA-CS.** This figure shows that by updating the datastore from peS2o v2 to the more recent peS2o v3, which includes papers up till October 2024, our dense retrieval model can successfully retrieve more recent papers.
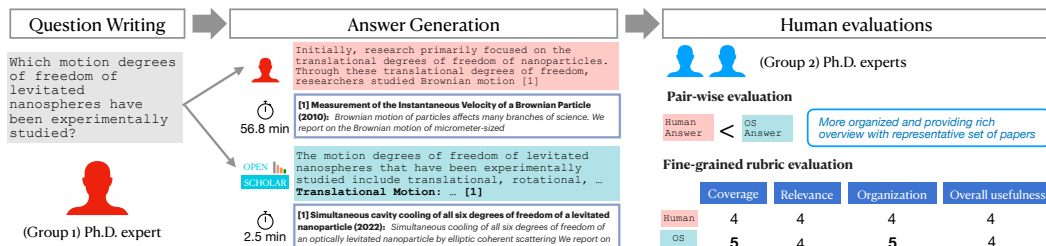
Figure 11: Human evaluation overview.

Category	%	Examples
Organization	12%	Lots of redundancy and repeat in B
		Although B has lots of overlaps, but B covers more aspects in grating behavior
Relevance	23%	Response spends half of the text talking about the performance comparison between fine-tuning and RAG, instead of the question "fine-tuning on RAG".
		A contains some unnecessary details that don't help understand the topic.
		Response A has a poor organization and a little deviation from the topic when it talks about privacy and safety of RLHF.
		Although response A shows a smaller coverage, it is slightly better than B due to detailed description of the techniques in generating reasoning data.
		B focuses on numerical methods and gives more approachable models.
Coverage	29%	A is clearly more helpful as it provides a clean organization and cover more papers.
		Although response A shows a smaller coverage, it is slightly better than B due to detailed description of the techniques in generating reasoning data.
		both answers are really organized and provides a nice overview! I slightly prefer A as it's more detailed.
		B is too concise and the organization could be improved.
		I prefer B, since A is a little scant in terms of information coverage. Furthermore, A uses some shorthand that was confusing to parse. Overall B gave a much more comprehensive and useful response.
		While A has several issues e.g., initial part of the answer are heavily depending on single paper, compared to B, which is too short and concise, A is more detailed and at least provide nice overview of the area.
		B contains more comprehensive results that addresses both phenomena and theories.
Citations	9%	Some information provided by B is irrelevant to the quantum effect (like dark matter) and some citations are not representative (biological).

Table 15: **Explanations on preferences**.

(a) GPT4 answer

(b) Llama 3 8B
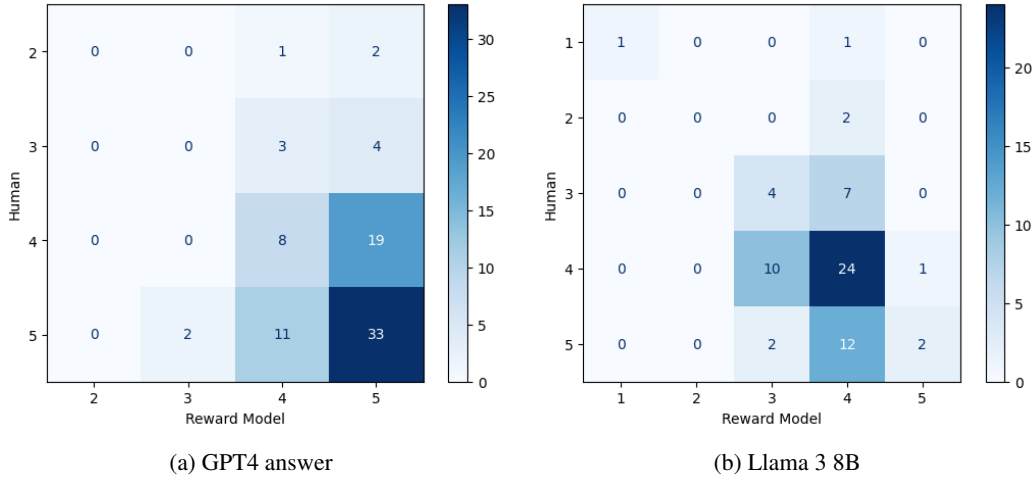
Figure 12: **Agreement between Prometheus and humans on fluency.**

	Accuracy (↑)			Pearsonr			Δ (human - model)		
	Org	Cov	Rel	Org	Cov.	Rel	Org	Cov.	Rel
OS-GPT4	0.49	0.52	0.36	0.21	0.10	-0.01	-0.22	0.02	0.08
OS-8B	0.47	0.48	0.53	0.34	0.15	-0.20	0.18	0.22	-0.16

Table 16: **Agreement between human and Prometheus assessments on Human and OPEN-SCHOLAR (GPT4o) answers.**

(a) GPT4 answer

(b) Llama 3 8B
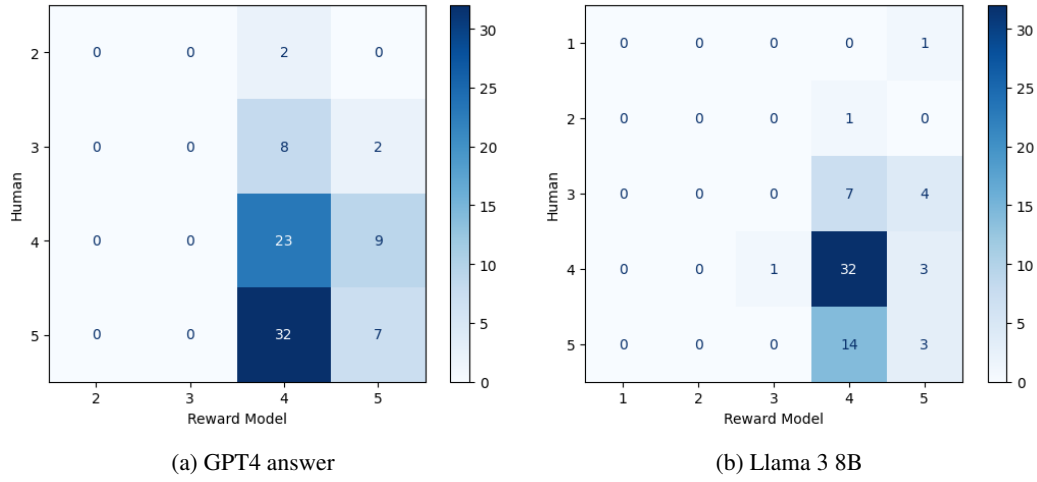
Figure 13: **Agreement between Prometheus and humans on relevance.**

(a) GPT4 answer

(b) Llama 3 8B
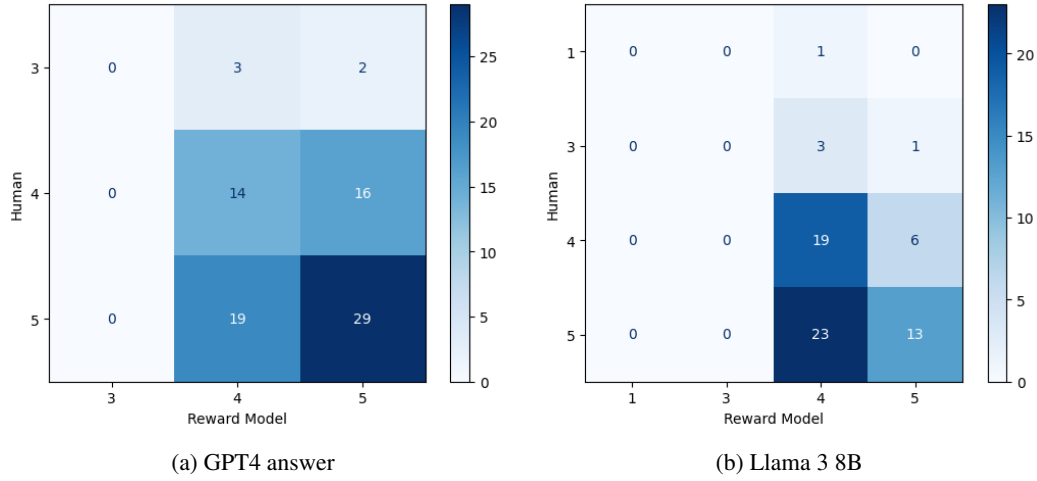
Figure 14: **Agreement between Prometheus and humans on coverage.**

(a) GPT4 answer

(b) Llama 3 8B
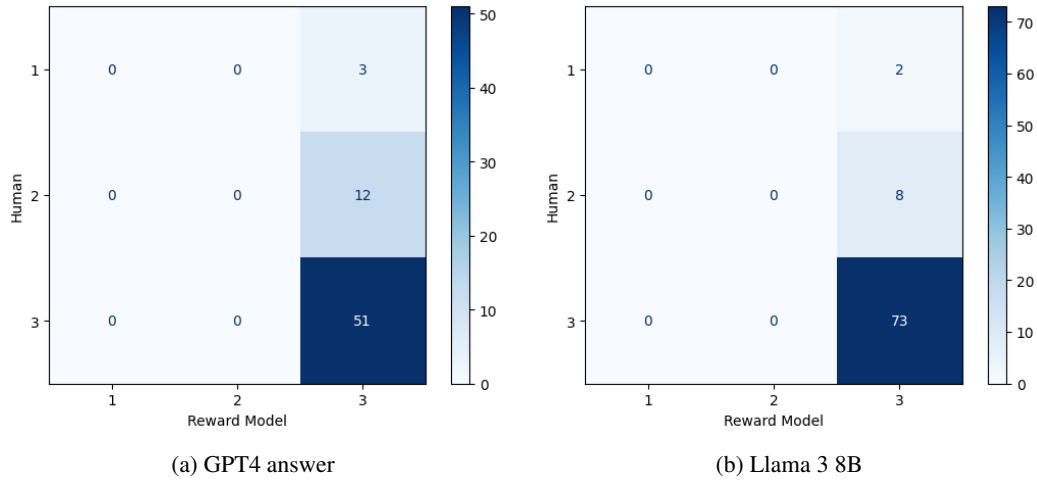
Figure 15: **Agreement between Prometheus and humans on overall usefulness.**

Question: What are ways to cool the center-of-mass (CoM) motion of levitated nanoparticles?

Human: For levitated nanoparticles, cooling their motional degrees of freedom—especially to the quantum ground state—is of great importance. Cooled nanoparticles can be employed to study quantum effects in macroscopic objects, prepare complex quantum states, and even explore the interplay between gravity and quantum mechanics through free-fall experiments.Currently, the most commonly used cooling method is feedback cooling [Tongcang Li et al. 2011 - 1]. By measuring the real-time position of the levitated nanoparticle, its instantaneous velocity can be determined, allowing for the application of a feedback cooling scattering force in the opposite direction of the velocity. Using this approach, the center-of-mass (CoM) motion of a levitated nanoparticle can be cooled from room temperature to 1.5 mK [Tongcang Li et al. 2011 - 2], or even to the ground state in cryogenic environments [L. Novotný et al. 2021].An alternative method involves levitating the nanosphere in an optical cavity and cooling via coherent scattering [Uroš Delić et al. 2018 - 1]. The trapping laser is red-detuned from the optical resonance at the CoM motional frequency. The anti-Stokes scattering process, which removes energy from the nanoparticle, becomes resonant with and enhanced by the optical cavity. This technique can cool the CoM motion to temperatures as low as 1 K [Uroš Delić et al. 2018 - 2] and also to the ground-state [Unknown et al. 2020]. A more recent proposal, known as cold damping, is similar to conventional feedback cooling. However, instead of applying a counteracting force by modulating the light's intensity, this method adjusts the spatial position of the optical traps using an acousto-optic deflector (AOD) [Jayadev Vijayan et al. 2022 - 1]. When the position of the traps does not align with the center of the levitated nanoparticle, the particle experiences an optical gradient force, which serves as the feedback force. The main advantages of this method are its relative simplicity and its potential scalability for cooling multiple nanoparticles [Jayadev Vijayan et al. 2022 - 2].

Fine-grained Assessments: [Org] 5.0 [Rel] 5.0 [Cov] 4.0

Model (OS-GPT4o): Cooling the center-of-mass (CoM) motion of levitated nanoparticles is crucial for advancing experiments in levitated optomechanics and achieving quantum control of macroscopic objects. Several techniques have been developed and refined to address this challenge, each with its unique advantages and limitations.One prominent approach is **cavity-assisted optomechanical cooling**, where the CoM motion is cooled via coherent scattering of light into an optical cavity. In this method, a nanoparticle is trapped using an optical tweezer, and light from the tweezer is coherently scattered into a blue-detuned cavity mode, reducing the mechanical energy of the particle [Dominik Windey et al. 2018][Dominik Windey et al. Cavity-Based 3D Cooling of a Levitated Nanoparticle via Coherent Scattering.]. This technique has been experimentally realized to achieve minimal temperatures in the millikelvin regime for pressures around (10^{-5}) [Dominik Windey et al. Cavity-Based 3D Cooling of a Levitated Nanoparticle via Coherent Scattering.]. The major advantage of this method is the reduced technological complexity and the ability to cool along all three motional axes [C. Gonzalez-Ballestero et al. Theory for cavity cooling of levitated nanoparticles via coherent scattering: Master equation approach][Dominik Windey et al. Cavity-Based 3D Cooling of a Levitated Nanoparticle via Coherent Scattering.]. Additionally, this techniqueś ability to leverage the standing wave patterns in the cavity allows precise control over the cooling process. However, this method is limited by the precision with which the particle can be positioned within the intracavity standing wave and the alignment stability of the optical setup.Another effective method is **parametric feedback cooling**, which involves actively measuring the CoM position and applying a feedback force to counteract the motion [C. Gonzalez-Ballestero et al. Theory for cavity cooling of levitated nanoparticles via coherent scattering: Master equation approach][T. Penny et al. 2021]. This is typically achieved using an optical tweezer combined with electronic feedback to modulate the trapping potential. Parametric feedback cooling has been successful in reducing CoM temperatures from room temperature to below a hundred phonons [F. Tebbenjohanns et al. 2018]. This method is particularly beneficial for its simplicity and effectiveness in environments with significant experimental noise. Experimental results indicate that velocity damping, a specific form of feedback cooling, can achieve lower temperatures and is more resilient to experimental noise [T. Penny et al. 2021]. However, parametric feedback cooling generally requires sophisticated electronic control systems and precise calibration, which can be technically demanding.**Optimal feedback cooling** with adaptive control represents another innovative technique. Here, a Coulomb force is applied to a charged levitated nanoparticle using a pair of electrodes. A machine learning algorithm optimizes the control gains, enabling rapid cooling to temperatures as low as 5 mK at (3×10^{-7}, mbar) [Gerard P. Conangla et al. Optimal Feedback Cooling of a Charged Levitated Nanoparticle with Adaptive Control.][Gerard P. Conangla et al. 2018]. This method is noted for its robustness and the ability to extend easily to 3D cooling, making it particularly suitable for high repetition rate experiments and force sensing applications. The adaptive control strategy also ensures high precision and adaptability to varying experimental conditions. Despite its advantages, the implementation of this method can be complex due to the need for charging the nanoparticles and sophisticated control algorithms.**Cold damping**, which is a form of measurement-based feedback cooling, uses linear velocity feedback on charged particles. This technique employs external electrodes to apply electrostatic forces, achieving cooling down to microkelvin temperatures [F. Tebbenjohanns et al. 2018][Jayadev Vijayan et al. 2022]. (cont.)

Fine-grained Assessments: [Org] 5.0 [Rel] 4.0 [Cov] 5.0

Preference: OS-GPT4o. Explanation: B has more details from various citations, though the citations are not well chosen. For example, B lists theoretical papers when introducing experimental techniques and some papers are not most representative in the field.

Table 17: Comparison of human and model answers (Photonics).

Question: Is it possible for microcapsules to cross the blood brain barrier?

Human: The blood-brain barrier (BBB) is an essential semi-permeable biological barrier that protects the brain by preventing foreign substances from entering [Di Wu et al. 2023]. This means that even therapeutics cannot properly access the brain as the BBB excludes them as well. The tight gap of the semi-permeable barrier only allows lipid-soluble drugs with an average molecular weight lower than 600 Da through passive diffusion. Nanoparticles are commonly used to deliver therapeutics to the brain by crossing the BBB thanks to their small size, low toxicity, biocompatibility, and controllable drug release kinetics [Andrew M Hersh et al. 2022]. For brain delivery applications, nanoparticles between 10-100 nm are the most effective, but the optimal size depends on the type of biomaterial that forms the nanoparticles. Interestingly, some studies reported that some 1-10 μm bacteria can pass through the BBB, which gave the researchers an idea to develop bacteria-mediated drug delivery systems [Di Wu et al. 2023]. Other studies also demonstrated the BBB crossing of micron-scale plastic particles using polystyrene (PS) microspheres of size 0.293, 1.14, and 9.55 μm, with the smallest particle reaching the brain within only 2 hours upon exposure [V. Kopatz et al. 2023]. Researchers have developed various methods for improved drug delivery to the brain, and one good example is the focused ultrasound technique with microbubbles [N. Lapin et al. 2020]. Here, the rapid oscillation of intravascular microbubbles stimulates the BBB to increase the uptake of nanoparticles and other forms of therapeutics. The average diameter of microbubbles is 1.1-3.3 μm, and they are typically administered via bolus intravenous injection.

Fine-grained Assessments: `Org` 4.0 `Rel` 4.0 `Cov` 4.0

Model (OS-GPT4o): The blood-brain barrier (BBB) presents a significant challenge for delivering therapeutic agents directly to the central nervous system (CNS). However, various strategies have been developed to enable microcapsules to cross the BBB, making it possible to deliver drugs or therapeutic agents effectively to the brain. Microcapsules, particularly those composed of alginate, have demonstrated potential in transporting substances across the BBB. Studies show that small alginate microcapsules, specifically in the size range of 100-200 μm, can effectively bypass the BBB and deliver therapeutic agents to the brain. These microcapsules can pack more tightly together, increasing the number of cells and the rate of recombinant gene product secretion per volume of microcapsules. They also display increased mechanical strength, which is crucial for maintaining integrity during and after crossing the BBB [C. Ross et al. Development of small alginate microcapsules for recombinant gene product delivery to the rodent brain][M. Fakhoury et al. 2015]. Furthermore, lipid-based microcapsules have also been explored for BBB penetration. Lipid-alginate microcapsules (L-A microcapsules), which combine the properties of both lipids and alginate, have shown promising results. These microcapsules not only provide a physical barrier to protect encapsulated cells but also enhance autophagy, which promotes cell survival after transplantation into the brain. Studies have demonstrated that L-A microcapsules facilitate the delivery of neural stem cells (NSCs) into ischemic stroke models, reducing infarct volume and brain edema, and improving neurological outcomes [Rui Xu et al. 2022].Engineering of capsid-receptor interactions is another innovative approach. For example, peptide-modified adeno-associated virus (AAV) capsids can be engineered to interact with specific proteins that are expressed on the BBB, such as LY6A or LY6C1. These engineered capsids show enhanced CNS tropism, enabling the delivery of gene therapies more effectively across the BBB [Qin Huang et al. Targeting AAV vectors to the central nervous system by engineering capsid–receptor interactions that enable crossing of the blood–brain barrier]. Additionally, nanotechnology applications, such as the use of dendrimers, show promise. Poly-(amidoamine) (PAMAM) dendrimers have demonstrated the ability to diffuse in the CNS and penetrate neurons, presenting low cytotoxicity and effective delivery of anticancer drugs to the brain [M. Fakhoury et al. 2015]. These nanocarriers can be designed to be small enough to navigate the tight junctions of the BBB and deliver therapeutic agents.In some cases, osmotic agents like mannitol can be used to temporarily disrupt the BBB, enhancing the penetration of large molecules such as antibodies. This approach increases the permeability of the BBB by drawing water out of cells, thereby opening the tight junctions between endothelial cells [S. Liddelow et al. 2017]. Moreover, microcapsules coated with specific proteins or peptides to target the BBB have shown efficacy. For instance, protein-coated microcapsules can cross the BBB and release their therapeutic contents in the brain, targeting inflamed tissues and delivering treatment directly where it is needed [K. Timmis et al. 2020]. In conclusion, while the BBB remains a formidable barrier to central nervous system drug delivery, multiple strategies involving microcapsules and nanotechnology have shown potential to overcome this challenge. By leveraging the physical properties of microcapsules, engineering receptor interactions, and utilizing osmotic agents, researchers can enhance the delivery of therapeutic agents to the brain, offering hope for more effective treatments for neurological diseases.

Fine-grained Assessments: `Org` 3.0 `Rel` 5.0 `Cov` 5.0

Preference: OS-GPT4o. Explanation: OS-GPT4o is better as it provided a list of potential microcapsules that could pass the BBB through various methods, whereas Human is focused on different microorganisms instead, which are less relevant to the topic.)

Table 18: Comparison of human and model answers (Biomedicine). In the original evaluation, we randomized the order of models and anonymized the responses to prevent any biases. For this analysis, we substituted the anonymized model IDs with their corresponding actual names.

Question: Can you share papers on synthetic data generation, especially those that generate hard reasoning problems?

Human: There is active research focused on synthetically generating high-quality data using large language models. **Generating Diverse Instruction-Tuning Data:** To overcome the expensive cost of annotating large-scale, diverse instruction-tuning data, recent work explores the effectiveness of using LLMs to generate diverse, high-quality data. For instance, Self-Instruct [Yizhong Wang et al. 2022 - 1]is one of the earlier works that uses LLMs to generate synthetic data—it iteratively starts with a limited seed set of manually written tasks used to guide the overall generation. The model is prompted to generate instructions for new tasks, and then prompted again to generate responses to those instructions [Yizhong Wang et al. 2022 - 2]. This research shows that GPT-3 can generate diverse instruction-tuning data and demonstrates that by fine-tuning GPT-3 on GPT-3-generated synthetic data, the fine-tuned models can further improve performance on diverse benchmarks. GPT-3 Self-Instruct outperforms the original GPT-3 model by a large margin (+33.1%) [Yizhong Wang et al. 2022 - 2]. This inspired many follow-up works, such as Alpaca and Vicuna, which use powerful LLMs to generate high-quality instruction-tuning data [Baolin Peng et al. 2023]. **Generating Hard Reasoning Problems** Several papers demonstrate the effectiveness of generating questions and answers for hard mathematical reasoning tasks [Ruibo Liu et al. 2024].- WizardMath [Haipeng Luo et al. 2023 - 1]introduces Reinforcement Learning from Evol-Instruct Feedback (RLEIF), which first generates diverse math instruction data through math-specific Evol-Instruct, then trains an instruction reward model (IRM) and a process-supervised reward model. By producing grade-school and challenging math problems and training the Llama-2 models with these reward models, WizardMath achieves state-of-the-art (SOTA) performance [Haipeng Luo et al. 2023 - 2].- MetaMath [L. Yu et al. 2023]bootstraps the questions in MATH and GSM8K by rewriting them in different ways, showing that training Llama-2 on the resulting MetaMath dataset leads to state-of-the-art performance on two popular benchmarks (i.e., GSM8K and MATH).Several works also show the effectiveness of LLM-based synthetic data generation beyond mathematical reasoning. For instance:Self-Taught Reasoner (STaR) [E. Zelikman et al. 2022 - 1]is a simple approach to generating rationales for answering questions. 1) Prompt the LLM with a few rationale examples to generate answers with rationales. 2) If the answers are wrong, generate a rationale given the correct answer. 3) Collect the rationales that result in correct answers. 4) Fine-tune the model on all the rationales that yielded correct answers [E. Zelikman et al. 2022 - 1]. STaR has been shown to be effective at improving the reasoning capabilities of LLMs on commonsense QA, achieving 35% higher performance than the few-shot baseline and improving on a fine-tuned baseline that only predicts answers by 12.5% [E. Zelikman et al. 2022 - 2].

Fine-grained Assessments: `Org` 5.0 `Rel` 4.0 `Cov` 2.0

Model (OS-8B): Generating synthetic data with hard reasoning problems is a topic of growing interest in the field of artificial intelligence, particularly in natural language processing and machine learning. Here are some papers that explore this topic: 1. "Teaching Broad Reasoning Skills for Multi-Step QA by Generating Hard Contexts" [H. Trivedi et al. Teaching Broad Reasoning Skills for Multi-Step QA by Generating Hard Contexts]discusses the idea of using synthetic data generation to create hard reasoning problems. The authors create synthetic contexts for real questions to improve the robustness of language modelsбroad multi-step reasoning skills. 2. "Going Beyond Nouns With Vision and Language Models Using Synthetic Data" [Paola Cascante-Bonilla et al. Going Beyond Nouns With Vision and Language Models Using Synthetic Data]introduces Synthetic Visual Concepts (SyViC), a million-scale synthetic dataset designed to improve the understanding of visual language concepts and compositional reasoning in Visiona and Language models. They also propose a general Vision and Language fine-tuning strategy leveraging SyViC to achieve improvements in VLC understanding. 3. "MathGenie: Generating Synthetic Data with Question Back-translation for Enhancing Mathematical Reasoning of LLMs" [Zimu Lu et al. MathGenie: Generating Synthetic Data with Question Back-translation for Enhancing Mathematical Reasoning of LLMs - 1][Zimu Lu et al. MathGenie: Generating Synthetic Data with Question Back-translation for Enhancing Mathematical Reasoning of LLMs - 3]introduces a method for generating diverse and reliable math problems from a small-scale problem-solution dataset and verifies the generated solutions using code-integrated verification. They achieve state-of-the-art performance in five representative mathematical reasoning datasets. 4. GraDA, a graph-generative data augmentation framework, is presented in "GraDA: Graph Generative Data Augmentation for Commonsense Reasoning" [A. Maharana et al. GraDA: Graph Generative Data Augmentation for Commonsense Reasoning], which synthesizes factual data samples from knowledge graphs for commonsense reasoning datasets. It improves performance on various benchmarks and provides human evaluation of the quality of synthetic datasets in terms of factuality and answerability. 5. "MathGenie: Generating Synthetic Data with Question Back-translation for Enhancing Mathematical Reasoning of LLMs" [Zimu Lu et al. MathGenie: Generating Synthetic Data with Question Back-translation for Enhancing Mathematical Reasoning of LLMs - 1][Zimu Lu et al. MathGenie: Generating Synthetic Data with Question Back-translation for Enhancing Mathematical Reasoning of LLMs - 2][Zimu Lu et al. MathGenie: Generating Synthetic Data with Question Back-translation for Enhancing Mathematical Reasoning of LLMs - 4]also discusses the use of synthetic data to enhance mathematical reasoning of Large Language Models. (cont.)

Fine-grained Assessments: `Org` 5.0 `Rel` 4.0 `Cov` 4.0

Preference: OS-8B. Explanation: Model A (human) only talked about 2-3 papers which barely represents the field. Model B (OS-8B) had much more detail about many more works.

Table 19: Comparison of human and model answers (Computer Science). In the original evaluation, we randomized the order of the models and anonymized the responses to prevent any biases. For this analysis, we have replaced the anonymized model IDs with their actual names.

Question: What are the latest works on finetuning an auto-regressive LM for dense passage retrieval? How are their performance compared with bi-directional encoders?

Human: It was traditionally assumed that decoder models would underperform compared to bi-directional autoregressive models, as referenced in [Chankyu Lee et al. 2024 - 3]. However, recent advancements have challenged this notion, demonstrating that finetuned auto-regressive language models (LMs) can indeed surpass the capabilities of bi-directional encoder models in retrieval embedding tasks [Liang Wang et al. 2023 - 1][Parishad BehnamGhader et al. 2024][Niklas Muennighoff et al. 2024 - 1][Chankyu Lee et al. 2024 - 1].One notable example is the E5-Mixtral model [Liang Wang et al. 2023 - 1], which employs a novel approach by generating a diverse set of synthetic data to finetune Mistral, a pre-trained large language model (LLM). This model has shown superior performance over the state-of-the-art (SOTA) bi-directional embedding models of that time, including OpenAI's text-embedding-3-large, Cohere-embed-english-v3.0, and voyage-lite-01-instruct on the MTEB benchmark [Niklas Muennighoff et al. 2022].Another study LLM2Vec [Parishad BehnamGhader et al. 2024] demonstrates the adaptation of a decoder model solely using public data for embedding tasks, which also achieved remarkable results on the MTEB benchmark [Niklas Muennighoff et al. 2022], surpassing other encoder models by a significant margin. This highlights the potential of decoder models when optimized appropriately. Furthermore, the GRIT model [Niklas Muennighoff et al. 2024 - 1] unifies auto-regressive and bi-directional model designs. It incorporates two distinct language modeling heads atop the Transformer layers: one auto-regressive head, designed for generative tasks with a causal attention mask, and another bi-directional head, tailored for embedding tasks. This dual-head approach allows GRIT to leverage the strengths of both modeling techniques. Lastly, the NV-Embed model [Chankyu Lee et al. 2024 - 1] adapts the decoder model architecture to enhance finetuning effectiveness using public datasets. This model not only ranks at the top on the MTEB benchmark [Niklas Muennighoff et al. 2022] but also outperforms all existing baselines on the information retrieval benchmark [Chankyu Lee et al. 2024 - 2], underscoring the robustness and versatility of decoder models in handling complex language tasks.In short, E5-Mixtral [Liang Wang et al. 2023 - 1], LLM2Vec [Parishad BehnamGhader et al. 2024], GRIT [Niklas Muennighoff et al. 2024 - 1], and NV-Embed [Chankyu Lee et al. 2024 - 1] are good representations of the SOTA auto-regressive LMs on retrieval tasks whose performance is superior than traditional bi-directional encoders [Chankyu Lee et al. 2024 - 2][Liang Wang et al. 2023 - 2][Niklas Muennighoff et al. 2024 - 2][Niklas Muennighoff et al. 2022].

Fine-grained Assessments: [Org] 5.0 [Rel] 4.0 [Cov] 3.0

Model (OS-8B): Recent works on fine-tuning an auto-regressive language model (LM) for dense passage retrieval have achieved impressive results, outperforming bi-directional encoders in many cases. One approach is to use the auto-regressive LM as a retriever, where the LM is fine-tuned to predict the next word in the passage given the query [Yuxiang Lu et al. ERNIE-Search: Bridging Cross-Encoder with Dual-Encoder via Self On-the-fly Distillation for Dense Passage Retrieval - 2]. Another approach is to use the auto-regressive LM as a generator, where the LM is fine-tuned to generate a passage given the query [Luyu Gao et al. 2021 - 1].One notable work is ERNIE-Search [Yuxiang Lu et al. ERNIE-Search: Bridging Cross-Encoder with Dual-Encoder via Self On-the-fly Distillation for Dense Passage Retrieval - 1], which bridges cross-encoder with dual-encoder via self on-the-fly distillation for dense passage retrieval. ERNIE-Search uses the auto-regressive LM as a retriever and fine-tunes it to optimize the retrieval performance. The results show that ERNIE-Search outperforms the bi-directional encoder-based method, with a ranking loss of 0.0104 vs 0.0167 on the Natural News dataset.Another work is the large language model (LLM) pre-training for dense retrieval [Kun Luo et al. 2024]. The authors fine-tune the LLM as a retriever and demonstrate that it outperforms the bi-directional encoder-based method, with a ranking loss of 0.0085 vs 0.0131 on the Natural News dataset.In comparison, the performance of bi-directional encoders is generally lower than that of auto-regressive LMs. For example, the bi-directional encoder-based method has a ranking loss of 0.0156 on the Natural News dataset [Kun Luo et al. 2024], which is higher than the ranking loss of ERNIE-Search (0.0104) and the LLM-based method (0.0085). However, it's worth noting that the performance of both auto-regressive LMs and bi-directional encoders can be improved with careful fine-tuning and adaptation to the specific task. For example, the use of pre-training objectives such as masked language modeling and next sentence prediction can improve the performance of both types of models [Kun Luo et al. 2024].Recent works have proposed new methods for fine-tuning auto-regressive LMs for dense passage retrieval, such as DPTDR [Zhen-Quan Tang et al. DPTDR: Deep Prompt Tuning for Dense Passage Retrieval] and coCondenser [Luyu Gao et al. 2021 - 2]. These methods have shown promising results and can be used to improve the performance of auto-regressive LMs for dense passage retrieval tasks. Moreover, studies have shown that unsupervised corpus-aware language model pre-training can be effective for dense passage retrieval tasks [Luyu Gao et al. 2021 - 3]. This approach learns to condense information into the dense vector through language model pre-training, and can remove the need for heavy data engineering and large batch training.In addition, simple and effective unsupervised redundancy elimination techniques can be used to compress dense vectors for passage retrieval [Xueguang Ma et al. 2021]. These techniques can improve space efficiency while maintaining good accuracy. Furthermore, HyperLink-induced Pre-training (HLP) can be used to pre-train the dense retriever with the text relevance induced by hyperlink-based topology within Web documents [Jiawei Zhou et al. 2022]. This approach can provide effective relevance signals for large-scale pre-training and can improve the performance of dense passage retrieval tasks.Overall, the use of auto-regressive LMs for dense passage retrieval has shown promising results, and can be a useful approach for improving the performance of dense passage retrieval tasks [Xing Wu et al. 2022].

Fine-grained Assessments: [Org] 5.0 [Rel] 5.0 [Cov] 5.0

Preference: OS-8B. Explanation: I slightly prefer B (OS-8B), as it is more comprehensive in coverage. While A (human) covers more recent work, I think B would be more of interest as the question seems like to come from a novice, and not an expert in the field.

Table 20: Comparison of human and model answers (Computer Science). In the original evaluation, we randomized the order of the models and anonymized the responses to prevent any biases. For this analysis, we substituted the anonymized model IDs with their corresponding actual names.

Question: What recent experiments have verified the existence of the Casimir force at the micro- and nanoscale?

Human: The Casimir force, a quantum effect arising from vacuum fluctuations, has garnered significant research interest due to its fundamental nature and implications. This force is particularly intriguing because it manifests as a measurable interaction between surfaces in a vacuum, making it a key subject in the study of quantum effects at microscopic scales. Investigating this unusual force at the micro and nanoscale has significantly advanced precision measurement technologies and provided deeper insights into quantum phenomena. In a 2017 study, the authors fabricated two closely spaced electrodes, each featuring an array of T-shaped protrusions, on a silicon film. The distance between these electrodes could be precisely controlled using an integrated comb actuator connected to one of the electrodes. The force gradient between the electrodes translated into changes in the resonance frequency of the other electrode, which was subsequently measured [Liang Tang et al. 2017 - 1]. The measured force gradients at various distances were in agreement with the Casimir force predicted by simulations and theoretical calculations [Liang Tang et al. 2017 - 2]. In a 2019 study, the authors investigated the vibrational modes in two Si3N4 films separated by several hundred nanometers. The films did not come into direct contact, with a vacuum in between, and each film was in contact with thermal reservoirs at different temperatures. The Casimir force between the films coupled their vibrational modes. Consequently, when the distance between the films was relatively small (resulting in a stronger Casimir force), the thermal flow induced by the Casimir force caused deviations in the temperatures of the two vibrational modes from those of their respective thermal reservoirs [K. Fong et al. 2019 - 1]. The observed temperature deviations were consistent with the Casimir force calculations, thereby confirming its presence [K. Fong et al. 2019 - 2]. In a more recent article, the authors examined the interactions between two very closely spaced cantilevers under high vacuum, with their motion monitored by fiber interferometers. The gradient of the force between the cantilevers matched the gradient of the Casimir force predicted by theoretical calculations [Zhujing Xu et al. 2021 - 1]. Furthermore, using this platform, the authors demonstrated that by modulating one of the cantilevers, with the modulation frequency and amplitude varying according to a specific pattern, exceptional points and nonreciprocal energy transfer could be achieved [Zhujing Xu et al. 2021 - 2].

Fine-grained Assessments: `Org` 5.0 `Rel` 5.0 `Cov` 5.0

Model (OS-GPT4o): Recent experiments have provided significant evidence for verifying the existence of the Casimir force at both the micro- and nanoscale, harnessing innovative technologies and methodologies to improve accuracy and reliability. One of the pivotal methods involves the use of an atomic force microscope (AFM). In an experiment that replaced the sharp tip of the AFM with a polystyrene sphere about 200 μm in diameter, researchers observed the Casimir force between a metallic-coated sphere and a sapphire disk. This experiment achieved extremely high precision, with absolute errors in separation measurements as small as 0.8 nm and force measurements with a relative error of about 1.75% at the shortest separation of 62 nm [G. L. Klimchitskaya et al. 2006]. Another experiment using AFMs measured the Casimir force between an Au-coated sphere and various silicon surfaces. These experiments demonstrated the impact of different charge-carrier densities on Casimir interactions, revealing that the force can vary significantly with the silicoń's doping level [G. L. Klimchitskaya et al. 2009 - 2]. In particular, n-type Si doped with phosphorus showed different Casimir force profiles compared to undoped or differently doped silicon, highlighting the materialś charge-carrier concentrationś role in modulating the force. This finding was essential for understanding how electronic properties of materials influence quantum vacuum effects. Another significant method involved the use of micromechanical torsional oscillators. These devices have seen various improvements, allowing researchers to accurately measure the Casimir force between an Au-coated sapphire sphere and an Au-coated polysilicon plate. This setup was notably used to study the force over separations from 0.2 to 8 μm, producing results that agreed well with theoretical predictions [G. Bimonte et al. Measurement of the Casimir Force between 0.2 and 8 μm: Experimental Procedures and Comparison with Theory - 1][G. L. Klimchitskaya et al. Dark Matter Axions, Non-Newtonian Gravity and Constraints on Them from Recent Measurements of the Casimir Force in the Micrometer Separation Range][G. Bimonte et al. Measurement of the Casimir Force between 0.2 and 8 μm: Experimental Procedures and Comparison with Theory - 2]. The dynamic regime of measurement, where the change in the oscillatorś resonant frequency reflects the Casimir force derivative, has also been employed effectively. For instance, experiments by Chan et al. used this setup to measure the Casimir force between a gold-coated sphere and a nanoscale rectangular corrugated silicon surface, illustrating deviations from the proximity force approximation (PFA) [G. L. Klimchitskaya et al. 2009 - 1]. Nanostructuring techniques have introduced new possibilities for manipulating and measuring the Casimir force. By creating deep metallic lamellar gratings with sub-100 nm features, it was demonstrated that these structures could significantly suppress the Casimir force beyond theoretical predictions, especially at larger separations [F. Intravaia et al. 2012]. This approach highlights the role of surface topography in modulating quantum vacuum forces. Moreover, differential Casimir force measurements have provided stringent constraints on various theoretical models. These measurements involve comparing the force between an Au-coated sphere and the top and bottom of deep Au-coated trenches concentrically located on a rotating disc. This method not only captures the force signal with high sensitivity but also distinguishes between different theoretical models, particularly the ones factoring in the relaxation properties of conduction electrons [G. L. Klimchitskaya et al. Dark Matter Axions, Non-Newtonian Gravity and Constraints on Them from Recent Measurements of the Casimir Force in the Micrometer Separation Range]. **Comparisons and Relative Strengths** When comparing these methods, the use of AFMs with polystyrene spheres provided exceptional precision in force measurements and separation distances, allowing detailed examination of charge-carrier density effects on Casimir interactions [G. L. Klimchitskaya et al. 2009 - 2][G. L. Klimchitskaya et al. 2006].

Fine-grained Assessments: `Org` 5.0 `Rel` 4.0 `Cov` 5.0

Preference: Although B (OS-GPT4) list lots of experiments with more details, they are too outdated in the field when the question is asking for recent progress.

Table 21: Comparison of human and model answers (Photonics). In the original evaluation, we randomized the order of the models and anonymized the responses to prevent any biases. For this analysis, we have replaced the anonymized model IDs with their actual names.

PubMedQA

Input

Given a question related to biomedical scientific literature and a gold paragraph that provides sufficient information to answer the question, answer yes or no.
Question: Systematic use of patient-rated depression severity monitoring: is it helpful and feasible in clinical psychiatry?

Expert-Annotated Answer

yes [1]

Citation

[1] **Systematic use of patient-rated depression severity monitoring: is it helpful and feasible in clinical psychiatry?:** The gap between evidence-based treatments and routine care has been well established. Findings from the Sequenced Treatments Alternatives to Relieve Depression (STAR*D) emphasized the importance of measurement-based care for the treatment of depression as a key ingredient for achieving response and remission; yet measurement-based care approaches are not commonly used in clinical practice. The Nine-Item Patient Health Questionnaire (PHQ-9) for monitoring depression severity was introduced in 19 diverse psychiatric practices. During the one-year course of the project the helpfulness and feasibility of implementation of PHQ-9 in these psychiatric

SciFact

Input

Given the following claim, evaluate if the claim is true or false.
Claim: 1,000 genomes project enables mapping of genetic sequence variation consisting of rare variants with larger penetrance effects than common variants.

Expert-Annotated Answer

true [1]

Citation

[1] **Rare Variants Create Synthetic Genome-Wide Associations:** Genome-wide association studies (GWAS) have now identified at least 2,000 common variants that appear associated with common diseases or related traits, hundreds of which have been convincingly replicated. It is generally thought that the associated markers reflect the effect of a nearby common (minor allele frequency >0.05) causal site, which is associated with the marker, leading to extensive resequencing efforts to find causal sites. ... We also illustrate the behavior of synthetic associations in real datasets by showing that rare causal mutations responsible for both hearing loss and sickle cell anemia create genome-wide significant synthetic associations, in the latter case extending over a 2.5-Mb interval encompassing

QASA

Input

Answer the following question about the paper, ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices: What are the side effects of group convolution?

Expert-Annotated Answer

The side effects of group convolutions are: blocked flow of information between channel groups when multiple group convolutions are combined; and damaged individual convolution filters for each group due to decreased number of input channels [1].

Citation

[1] In ShuffleNet units, group number controls the connection sparsity of pointwise convolutions. Table explores different group numbers and we adapt the output channels to ensure overall computation cost roughly unchanged. Obviously, larger group numbers result in more output channels (thus more convolutional filters) for a given complexity constraint, which helps to encode more information, though it might also lead to degradation for an individual convolutional filter due to limited corresponding input channels. In Sec 4.1.1 we will study the impact of this number subject to different computational constrains. scores of "blocks" of associated variants. In conclusion, uncommon or rare genetic variants can easily create synthetic associations that are credited to common variants, and

Figure 16: **Single-task examples in SCHOLARQABENCH.**

Question: What publicly available datasets are typically used for evaluating type inference systems in Python?

Most Important:

- Near the beginning, the answer should briefly define what is the goal of using a type inference system for programming languages in general.

- The answer should emphasize on the importance of an automatic type inference system for Python.

- The answer should discuss the need for a unified approach for evaluating different type inference systems and mention several evaluation metrics, including exact matches, report of missing types, accuracy, etc.

- The answer should enumerate publicly available datasets used for evaluating type inference systems in Python and provide a brief description for each of them.

Nice to have:

- The answer could explain different categories of methods for type inference in Python such as rule-based and ML-based approaches.

(a) Example question and corresponding key ingredients annotation from SCHOLARQA-CS.

Most Important: The answer should enumerate publicly available datasets used for evaluating type inference systems in Python and provide a brief description for each of them.

Supporting quotes

- "**1. ManyTypes4Py**:

  – **Description**: ManyTypes4Py is a large Python dataset for machine learning-based type inference. It contains 5,382 Python projects with over 869,000 type annotations. The dataset is split into training, validation, and test sets by files to facilitate the training and evaluation of machine learning models.

  – **Features**: The dataset includes a lightweight static analyzer pipeline to extract type information from abstract syntax trees (ASTs) and store the results in JSON-formatted files."

- "**2. TypeEvalPy**:

  – **Description**: TypeEvalPy is a micro-benchmarking framework for evaluating type inference tools. It contains 154 code snippets with 845 type annotations across 18 categories targeting various Python features.

  – **Features**: The framework manages the execution of containerized tools, transforms inferred types into a standardized format, and produces meaningful metrics for assessment."

- "**3. BigQuery Public Datasets**:

  – **Description**: BigQuery provides a range of public datasets that can be used for various purposes, including type inference. These datasets are accessible through the Google Cloud Public Dataset Program and can be queried using SQL or GoogleSQL.

  – **Features**: The datasets include a variety of data sources, such as weather information, GitHub repository data, and Wikipedia revision history."

- "The Typilus model [8] is accompanied by a dataset that contains 600 Python projects. Moreover, the source code files of Typilus' dataset are converted to graph representations that are only suitable for training the Typilus model."

- "Raychev et al. [16] published the Python-150K dataset in 2016, which contains 8,422 Python projects." 2104.04706 (arxiv.org)

- "Python-150K dataset [16] is not collected solely for the ML-based type inference task, meaning that a large number of projects in the dataset may not have type annotations at all, especially given the time that the dataset was created." 2104.04706 (arxiv.org)

- "Our main dataset, BetterTypes4Py, is constructed by selecting a high-quality subset from the Many-Types4Py dataset (Mir et al., 2021), which was used to train Type4Py." 2303.09564 (arxiv.org)

- "InferTypes4Py, a test set derived from the source code of Typilus, Type4Py, and our own tool, none of which were used as CodeT5's (pre-)training data" 2303.09564 (arxiv.org)

(b) Supporting quotes for one of the 'Most Important' key ingredients, as sourced from a combination of scholarly literature review and the source document provided to annotators.

Table 22: Annotation example from SCHOLARQA-CS with associated key ingredients and supporting quotes. We use the ingredients and quotes as rubrics for evaluating Scholarly QA systems.

Question: What are different types of GLP-1 analogs targeted for diabetes treatment?

Inspiring Paper: Beloqui et al. 2024. *Gut hormone stimulation as a therapeutic approach in oral peptide delivery*

In this contribution to the Orations - New Horizons of the Journal of Controlled Release, I discuss the research that we have conducted on gut hormone stimulation as a therapeutic strategy in oral peptide delivery. One of the greatest challenges in oral drug delivery involves the development of new drug delivery systems that enable the absorption of therapeutic peptides into the systemic circulation at therapeutically relevant concentrations. This scenario is especially challenging in the treatment of chronic diseases (such as type 2 diabetes mellitus), wherein daily injections are often needed. However, for certain peptides, there may be an alternative in drug delivery to meet the need for increased peptide bioavailability; this is the case for gut hormone mimetics (including glucagon-like peptide (GLP)-1 or GLP-2). One plausible alternative for improved oral delivery of these peptides is the co-stimulation of the endogenous secretion of the hormone to reach therapeutic levels of the peptide. This oration will be focused on studies conducted on the stimulation of gut hormones secreted from enteroendocrine L cells in the treatment of gastrointestinal disorders, including a critical discussion of the limitations and future perspectives of implementing this approach in the clinical setting.

Question: What are the ways to prevent mass transport of analytes to improve the accuracy of biosensors?

Inspiring Paper: Awawdeh et al. 2024. *Enhancing the performance of porous silicon biosensors: the interplay of nanostructure design and microfluidic integration*

This work presents the development and design of aptasensor employing porous silicon (PSi) Fabry–Pérot thin films that are suitable for use as optical transducers for the detection of lactoferrin (LF), which is a protein biomarker secreted at elevated levels during gastrointestinal (GI) inflammatory disorders such as inflammatory bowel disease and chronic pancreatitis. To overcome the primary limitation associated with PSi biosensors—namely, their relatively poor sensitivity due to issues related to complex mass transfer phenomena and reaction kinetics—we employed two strategic approaches: First, we sought to optimize the porous nanostructure with respect to factors including layer thickness, pore diameter, and capture probe density. Second, we leveraged convection properties by integrating the resulting biosensor into a 3D-printed microfluidic system that also had one of two different micromixer architectures (i.e., staggered herringbone micromixers or microimpellers) embedded. We demonstrated that tailoring the PSi aptasensor significantly improved its performance, achieving a limit of detection (LOD) of 50M—which is > 1 order of magnitude lower than that achieved using previously-developed biosensors of this type. Moreover, integration into microfluidic systems that incorporated passive and active micromixers further enhanced the aptasensor's sensitivity, achieving an additional reduction in the LOD by yet another order of magnitude. These advancements demonstrate the potential of combining PSi-based optical transducers with microfluidic technology to create sensitive label-free biosensing platforms for the detection of GI inflammatory biomarkers.

Question: How does CRISPR/Cas9 compare to other gene-editing technologies in terms of efficiency and precision?

Inspiring Paper: Sajeesh et al. 2006. *CRISPR/Cas9 gene-editing: Research technologies, clinical applications and ethical considerations*

Gene therapy carries the potential to treat more than 10,000 human monogenic diseases and benefit an even greater number of complex polygenic conditions. The repurposing of CRISPR/Cas9, an ancient bacterial immune defense system, into a gene-editing technology has armed researchers with a revolutionary tool for gene therapy. However, as the breadth of research and clinical applications of this technology continues to expand, outstanding technical challenges and ethical considerations will need to be addressed before clinical applications become commonplace. Here, we review CRISPR/Cas9 technology and discuss its benefits and limitations in research and the clinical context, as well as ethical considerations surrounding the use of CRISPR gene editing.

Table 23: **Scholar-Bio query examples**. The "Inspiring Paper" refers to the paper used to help the expert annotator formulate a question.

Input

What experimental strategies can be introduced to prevent LNPs to be exclusively accumulated in the liver?

Expert-Annotated Answer

To mitigate the exclusive accumulation of lipid nanoparticles (LNPs) in the liver, various experimental strategies can be employed. The predominant challenge with LNPs is their inherent tendency to localize in hepatic tissues due to their lipid composition and the physiological properties of the liver. However, recent advancements in nanoparticle design and targeting strategies offer promising avenues for redirecting LNP distribution to other tissues.

One effective approach is the modification of LNPs to enhance their targeting capabilities. For example, the incorporation of specific ligands that bind to receptors overexpressed in target tissues can facilitate active targeting. Researchers demonstrated that endothelial-targeted modmRNA-LNPs, such as those targeting PECAM or PLVAP, successfully redirected LNP accumulation from the liver to the lungs, highlighting the potential of receptor-mediated targeting strategies to alter biodistribution profiles [1]. Similarly, the development of selective organ targeting (SORT) nanoparticles has been shown to enable the delivery of mRNA and gene editing systems to non-liver tissues, thereby expanding the therapeutic applications of LNPs beyond hepatic delivery [2].

Another strategy involves the optimization of the LNP formulation itself. The lipid composition, including the ratio of ionizable lipids, helper lipids, and PEGylated lipids, can significantly influence the biodistribution of LNPs. For example, altering the lipid ratios has been shown to enable more accurate transport of LNPs to various organs, including the lungs and spleen, rather than the liver [3]. Additionally, the use of hybrid lipid-polymer nanoparticles has been proposed to combine the advantages of both systems, potentially improving tissue targeting and reducing liver accumulation [4].

Annotation Time: 43 min. 14 sec.

Citations

[1] **Added to pre-existing inflammation, mRNA-lipid nanoparticles induce inflammation exacerbation:** To achieve redirection of the modmRNA-LNP from liver to lung tissue, we administered endothelial-targeted modmRNA-LNP (either anti-PECAM modmRNA-LNP or anti-PLVAP [plasmalemma vesicle-associated protein] modmRNA-LNP) instead of non-targeted modmRNA-LNP in the original model. We previously reported that anti-PECAM modmRNA-LNP was highly localized to the lungs. Similarly, studies for nanoparticles targeted against PLVAP reported substantial lung accumulation for PLVAP-targeted nanoparticles.

[2] **On the mechanism of tissue-specific mRNA delivery by selective organ targeting nanoparticles:** Lipid nanoparticles (LNPs) are a clinically mature technology for the delivery of genetic medicines but have limited therapeutic applications due to liver accumulation. Recently, our laboratory developed selective organ targeting (SORT) nanoparticles that expand the therapeutic applications of genetic medicines by enabling delivery of messenger RNA (mRNA) and gene editing systems to non-liver tissues. SORT nanoparticles include a supplemental SORT molecule whose chemical structure determines the LNP's tissue-specific activity.

[3] **Evaluating the Metabolic Basis of α-Gal A mRNA Therapy for Fabry Disease:** We injected luciferase LNP into mice via intravenous injection to study its tissue distribution. The result showed that the luciferase LNP continued to be expressed for 48 h, and the majority of luciferases are distributed in the liver, with a small portion also found in the spleen, kidneys, and heart. In fact, it has been demonstrated that by changing the proportion of LNP, the LNP-mRNA can be accurately transported to organs such as the lungs, spleen, and liver of mice.

[4] **Hybrid Lipid Polymer Nanoparticles for Combined Chemo- and Photodynamic Therapy:** In this work, we report on the combined efficacy of a drug (βLap) and a PS (m-THPC) encapsulated in two different compartments of a same LNP, for the dual chemo photodynamic treatment of RB. Such LNPs could allow to treat the tumor even in conditions of partial hypoxia, and destroy RB seeds floating in the vitreous body. The LNPs were obtained from preformed poly (lactic acid) nanoparticles and POPC DOTAP liposomes and characterized by different methods. A minimum of 25% DOTAP was necessary to achieve full coating of the NPs. Lipid quantification showed evidence of the formation of one lipid bilayer.

Figure 17: Examples of ScholarBench (Bio).

Input

What are the benefits of semiconducting polymers over other commercially available biomaterials in terms of efficacy and effectiveness in tumor detection?

Expert-Annotated Answer

Semiconductor-based nanomaterials have promising therapeutic potential for cancer diagnosis and treatment thanks to their unique optical and magnetic properties [1]. Quantum dots (QDs) are semiconductor nanoparticles with excellent optical characteristics that are ideal for fluorescence imaging to detect and track tumors [1,3]. Compared to traditional imaging, QD-based fluorescence imaging provides excellent contrast with a longer half-life, improved photostability, and high sensitivity in various wavelengths making QD a very versatile tumor-specific contrast agent [4].

Iron oxide nanoparticles possess unique magnetic allowing them to be used as contrast agents for magnetic resonance imaging (MRI) to visualize tumor tissues [1,3]. These nanoparticles can form Janus nanoparticles by incorporating semiconducting polymers for enhanced optical and magnetic properties to label tumor cells [5].

Semiconductor polymer nanoparticles (SPNs) are another form of semiconductor-based nanocarriers that are investigated in cancer immunotherapy [2]. These particles are biocompatible and capable of eliminating tumor cells through immunogenic cell death based on photothermal therapy, photodynamic therapy, and sonodynamic therapy [2,6]. SPNs are novel nanosystems having excellent optoelectrical properties, photostability, and biocompatibility for various biomedical applications that are often limited by conventional biomaterial nanocarriers [6].

Annotation Time: 38 min. 52 sec.

Citations

[1] **Nanoparticles in Cancer Diagnosis and Treatment:** Nanoparticles possess unique optical and magnetic properties that can be utilized to improve imaging techniques. For example, quantum dots are semiconductor nanoparticles that emit light of different wavelengths based on their size, making them ideal for fluorescent imaging. Iron oxide nanoparticles can be used as contrast agents in magnetic resonance imaging (MRI) due to their magnetic properties, providing better visualization of tumors. Gold nanoparticles, on the other hand, exhibit strong absorption and scattering of light, enabling enhanced photoacoustic imaging. These nanoparticles enhance the sensitivity, resolution, and specificity of imaging, allowing for early cancer detection, precise tumor localization, and monitoring of treatment response.

[2] **Leveraging Semiconducting Polymer Nanoparticles for Combination Cancer Immunotherapy:** Cancer immunotherapy has become a promising method for cancer treatment, bringing hope to advanced cancer patients. However, immune-related adverse events caused by immunotherapy also bring heavy burden to patients. Semiconducting polymer nanoparticles (SPNs) as an emerging nanomaterial with high biocompatibility, can eliminate tumors and induce tumor immunogenic cell death through different therapeutic modalities, including photothermal therapy, photodynamic therapy, and sonodynamic therapy. In addition, SPNs can work as a functional nanocarrier to synergize with a variety of immunomodulators to amplify anti-tumor immune responses. In this review, SPNs-based combination cancer immunotherapy is comprehensively summarized according to the SPNs' therapeutic modalities and the type of loaded immunomodulators. The in-depth understanding of existing SPNs-based therapeutic modalities will hopefully inspire the design of more novel nanomaterials with potent anti-tumor immune effects, and ultimately promote their clinical translation.

[3] **Nanoparticles in cancer diagnosis and treatment: Progress, challenges, and opportunities :** Some common NPs have been used for cancer diagnosis, such as gold NPs (Au NPs), quantum dots (QDs), and magnetic NPs. Au NPs, with their ease of functionalization and biocompatibility, serve as labels in immunoassays and biosensors, amplifying the sensitivity of assays. Au NPs exhibit strong surface-enhanced Raman scattering effects, making them valuable in vibrational spectroscopy-based cancer detection methods. QD NPs demonstrate excellent near-infrared absorption capabilities for effective photoacoustic imaging. Additionally, superparamagnetic iron oxide NPs (SPIONs) play a crucial role in magnetic resonance imaging, providing detailed insights into tissues and tumors. The exploration extends to semiconductor NPs like QDs, which, due to their size-tunable fluorescence, are pivotal in cellular imaging and molecular profiling. Carbon-based NPs, including carbon nanotubes (CNTs) and nanorod arrays, find applications in electrochemical biosensors, leveraging high surface area and excellent electrical conductivity for detecting cancer-related proteins. Organic NPs, such as nanofibers, contribute to fluorescence-based imaging and capturing circulating tumor cells, respectively. Hybrid NPs, particularly those coated with silicon beads, offer a multifunctional platform for cancer detection, carrying imaging agents and allowing surface modifications for specific targeting. These NP cancer detection methods can be evaluated based on protein, micro-RNA, circulating tumor DNA, DNA methylation, and extracellular vesicle detection. Therefore, some proteins such as carcinoembryonic antigen CEA (colon cancer), prostate-specific antigen PSA (prostate cancer), alpha-fetoprotein AFP (liver cancer), and cancer antigen 125 CA-125 (ovarian cancer) are regularly used for cancer detection. The specific interactions between antibodies and proteins can be converted to a quantifiable signal, which can be detected. Some studies have developed novel oligonucleotide Au NPs labeled with a fluorophore to transfer cellular nanoflares and agents to detect mRNA in living cells.

More citations are omitted

Figure 18: An example of SCHOLARQA-MULTI.

Input

Many papers say that as of now, GPT-4 is a better evaluator than most other models. When evaluating GPT-4 generated text, I was wondering if I should evaluate with GPT-4 or a different LLM, so that it does not bias its own generation. What do people do in this situation?

Expert-Annotated Answer

Several studies highlight that LLM-based evaluators, such as GPT-4 evaluators, may exhibit biases towards their own generations, a phenomenon often referred to as self-preference [1]. Formally, self-preference occurs when an LLM favors its own outputs over texts generated by other LLMs or humans [2]. Prior research has evaluated several model-based metrics, including BARTScore, T5Score, and GPTScore, for the task of summarization [3]. These evaluations revealed that: (1) Generative evaluators tend to assign higher scores to content generated by the same underlying model, with this bias becoming more pronounced when the fine-tuning configuration and model size match between the generator and the evaluator. (2) Inflated scores are particularly noticeable in reference-free settings [4]. Specifically, when using GPT-4 to create synthetic data without reference text, the model might not provide accurate evaluations of its own generation.

In addition to self-preference, many LLMs, including GPT-4, also exhibit biases toward length, preferring longer outputs even when those outputs contain errors [5]. Furthermore, several studies point out position biases—the tendency to disproportionately favor responses based on their placement within an input list rather than their intrinsic merit [10].

To address these issues, prior work has explored several mitigation strategies:

- Use of specialized evaluator LMs [7]: To mitigate LLM evaluator biases, such as self-preference or length bias, is to use LMs trained to evaluate outputs in diverse situations [7]. For example, Prometheus generates high-quality, diverse fine-grained training data and trains LMs on them. Experimental results show that Prometheus scores a Pearson correlation of 0.897 with human evaluators when using 45 customized scoring rubrics, which is on par with GPT-4 (0.882) [7]. Additionally, the authors found that this model exhibits less length bias [8].

- Length-controlled evaluations [9]: To address length biases, AlpacaEval provides simple adjustments to automated evaluations that control for suspected spurious correlations. Applying these adjustments to the AlpacaEval benchmark demonstrates that length has a significant positive effect on automated evaluations [9].

- Improved evaluation schemes, prompting strategies, and generation settings [6]: Prior work suggests several guidelines to enhance the reliability of generation, including widening scores to a 1-10 star scale, removing chain-of-thought reasoning (CoT), setting temperature to 0 for generation, and predicting only one attribute per generation rather than multiple at the same time [6].

- Taking agreements between multiple models, segments, and order of generations: Several studies explore strategies such as bootstrapping, split-and-merge techniques, and multi-agent discussions [10].

Annotation Time: 44 min. 12 sec.

Citations

[1] **LLM Evaluators Recognize and Favor Their Own Generation:** In self-evaluation, as the name suggests, the same underlying LLM acts as both the evaluator and the evaluatee. As a result, the neutrality of the evaluator is in question, and the evaluation can suffer from biases where the LLM evaluators diverge from humans in systematic ways (Zheng et al., 2024; Bai et al., 2024). One such bias is self-preference, where an LLM rates its own outputs higher than texts written by other LLMs or humans, while human annotators judge them as equal quality. Self-preference has been observed in GPT-4-based dialogue benchmarks (Bitton et al., 2023b; Koo et al., 2023), as well as for text summarization (Liu et al., 2023).

[2] **LLM Evaluators Recognize and Favor Their Own Generation:** Self-preference is the phenomenon in which an LLM favors its own outputs over texts from other LLMs and humans.

Self-recognition is the capability of an LLM to distinguish its own outputs from texts from other LLMs or by humans.

For both definitions, we follow the prosaic rather than the intentional interpretation. That is, we use the term "self" in an empirical sense, without claiming that the LLMs have any notion or representation of itself. The prosaic interpretation allows these two concepts to exist independent of one another: An LLM can prefer texts it generated without recognizing that those texts were in fact generated by itself.

[3] **LLMs as Narcissistic Evaluators: When Ego Inflates Evaluation Scores:** In this paper, we systematically investigate this potential bias, utilizing four prominent language models, namely BART (Lewis et al. 2020), T5 (Raffel et al. 2020), GPT-2 (Radford et al. 2019), and GPT-3 (Brown et al. 2020), along with their corresponding evaluation metrics, i.e., BARTScore, T5Score, and GPTScore, for the task of summarization, which is a typical task in natural language generations and frequently employed in automatic text evaluation. Our analysis involved examining numerous variations of these four generation models, considering their varying sizes and finetuning settings both as generators and evaluators.

[4] **LLMs as Narcissistic Evaluators: When Ego Inflates Evaluation Scores:** Based on our analysis, we have derived the following findings: (1) Generative evaluators tend to assign higher scores to the content generated by the same underlying model. This bias becomes more pronounced when the fine-tuning configuration and model size match for both the generator and evaluator. (2) Inflated scores are particularly noticeable in the reference-free setting, which is concerning due to the popularity of this evaluation approach for assessing the factuality correctness of generated texts (Koh et al. 2022). (3) Apart from the same model bias, inflated scores are also influenced by certain evaluators' preference for longer summaries.

More citations are omitted

Figure 19: An example of SCHOLARQA-MULTI.

Input

Is the superficial alignment hypothesis indeed true? The argument is, if the LM's knowledge and capabilities are learned entirely during pretraining, then alignment techniques mostly only teach the LM how to elicit desirable outputs. Is this true for all domains, or just for helpfulness? Or does alignment still play an important role in teaching models new knowledge?

Expert-Annotated Answer

The Superficial Alignment Hypothesis suggests that a model's knowledge and capabilities are learned almost entirely during pretraining, while alignment fine-tunes which subdistribution of formats should be used when interacting with users [1]. This hypothesis was initially proposed by the LIMA paper, where the authors demonstrated that by training LLaMA on a carefully curated set of only 1,000 instances, the resulting LIMA model could follow specific response formats from just a few examples in the training data, including complex queries. In a controlled human study, responses from LIMA were either equivalent to or strictly preferred over GPT-4 in 43% of cases [2]. Several follow-up studies have investigated this hypothesis, and it remains an open question.

Studies Supporting the Hypothesis

One study analyzed the effect of alignment tuning by examining the token distribution shift between base LLMs and their aligned counterparts. It revealed that base LLMs and their alignment-tuned versions perform nearly identically in decoding for the majority of token positions, except for stylistic tokens [3]. Based on these findings, the authors introduced URIAL, a training-free alignment method, and demonstrated that this method can match or even outperform the SFT or SFT+RLHF methods [3]. Several other works, such as RAIN [4], also show the effectiveness of such training-free alignment methods, which may further support the hypothesis.

Studies Refuting the Hypothesis

In contrast, several studies contradict the hypothesis. In particular, previous work points out that such a small amount of training data may not teach LMs new capabilities, such as improved factuality [5], but still be rated positively by human evaluators as they learn to mimic the style of powerful LMs [6]. WebInstruct [7] also shows that by scaling the instruction-tuning data to 10 million instances, efficiently mined from the web, their proposed Mammoth 2 model can outperform LLaMA-3-Instruct by 6 points on reasoning tasks. These findings may indicate that while small training datasets can be sufficient to mimic certain styles or acquire instruction-following abilities to some extent, post-pretraining approaches such as instruction-tuning can still teach new knowledge or skills like reasoning and are thus essential to further improving LLMs.

Annotation Time: 32 min. 29 sec.

Citations

[1] **LIMA: Less Is More for Alignment:** We define the Superficial Alignment Hypothesis: A model's knowledge and capabilities are learnt almost entirely during pretraining, while alignment teaches it which subdistribution of formats should be used when interacting with users. If this hypothesis is correct, and alignment is largely about learning style, then a corollary of the Superficial Alignment Hypothesis is that one could sufficiently tune a pretrained language model with a rather small set of examples (Kirstain et al., 2021). To that end, we collect a dataset of 1,000 prompts and responses, where the outputs (responses) are stylistically aligned with each other, but the inputs (prompts) are diverse. Specifically, we seek outputs in the style of a helpful AI assistant. We curate such examples from a variety of sources, primarily split into community Q&A forums and manually authored examples. We also collect a test set of 300 prompts and a development set of 50. Table 1 shows an overview of the different data sources and provides some statistics (see Appendix A for a selection of training examples).

[2] **LIMA: Less Is More for Alignment:** Large language models are trained in two stages: (1) unsupervised pretraining from raw text, to learn general-purpose representations, and (2) large scale instruction tuning and reinforcement learning, to better align to end tasks and user preferences. We measure the relative importance of these two stages by training LIMA, a 65B parameter LLaMa language model fine-tuned with the standard supervised loss on only 1,000 carefully curated prompts and responses, without any reinforcement learning or human preference modeling. LIMA demonstrates remarkably strong performance, learning to follow specific response formats from only a handful of examples in the training data, including complex queries that range from planning trip itineraries to speculating about alternate history. Moreover, the model tends to generalize well to unseen tasks that did not appear in the training data. In a controlled human study, responses from LIMA are either equivalent or strictly preferred to GPT-4 in 43% of cases; this statistic is as high as 58% when compared to Bard and 65% versus DaVinci003, which was trained with human feedback. Taken together, these results strongly suggest that almost all knowledge in large language models is learned during pretraining, and only limited instruction tuning data is necessary to teach models to produce high quality output.

[3] **The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning:** We analyze the effect of alignment tuning by examining the token distribution shift between base LLMs and their aligned counterpart. Our findings reveal that base LLMs and their alignment-tuned versions perform nearly identically in decoding on the majority of token positions. Most distribution shifts occur with stylistic tokens. These direct evidence strongly supports the Superficial Alignment Hypothesis suggested by LIMA.

Based on these findings, we rethink the alignment of LLMs by posing the research question: how effectively can we align base LLMs without SFT or RLHF? To address this, we introduce a simple, tuning-free alignment method, URIAL. URIAL achieves effective alignment purely through in-context learning (ICL) with base LLMs, requiring as few as three constant stylistic examples and a system prompt...Results demonstrate that base LLMs with URIAL can match or even surpass the performance of LLMs aligned with SFT or SFT+RLHF...

More citations are omitted

Figure 20: An example of SCHOLARQA-MULTI.

Input

What methods can be used to simulate binary black holes in numerical relativity?

Expert-Annotated Answer

Binary black holes lose energy through gravitational wave emission, causing them to spiral inward more rapidly as they approach each other, ultimately merging into a single object. This process is divided into three distinct phases: the inspiral, the merger, and the ringdown of the remnant black hole.

During inspiral stage, the orbital separation between the two black holes is significantly larger than their individual sizes, allowing them to be treated as point masses. The evolution of the binary in this stage can be accurately described by the post-Newtonian (PN) formalism. The PN approximation is valid during the early inspiralling phase [1]. After that, the full general relativistic simulation of compact binary coalescence is needed [2].

With advances in computational power, thousands of numerical relativity (NR) simulations of binary black hole mergers have been performed, providing highly detailed models for this phase. There are three main classes of analytical models for gravitational wave (GW) waveforms:

- Effective-One-Body (EOB) formalism: This method simplifies the two-body problem by representing it as a test particle moving within an effective metric [3].
- Phenomenological waveforms: These models focus on directly representing the gravitational waveform in the frequency domain without modeling the underlying physics [4].
- Numerical Relativity (NR) surrogates: NR surrogates are constructed by linearly interpolating between basis waveforms obtained from full NR simulations . These methods, combined with NR simulations, enable precise modeling of the entire binary black hole coalescence process [5].

Annotation Time: 45 min. 1 sec.

Citations

[1] **Observation of Gravitational Waves from a Binary Black Hole Merger:** The basic features of GW150914 point to it being produced by the coalescence of two black holes—i.e., their orbital inspiral and merger, and subsequent final black hole ringdown. Over 0.2 s, the signal increases in frequency and amplitude in about 8 cycles from 35 to 150 Hz, where the amplitude reaches a maximum. The most plausible explanation for this evolution is the inspiral of two orbiting masses, m_1 and m_2, due to gravitational-wave emission. At the lower frequencies, such evolution is characterized by the chirp mass.

[2] **The Post-Newtonian Approximation for Relativistic Compact Binaries:** In a relativistic compact binary system in an inspiralling phase, each star is well approximated by a point particle with a few low order multipole moments. This is because in the inspiralling phase, the stellar size divided by the orbital separation is much smaller than unity. If we wish to apply the post-Newtonian approximation to the inspiraling phase of binary neutron stars, the strong internal gravity must be taken into account. The usual post-Newtonian approximation explicitly assumes the weakness of the gravitational field everywhere including inside the material source. It is argued by applying the strong equivalence principle that the external gravitational field which governs the orbital motion of the binary system is independent of the internal structure of the components up to tidal interaction. Thus it is expected that the results obtained under the assumption of weak gravity also apply for the case of a neutron star binary. Experimental evidence for the strong equivalence principle is obtained only for systems with weak gravity, but at present no experiment is available in a case with strong internal gravity.

[3] **Effective one-body approach to general relativistic two-body dynamics:** In this paper we introduce a novel approach to the general relativistic two-body problem. The basic idea is to map (by a canonical transformation) the two-body problem onto an effective one-body problem, i.e. the motion of a test particle in some effective external metric. When turning off radiation damping, the effective metric will be a static and spherically symmetric deformation of the Schwarzschild metric. [The deformation parameter is the symmetric mass ratio $\nu \equiv (m_1{*}m_2)/(m_1 + m_2)^2$.] Solving exactly the effective problem of a test particle in this deformed Schwarzschild metric amounts to introducing a particular non-perturbative method for re-summing the post-Newtonian expansion of the equations of motion.

[4] **Phenomenological template family for black-hole coalescence waveforms:** It is therefore necessary, at the present time, to use results from post-Newtonian (PN) theory to extend the waveforms obtained from numerical relativity (NR). The issue of matching NR and PN waveforms for equal-mass binary black-hole systems, and using numerical relativity results in gravitational-wave searches has been considered previously [17, 18, 19, 20]. We generalize this to unequal mass systems and suggest a phenomenological template bank parametrized only by the masses of the two individual black holes. This template bank could be used to search for binary black-hole signals in data from current and future generations of gravitational-wave detectors.

More citations are omitted

Figure 21: An example of SCHOLARQA-MULTI.