nature portfolio

https://doi.org/10.1038/s41586-025-10072-4

## Supplementary information

# Synthesizing scientific literature with retrieval-augmented language models

In the format provided by the authors and unedited

# Supplementary Methods

## 1 Released Artifacts

We release the following 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` |
| **Public Demo Interface** | `allenai/open-scholar-demo` |

## 2 Additional ScholarQABench Details

### 2.1 Goal of ScholarQABench

ScholarQABench satisfies two principles: it serves as a realistic benchmark for literature review, and as a reproducible, multifaceted evaluation pipeline.

**A realistic benchmark for literature review.** ScholarQABench integrates tasks from two sources: (i) curated and adapted existing literature review tasks that are annotated by scientists, and (ii) three new datasets from different scientific domains that reflect realistic literature review scenarios (e.g., requiring information synthesis from multiple papers) and are annotated by Ph.D. experts. ScholarQABench tasks encompass different output formats and disciplines.

For multi-paper tasks, we instruct our expert annotators to formulate information-seeking questions that they are genuinely interested in finding answers to, instead of questions they already know the answers to or that could be answered using small text chunks from a single paper (Choi et al., 2018; Asai & Choi, 2021). We found this approach crucial for curating realistic questions that scientists might ask in reality. These questions are typically more detailed, contextualized (e.g., "I am planning to generate and filter synthetic training data using GPT4o, but I'm concerned that GPT4o may be biased toward its own generations."), and require nuanced, long-form answers instead of binary or multiple-choice responses. We collected human-written answers or rubrics to ensure reliable evaluation, instead of relying on answers generated by state-of-the-art proprietary LMs, as in prior work (Xu et al., 2024; Malaviya et al., 2023). This is because proprietary models still face limitations such as hallucination due to lack of domain knowledge, biases, and outdated information, making them unsuitable for consistent evaluation with newer models. Additionally, using model-generated answers as references can unfairly favor models from the same family and introduce possible evaluation biases. To avoid these issues, we curated expert-written answers for ScholarQA-CS and ScholarQA-Multi.

**A reproducible, multifaceted evaluation pipeline.** Due to the low correlation between automated similarity-based metrics such as ROUGE (Xu et al., 2023; Malaviya et al., 2023) and human judgment, evaluations of long-form generation tasks in expert domains typically rely on small- to medium-scale expert annotations (Singhal et al., 2023; Si et al., 2024; Zheng et al., 2024). Human expert evaluation constitutes the gold standard, but human annotations can be expensive to obtain and difficult 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.

### 2.2 SINGLE-PAPER TASK DATA CURATION

**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 1000 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 answer a detailed question about the target paper sufficiently. 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.

### 2.3 MULTI-PAPER TASK DATA COLLECTION

While some existing datasets require retrieval or reasoning over single, pre-specified papers (Wadden et al., 2020), we found a lack of resources that support multi-paper synthesis and long-form generation. To address this gap, we curated a new set of datasets that require long-form generation across multiple disciplines, unlike prior datasets, which primarily focus on short-form generation or multiple-choice questions (Skarlinski et al., 2024; Phan et al., 2025). For our data curation, we recruited annotators who have completed or are currently pursuing a Ph.D.

**Recruiting annotators.** For data curation, we recruit expert annotators through UpWork and interinstitution 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.

**IRB and compensation.** We obtained IRB approval from the University of Washington STUDY00021379 (exempt research[1]), to conduct all human-facing evaluations. For benchmark creation, where annotators generate questions, reference answers, and/or rubrics, the time required varies by task. For CS and Multi, each instance takes on average 45 minutes to 1 hour. For Neuro and Bio, each instance takes on average 5 minutes. Similar to the expert evaluations, we ensure fair compensation across all subsets, paying annotators between USD 30–40 per hour.

**Statistics of human-written questions and answers in SCHOLARQA-MULTI.** Figure SI.1a shows the subject distribution, and Figure SI.1b 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.

**Annotation instructions.** Table SI.1 outlines our annotations instructions to collect rubrics for SCHOLARQA-CS, and Table SI.2 shows annotation instructions given to annotators for SCHOLARQA-BIO and SCHOLARQA-NEURO. Table SI.3 shows the instructions given to annotators for SCHOLARQA-MULTI.

#### 2.3.1 SCHOLARQA-CS EVALUATION

**Overview of score calculations.** 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,

[1] `https://www.washington.edu/research/hsd/guidance/exempt/`

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 between individual ingredients of the same type. Specifically, For the general measures, we considered four sub-aspects, length penalty, expertise, citation count, and citation excerpts, and assigned each a uniform weight of 0.1, totaling 0.4. The remaining weight (0.6) was allocated to the rubric-based evaluation. Within the rubric, each criterion was labeled as either "important" or "nice to have." To reflect their relative importance, we assigned twice the weight to each "important" criterion compared to "nice to have" ones, such that $2w \times$ #important $+ w \times$ #nice $= 0.6$. Each criterion's weight was then equally split between rewarding satisfaction of the overall criterion and the corresponding examples, if present.

This score emphasizes the presence of citations and excerpts, such that systems lacking either would underperform even when they score relatively high on the remaining criteria, as seen in our experiments.

**Evaluation instructions.** Table SI.4 presents the instructions given to the evaluator LM for determining whether a model-generated answer satisfies each evaluation criterion (rubric).

**Human agreement with rubric evaluations.** Accurately evaluating whether a given response satisfies a rubric is essential for computing rubric scores. To assess the reliability of the LLM judge, we conducted human evaluations. Specifically, we sampled 12 questions from the SCHOLARQABENCH-CS dataset and obtained answers from two systems for each question. Each question–answer pair was scored by two independent expert annotators with Ph.D. degrees, using the same instructions as the LLM judge (Table SI.4). The average agreement between the two human annotators on rubric items was 0.80, while the average agreement between a human annotator and the LLM judge was 0.79 (see Table SI.5 for detailed correlation results). These results show that the LLM judge is highly reliable in assessing rubric-item satisfaction, performing on par with human–human agreement, indicating that humans generally concur with its evaluations.

### 2.3.2 CONTENT QUALITY AND ORGANIZATION

Table SI.6 shows an overview of evaluation dimensions. We use relevance, coverage, and organization for automatic evaluation. Each aspect is evaluated on a five-point scale, based on detailed criteria. We applied the same evaluation scheme as for human evaluations. We found that existing evaluator LMs struggle with evaluating overall usefulness, and tend to be over-optimistic. Overall usefulness is additionally used in human evaluation.

**Evaluation instructions and rubrics.** We show annotator rubrics for organization, coverage, relevance and overall usefulness in Tables SI.7,SI.8, SI.9, and SI.10, 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 relevance, it sometimes gives scores that are much higher than human assessments, especially for GPT4o. Consequently, we use Prometheus BGB for organization and coverage, 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.

**Agreements between humans and Prometheus on SCHOLARQABENCH-Multi.** We examine the alignment between human assessments and LLM assessments for evaluating the fine-grained quality of responses.

Figures SI.2, SI.3, and SI.4 show the confusion matrices between human annotations and Prometheus predictions. Most disagreements occur between adjacent classes (e.g., 4 vs. 5), and the evaluator LMs rarely confuse negative classes (scores below 3) with positive classes.

Table SI.11 provides a more quantitative analysis of the agreement between the LLM judge and expert annotators for three models (GPT-4o, OS-8B, and OS-GPT-4o—based) on five-point scale assessments. For this analysis, we compute prediction accuracy on a collapsed three-point scale, merging scores 4 and 5 as positive, 3 as neutral, and 1 and 2 as negative (Artstein & Poesio, 2008).

Across aspects and models, the LLM judge generally predicts the correct category, with most confusion occurring between neighboring categories, particularly between 4 and 5.

The average score differences between the LLM judge and expert evaluators across the three evaluated baselines are 0.28, -0.09, and -0.16, respectively. While there is a small score gap, especially for more subjective, adjacent labels, the LLM judge correctly predicts the relative ranking of the three models for all aspects except for GPT-4o vs. OS-GPT-4o on Relevance, where expert evaluations also show a small gap. These results suggest that Prometheus-based evaluations are generally reliable.

#### 2.3.3 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 SI.12.

#### 2.3.4 COMPARISON WITH OTHER DATASETS

Table SI.13 provides comparison of SCHOLARQABENCH and other datasets, showing their example instances.

## 3 ADDITIONAL OPENSCHOLAR DETAILS

### 3.1 INFERENCE PIPELINE DETAILS

**Details of each inference pipeline.** During OPENSCHOLAR inference, there are three core components: (1) generating a response based on the input question and retrieved context (Response Generation), (2) providing feedback on the response (Feedback and Retrieval) and (3) incorporating it (Feedback Incorporation), and (4) verifying the citations in the final response (Citation Verification).

- (1) **Response generation:** We insert the retrieved references and the query immediately after the instructions and prompt the LM to generate an answer, following the standard RAG paradigm.
- (2) **Feedback and retrieval:** If the feedback indicates missing information in the original answer, we optionally generate a set of search queries and invoke the retrieval pipeline to identify additional relevant papers.
- (3) **Feedback incorporation:** The model is instructed to revise the response by incorporating the feedback and, if applicable, the newly retrieved documents from step (2), while preserving parts of the original answer that are unrelated to the feedback.
- (4) **Citation verification:** We first split the model response into paragraphs. For each paragraph, we prompt the model to verify the correctness of existing citations and insert any missing citations as needed.

For each response, we instruct the LM to mark the response boundaries using two special tokens: `[Response_Start]` and `[Response_End]`.

**Instructions.** Table SI.14, Table SI.15, Table SI.16, and Table SI.17 show instructions for (1) Response generation, (2) Feedback and retrieval, and (3) Feedback incorporation, and (4) Citation verification, respectively.

### 3.2 SCIENTIFIC BI-ENCODER $\theta_{\text{bi}}$ TRAINING

For $\theta_{\text{bi}}$, 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.

### 3.3 SCIENTIFIC CROSS-ENCODER $\theta_{\text{cross}}$ TRAINING

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 $\mathbf{P}$ to the generator LM. For $\theta_{\text{cross}}$, 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.

### 3.4 GENERATOR $\mathcal{G}$ TRAINING

**Training data statistics.** Figure SI.5 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 data used for evaluation, namely PubMedQA, SciFact, and QASA. In total, this leads to 130,135 training instances.

**Training hyperparameters.** For OS-8B, we initialize model checkpoints from Llama 3.1 8B Instruct (Dubey et al., 2024), respectively. We use `torchtune`[2] 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, a batch size of 1, a gradient accumulation step of 2, and in bf16. We use AdamW (Loshchilov & Hutter, 2019) as an optimizer. We use 8 A100 GPUs with 80 GB of memory for training.

## 4 ADDITIONAL EXPERIMENTAL SETTINGS

### 4.1 BASELINE MODEL DETAILS

**Parametric LM Baselines.** We prompt language models (LMs) to generate long-form responses that include citations, followed by a list of paper titles. The full instruction is provided in Table SI.18. Example outputs from the LLama3.1 8B and GPT4o are shown in Tables SI.19 and SI.20. We observe that the generated responses frequently include fabricated paper titles, leading to a high rate of hallucination.

**RAG Baselines.** For our OSDS RAG baselines, we retrieve the top 15 passages using our peS2o retriever without applying additional reranking, following standard practice in RAG settings. We then prompt language models to generate responses while citing the corresponding evidence from the retrieved context. Table SI.21 presents the exact instructions provided to the RAG baseline systems.

**PaperQA2.** We performed a good-faith replication of PaperQA2 using the public codebase[3] released by Skarlinski et al. (2024). We use OpenAI's `hybrid-text-embedding-3-large` as the embedding model and `gpt-4o-2024-08-06` as the LLM agent backbone, and employ a slightly longer response word length of 300 instead of 200, which improves end-task performance; otherwise, we follow default hyper-parameter configurations. One key difference is that PaperQA2 did not release their full retrieval corpus, which contains private or license-protected papers. We restricted our retrieval datastore to only open-access papers, and downloaded PDFs of papers suggested by our peS2o-based retrieval pipelines and the Semantic Scholar API, resulting in 2747 PDFs in total.

### 4.2 HUMAN EXPERT EVALUATION

We evaluated human expert performance on SCHOLARQABENCH-CS and Multi. To measure human expert performance on SCHOLARQABENCH-CS, we recruited two Ph.D. holders in Computer Science and one Ph.D. candidate to produce long-form expert-written answers. Experts were instructed to select questions from the SCHOLARQABENCH-CS that they felt aligned the best with their own expertise. In order to discourage annotators from choosing all "easy" questions, we segmented the questions into three bins: simple, complex, and highly complex; annotators were asked to choose questions from each level of complexity. Keeping with SCHOLARQABENCH-Multi expert answer writing set up, the answer writers were asked to write the report, ensuring that all citation-worthy sentences are supported with citations from publications. They were allowed to use LLM support

[2] `https://github.com/pytorch/torchtune`

[3] `https://github.com/Future-House/paper-qa`

for isolated sentence writing assistance, but otherwise, they were given explicit instructions not to resort to LLMs or research agents for response composition. The annotators were recruited via Upwork from a pool of Ph.D. annotators with strong track records. Each annotator was compensated at a pre-negotiated hourly pay rate ($28-35). In total, we collected 31 query–response pairs, with annotators spending an average of one hour per response.

# 5 Additional Results

## 5.1 peS2o v2 and v3 comparison

Figure SI.6 shows the distributions of the top retrieved papers using our retrieval systems on different datastores: peS2o v2 and v3.

In this section, we analyze the datastores that OpenScholar can use. Figure SI.6 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 updated datastore (peS2o v3), it is still able to retrieve relevant papers from the newer datastore at test time, resulting in many papers from 2023 to 2024.

## 5.2 Selecting embedding and retrieval models

Effective retrieval is critical for accurate and comprehensive scientific literature synthesis. To evaluate the performance of different retrieval and reranking models, we constructed a synthetic retrieval benchmark tailored to the scientific domain.

**Evaluation setup.** We generated the benchmark by prompting GPT4o to create scientific questions or claims based on randomly sampled paper abstracts from the OpenScholar datastore. Each generated query was then paired with its corresponding source abstract to form a set of query–gold document pairs. This setup results in a synthetic retrieval task where the goal is to retrieve the correct (gold) abstract from a large collection, given the generated query. We created a synthetic dataset with 843 queries and 100k randomly sampled abstracts from OSDS. We evaluate performance using standard information retrieval metrics NDCG@k, following the protocol used in widely adopted benchmarks such as BEIR (Thakur et al., 2021).

**Retrieval results.** Table SI.22 presents the results. The top section reports bi-encoder performance (i.e., without reranking). Our experiments demonstrate that the OpenScholar-Retriever (OS Retriever) significantly outperforms the base Contriever-base, showing the effectiveness of domain-focused continued pre-training. OpenScholar-Retriever also surpasses stronger baselines such as GIST (Solatorio, 2024) and BGE-base (Xiao et al., 2023), despite those models being trained on much larger retrieval datasets.

**Reranking results.** We also evaluate two reranking models on our synthetic benchmark: the base BGE reranker and our OpenScholar-Reranker (OS-Reranker). This analysis serves two purposes: (1) to assess the impact of incorporating a reranker on top of retrieval models, and (2) to evaluate the effectiveness of domain-specific fine-tuning. The bottom section of Table SI.22 shows cross-encoder reranker performance applied on top of the OpenScholar retriever. Adding OpenScholar-Reranker on top of OpenScholar-Retriever improves NDCG, surpassing the gain achieved by a state-of-the-art BGE reranker. These results demonstrate the value of domain-specific training for reranking.

## 5.3 Retrieval pipeline ablations

To further examine the contribution of each initial retrieval component—OpenScholar-Retriever–based retrieval (embedding-based), Semantic Scholar API retrieval, and web search—we run OpenScholar-8B with one retrieval component ablated at a time. In each setting, we apply OpenScholar-Reranker to the top passages retrieved by the remaining two pipelines, and then run inference on ScholarQABench-CS without additional feedback or verification.

Table SI.23 presents the results. We found that all ablations result in substantial Rub or Cite deterioration from the baseline using all retrieval pipelines (top row), indicating that all three components

contribute to producing more reliable and better-supported results. We find that removing the embedding-based pipeline (SCHOLARQABENCH-Retriever and DataStore) causes the largest performance drop. Removing either of the other two APIs also reduces performance, particularly citation accuracy.

### 5.4 QUALITATIVE MODEL OUTPUT COMPARISON

**Effectiveness of self-feedback on OS-8B.** In Table SI.24, we present a qualitative comparison between our standard RAG baseline ($RAG^{OSDS}$) and OPENSCHOLAR-8B. Both responses are generated using our `Llama-3.1_OpenScholar-8B`. The example illustrates that, with additional feedback, OPENSCHOLAR-8B produces a more detailed and better-organized response. In contrast, the original RAG output presents various approaches from individual papers in a less structured manner.

However, we also observed cases where additional feedback introduced repetitive or less relevant information, leading to a less structured overall organization. Table SI.25 shows such examples. In these cases, the first paragraph already discusses retrieval-augmented approaches, yet the response later repeats this content when presenting their results. It also briefly mentions vision–language models with caption rewrites for reducing hallucinations, which is not directly relevant to the original question. While the model can generate useful feedback—such as suggesting new methods or adding concrete experimental results—during the iteration and editing stages, the trained 8B model often struggles to integrate new information effectively, reorganize sections based on updated search results, or filter out unnecessary facts.

**Analysis on PaperQA2 response.** Table SI.26 shows PaperQA2 response. We found that the model tends to retrieve from a small set of papers, resulting in less cited papers as well as shorter responses than OPENSCHOLAR.

## 6 EXPERT EVALUATIONS

### 6.1 EVALUATION SETUP DETAILS

**Compensation and IRB.** We obtained IRB approval from the University of Washington STUDY0002137. Annotators are presented with a model response and an expert-written reference response for an information-seeking query in their domain of expertise. They are asked to (1) select their preferred response, (2) provide fine-grained evaluations across multiple aspects, and (3) give free-form explanations for their preference. We estimate that each instance takes approximately 6 minutes to complete, so annotators complete about 10 instances per hour. Accordingly, we pay USD 30 per 10 instances.

**Annotation interface.** Figures SI.7 and SI.8 illustrate the annotation interface used for expert evaluations. Annotators are first presented with two anonymized responses—one from a model and one from a human—as shown at the top of Figure SI.7. They then assess each response according to three fine-grained criteria, as well as overall usefulness (Figure SI.8, middle). Finally, the expert annotators select their preferred response (Figure SI.8, bottom).

### 6.2 EXPERT EVALUATION RESULTS

**Analyses on experts' explanations of pairwise preference.** We analyzed the explanations provided by expert annotators for their preferred responses. Table SI.27 summarizes the top four categories identified in these explanations. We found that expert annotators most frequently emphasized overall coverage and relevance when making their selections. Additionally, annotators often highlighted the importance of citations, noting cases where citations were incorrect or less relevant.

**Qualitative examples of model and human answers.** Tables SI.28, SI.29, SI.30, SI.31, SI.32 show qualitative examples of human versus model answers in Photonics, Biomedicine, and Computer Science.

## 7 PUBLIC DEMO

### 7.1 PUBLIC DEMO OVERVIEW

**Public demo interface.**

Figure SI.9 presents the public demo interface for OPENSCHOLAR, which is available at `https://openscholar.allen.ai/`. The demo is powered by our 8B model and uses a modified pipeline that replaces the static pes2o index with a live public index of 11M+ paper full text snippets and 100M+ paper abstracts, which is updated weekly and can be accessed for free through the Semantic Scholar API. The index is provision via a `vespa`[4] vector DB cluster and combines sparse and dense retrieval, enabling efficient, large-scale, real-time query handling (Singh et al., 2025). As illustrated in Figure SI.9, when given a question such as "Is there any study showing if LMs can help scientists synthesize scientific literature?", OPENSCHOLAR is able to cite newly published papers due to the frequently updated index. The demo also provides fine-grained citations, allowing users to view the exact snippets used by the model during inference, which facilitates easy verification of the model's responses.

**Infrastructure.** We host our OPENSCHOLAR-8B LM (Llama 3.1-OpenScholar 8B) using single H100, and OPENSCHOLAR-reranker using a single T4, via Modal,[5] which provides serverless model deployment on cloud. The retrieval indexes are hosted as end points on the Semantic Scholar api, with the demo app itself hosted on GCP. Using these production-ready components as part of our pipeline allow us to scale automatically. In our load tests, we found that without any scaling, the system is able to handle 100 concurrent users without latency and performance issues.

### 7.2 PUBLIC DEMO STATISTICS

Since the release of the public demo, we have received over 81,000 queries from 37,000 users (as of August 2025). After postprocessing to remove queries less relevant to research, 57,000 fully valid queries remained, spanning a wide range of fields.

We used Claude to classify the subject and query intent. For subjects, computer science was the most common category (41.6%), followed by medicine (9.8%), engineering (8.0%), biology (6.4%), and psychology (5.0%).

For query intent, literature understanding (e.g., comparing two methods, gathering background knowledge) was the most common category (61.6%), followed by finding specific paper findings (e.g., identifying a paper on a particular idea) at 27.2%. We also observed notable portions of queries focused on understanding a specific paper (2.5%) and ideation (1.7%).

Example queries are in Table SI.33.

We have released the post-processed data to facilitate future research.

# Supplementary Discussions

## 8 COMPARISON WITH EXISTING SCIENTIFIC BENCHMARKS

Table SI.34 compares SCHOLARQABENCH with widely used benchmarks in scientific domains, focusing on non-coding tasks such as MMLU (Hendrycks et al., 2020), GPQA (Rein et al., 2024), Humanity's Last Exam (Phan et al., 2025), SciBench (Wang et al., 2024), LITQA (Skarlinski et al., 2024), and KIWI (Xu et al., 2024). Several features distinguish SCHOLARQABENCH from prior work.

First, most existing datasets evaluate LMs' scientific domain knowledge and reasoning abilities (gray colored rows in the table), rather than their ability to perform scientific literature synthesis—a task that requires integrating and summarizing up-to-date research. This makes data collection even more challenging, as it requires research experience and an understanding of the latest literature, rather

[4] `https://vespa.ai/`

[5] `https://modal.com/`

than sourcing questions from existing textbooks (e.g., SciBench QA pairs are partially sourced from existing textbooks).

Second, prior datasets are often limited to short-form outputs (e.g., multiple-choice or brief answers). In contrast, literature synthesis requires detailed, long-form responses with accurate citations, posing significantly greater challenges for both data collection and evaluation (Xu et al., 2024). Standard evaluation approaches such as string matching or simple LLM-based scoring are ill-suited for this setting, as they assume concise, single-aspect answers, and their reliability for complex, multi-aspect synthesis remains uncertain.

While KIWI and LITQA aim to assess literature synthesis capabilities, neither includes an evaluation pipeline for long-form answers. KIWI lacks expert-written reference answers, providing only LM-generated, expert-rated responses, which hinders automatic evaluation. LITQA simplifies the task to multiple-choice QA, even though such predefined answer options rarely exist in real-world scientific inquiry and can introduce artificial constraints that do not reflect authentic information-seeking scenarios (Li et al., 2024).

To our knowledge, SCHOLARQABENCH is the first benchmark to curate realistic, information-seeking questions specifically for literature synthesis, each paired with an expert-authored, long-form answer. It also introduces a comprehensive, multi-faceted evaluation protocol aimed at more reliably assessing synthesis quality. We validate these protocols through large-scale expert evaluations and detailed analyses of LM–human agreement.

## 9 SCHOLARQABENCH EXAMPLES

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

### 9.1 SINGLE-PAPER TASKS

Figure SI.10 shows examples from single-paper tasks.

### 9.2 SCHOLARQA-CS

Table SI.35 shows an example from SCHOLARQA-CS.

### 9.3 SCHOLARQA-BIO

Table SI.36 shows several expert-annotated biomedical queries, along with the papers provided to the annotators for inspiration.

### 9.4 SCHOLARQA-MULTI

Figures SI.11, SI.12, SI.13, SI.14, and SI.15 show expert-annotated examples from different domains. In the original evaluation, we randomized the order of the models and anonymized the responses to prevent any biases. For the examples shown, we substituted the anonymized model IDs with their corresponding actual names.

# Supplementary Tables

**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 briefly define state space models and transformers and detail their technical differences."
- (do not read right away) A source 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.

Supplementary Table SI.1: **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.

Supplementary Table SI.2: **SCHOLARQA-BIO, NEURO annotation instructions**: Instructions given to our expert annotators for authoring questions in biomedicine. Example questions in this passage are composed with assistance from GPT.

**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.).

Supplementary Table SI.3: **SCHOLARQAMULTI annotation instructions**: Instructions given to our expert annotators to annotate questions and answers.

You will be given a question someone asked (in `<question></question>` tags) and the corresponding response (in `<response></response>` tags) given to them by an assistant. You will then be given a specific criterion of the response to evaluate (in `<criterion></criterion>` tags).
Return a score on a scale of 0 to 10 indicating how appropriate the response is based on the given criterion. Judge only the specified aspect(s), not any other qualities of the answer. Output JSON in the format: `"score": x`.

Supplementary Table SI.4: **Instructions for rubric evaluations.** Evaluation instructions to assess if the model output includes certain rubric items for SCHOLARQABENCH-CS.

| | Comparison | Pearson Correlation | p-value |
|---|---|---|---|
| System 1 | Human A vs Human B | 0.74 | 0.00 |
| | Human A vs Model | 0.87 | 0.00 |
| | Human B vs Model | 0.73 | 0.00 |
| System 2 | Human A vs Human B | 0.87 | 0.00 |
| | Human A vs Model | 0.79 | 0.00 |
| | Human B vs Model | 0.80 | 0.00 |

Supplementary Table SI.5: **Pearson correlations and p-values for agreement on rubric scores.** We recruited and asked expert annotators with Ph.D. to evaluate two systems (system 1 and 2) responses using rubric accuracy on SCHOLARQABENCH-CS. Each question is annotated by two expert annotators, and we compute Pearson correlations between human annotators, as well as Pearson correlations between the LLM judge and one expert annotators.

| Aspect | Definition | Instruction |
|---|---|---|
| **Organization** | Evaluate if the output is well-organized and logically structured. | Table SI.7 |
| **Coverage** | Evaluate if the output provides sufficient coverage and amount of information. | Table SI.8 |
| **Relevance** | Does the response stay on topic and maintain a clear focus to provide a useful response to the question? | Table SI.9 |
| **Usefulness** | Evaluate if the output is useful overall. | Table SI.10 |

Supplementary Table SI.6: **Evaluation protocols for writing quality**. We use the same evaluation scheme and instructions for human and evaluator LMs.

**Instructions**
Evaluate if the answer is well-organized and logically structured. An acceptable response should: 1. be clearly structured, grouping related points 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.

Supplementary Table SI.7: **Evaluation instructions and five-scale rubric used to evaluate organization.** We use the same scheme for both LM judge and expert evaluation.

**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.

Supplementary Table SI.8: **Evaluation instructions and five-scale rubric used to evaluate coverage.** We use the same scheme for both LM judge and expert evaluation.

**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 provide 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.

Supplementary Table SI.9: **Evaluation instructions and five-scale rubric used to evaluate relevance.** We use the same scheme for both LM judge and expert evaluation.

**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 a great summary of the area of my interest.**
- The answer gives a great set of papers that are 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.

Supplementary Table SI.10: **Evaluation rubrics for overall helpfulness.** This overall usefulness evaluation is applied only in expert evaluations, as we found that LM-judge assessments of overall usefulness are often much more optimistic than expert human assessments.

| | Accuracy (%, ↑) | | | Δ (Model-Human) | | |
|---|---|---|---|---|---|---|
| | Relevance | Coverage | Organization | Relevance | Coverage | Organization |
| GPT4o | 87.2 | 86.0 | 77.9 | 0.10 | -0.16 | -0.60 |
| OS-GPT4o | 84.9 | 86.0 | 77.9 | 0.57 | 0.20 | 0.12 |
| OS-8B | 73.2 | 75.5 | 83.8 | 0.17 | -0.31 | -0.02 |
| Average | 81.8 | 82.6 | 84.1 | 0.28 | -0.09 | -0.16 |

Supplementary Table SI.11: **Agreement between human and Prometheus assessments on Human and OPENSCHOLAR (GPT4o) answers**. We compute accuracy, after converting the original five-scale prediction to a three-scale prediction, and compute the averaged score difference between Model (Prometheus) and Human.

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:**

Supplementary Table SI.12: **Instruction for citation evaluation.** The instruction is adapted from Yue et al. (2023)).

| |
|---|
| **GPQA**<br>**Input:** Areaction of a liquid organic compound, which molecules consist of carbon and hydrogen atoms, is performed at 80 centigrade and 20 bar for 24 hours. In the proton nuclear magnetic resonance spectrum, the signals with the highest chemical shift of the reactant are replaced by a signal of the product that is observed about three to four units downfield. Compoundsfromwhichposition in the periodic system of the elements, which are also used in the corresponding large-scale industrial process, have been mostly likely initially added in small amounts? A) Ametal compound from the fifth period. B) Ametal compound from the fifth period and a non-metal compound from the third period. C) Ametal compound from the fourth period. D) Ametal compound from the fourth period and a non-metal compound from the<br>**Output:** C |
| **Humanity's Last Exam**<br>**Input:** Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.<br>**Output:** 2 |
| **SciBench**<br>**Input:** Halley's comet, which passed around the sun early in 1986, moves in a highly elliptical orbit with an eccentricity of 0.967 and a period of 76 years. Calculate its minimum distances from the Sun.<br>**Output:** $8.8 \times 10^{10}$ m |
| SCHOLARQABENCH-CS<br>**Input:** What are ways to cool the center-of-mass (CoM) motion of levitated nanoparticles?<br>**Output:** 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. Novotn ´y et al. 2021]. An alternative method involves levitating the nanosphere in an optical cavity and cooling via coherent scattering [Uro ˇs Deli ´c 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 [Uro ˇs Deli ´c 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 (answer continues) |

Supplementary Table SI.13: **Comparison of example instances from recent scientific benchmarks.** We provide examples from SCHOLARQABENCH and other datasets. Unlike prior benchmarks, which often require short-form answers, SCHOLARQABENCH requires long-form responses.

Provide a detailed, informative answer to the following research-related question. Your answer should be more than one paragraph, offering a comprehensive overview.
Base your answer on multiple pieces of evidence and references, rather than relying on a single reference for a short response.
Aim to give a holistic view of the topic. Ensure the answer is well-structured, coherent, and informative so that real-world scientists can gain a clear understanding of the subject. Rather than simply summarizing multiple papers one by one, try to organize your answers based on similarities and differences between papers.
Make sure to add citations to all citation-worthy statements using the provided references (References), by indicating the citation numbers of the corresponding passages. More specifically, add the citation number at the end of each relevant sentence e.g., "This work shows the effectiveness of problem X [1]." when the passage [1] in References provides full support for the statement.
You do not need to add the author names, title or publication year as in the ordinal paper writing, and just mention the citation numbers with your generation.
Not all references may be relevant, so only cite those that directly support the statement.
You only need to indicate the reference number, and you do not need to add the Reference list by yourself. If multiple references support a statement, cite them together (e.g., [1][2]). Yet, for each citation-worthy statement, you only need to add at least one citation, so if multiple evidence support the statement, just add the most relevant citation to the sentence.
Your answer should be marked as [Response_Start] and [Response_End].

Here's an example:
**References:** *A set of references*
**Question:** *A question*
[Response_Start] *An answer* [Response_End]

Now, please answer this question:

Supplementary Table SI.14: **The instruction for generating an initial response.** We instruct an LM to generate a response given the original query and retrieved passages.

Given an answer to a scientific question based on the most recent scientific literature, make a list of feedback.
Prioritize the feedback by listing the most critical improvements first. Regarding the content improvements, it is often helpful to ask for more results on applications to different tasks, elaborate on details of crucial methods, or suggest including other popular methods.
Stylistic improvements can include better organization or writing enhancements. For each suggested improvement requiring additional information from the literature that is not discussed in passages, formulate a question to guide the search for missing details.
If the feedback primarily concerns stylistic or organizational changes, omit the need for an additional question. Your answer should be marked as [Response_Start] and [Response_End]. Each feedback should be preceded by 'Feedback: ', and additional questions should be preceded by 'Question: '. Your question will be used to search for additional context, so they should be self-contained and understandable without additional context.

Here's an example.

**Question**: *example question*
**Answer**: *example answer*
[Response_Start]*example feedback*[Response_Start]

Now, please generate feedback for this question.

Supplementary Table SI.15: **The instruction for generating feedback given an initial response.** We prompt an LM to generate a set of feedback as well as optional search queries for further retrieval, based on the initial response.

We provide a question related to recent scientific literature, an answer from a strong language model, and feedback on the answer.
Please incorporate the feedback to improve the answer.
Only modify the parts that require enhancement as noted in the feedback, keeping the other sentences unchanged.
Do not omit any crucial information from the original answer unless the feedback specifies that certain sentences are incorrect and should be removed.
If you add new paragraphs or discussions, ensure that you are not introducing repetitive content or duplicating ideas already included in the original response.
Use existing references presented under References to support the new discussions, referring to their citation numbers.
Do not remove new lines or paragraphs in the original answer, unless the feedback specifies that certain sentences are incorrect and should be removed, or the paragraph organization should be changed.
Your answer should be marked as [Response_Start] and [Response_End].

Here's an example.
**References**: *Set of references*
**Question**: *example question*
**Answer**: *example original answer*
**Feedback**: *example feedback*
[Response_Start]*example updated answer*[Response_Start]

Now, please generate feedback for this question.

Supplementary Table SI.16: **Instruction for incorporating feedback.**
We prompt an LM to edit the response, given feedback and additional retrieval.

We give you a short paragraph extracted from an answer to a question related to the most recent scientific literature, and a set of evidence passages. Find all of the citation-worthy statements without any citations, and insert citation numbers into the statements that are fully supported by any of the provided citations listed as References. If none of the passages support the statement, do not insert any citation, and leave the original sentence as is, but do your best to insert citation. If multiple passages provide sufficient support for the statement, you only need to insert one citation, rather than inserting all of them. Your answer should be marked as [Response_Start] and [Response_End].

Here's an example:
**Statement**: *example statement*
**References**: *example reference*
[Response_Start]*example statement with citation*[Response_Start]

Now, please generate feedback for this question.

Supplementary Table SI.17: **Instruction for verifying citations.** We first split the final response into paragraphs, and then verify and insert missing citations given the retrieved context.

**Instructions**
Answer the following question based on up-to-date academic literature. Your answer should be detailed and informative, and is likely to be more than a single sentence. Your answer should come with citation numbers, and your answer should be followed by a list of the references. All of your citation worthy statements must be supported by one of the references. The references are numbered [1], [2], ..., [n].
One really important thing is, for references, you must only list the titles of the papers, and DO NOT include the author names or published venues.

Question: How do language models leverage parametric and non-parametric knowledge?
Answer: There are several ways in which language models leverage parametric and non-parametric knowledge.
Large language models can memorize a wealth of web knowledge during pre-training time [1], and can generate informative responses to information-seeking context and perform competitively on knowledge intensive tasks such as TriviaQA, without relying on retrieved context [4]. They have the advantage of not relying on any external corpus and are more efficient than their retrieval-augmented counterparts [6].
Retrieval-augmented language models, such as Atlas [3], combine a language model with a retrieval model and have shown to be effective to generate more factually accurate text with explicit attributions to existing context [2]. Recent work found that retrieval-augmented language models perform competitively especially in long-tail knowledge, where language models relying on parametric knowledge often struggle with [5].
Overall, the use of nonparametric knowledge and memorization can improve the performance of language models on various NLP tasks, and there are several methods that effectively leverage these aspects. Researchers can explore these methods and find ways to apply them to their work in order to improve the performance of their language models.
References:
[1] Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models
[2] Atlas: Few-shot Learning with Retrieval Augmented Language Models
[3] Language Models are Few-Shot Learners
[4 When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories
[5] Democratizing Large Language Models via Personalized Parameter-Efficient Fine-tuning
[6] RECOMP: Improving Retrieval-Augmented LMs with Context Compression and Selective Augmentation

Supplementary Table SI.18: **The instruction for parametric LM baselines.** We prompt an LM to generate a response with paper titles that support scientific claims.

**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?
**Response:** Finetuning an auto-regressive language model (LM) for dense passage retrieval is a recent area of research in natural language processing. Recent works have explored the use of auto-regressive LMs for dense passage retrieval, leveraging their ability to model long-range dependencies and generate coherent text.
One notable approach is the use of a masked language modeling objective during finetuning, where some input tokens are randomly replaced with a [MASK] token, and the LM is trained to predict the original token [1]. This approach has been shown to be effective in dense passage retrieval tasks, as it allows the LM to focus on learning contextualized representations of the input text.
Another approach is the use of a named entity recognition (NER) task during finetuning, where the LM is trained to identify and predict named entities in the input text [2]. This approach has been shown to improve the performance of the LM in dense passage retrieval tasks, as it allows the LM to learn contextualized representations of named entities.
In terms of comparison with bi-directional encoders, recent works have shown that auto-regressive LMs can achieve competitive or even superior performance to bi-directional encoders in dense passage retrieval tasks [3]. This is because auto-regressive LMs can model long-range dependencies and generate coherent text, while bi-directional encoders are limited to modeling fixed-length contextualized representations.
For example, a recent study found that a finetuned auto-regressive LM outperformed a bi-directional encoder in a dense passage retrieval task, with a significant improvement in terms of mean average precision and recall [4]. Another study found that a finetuned auto-regressive LM was able to achieve state-of-the-art performance in a dense passage retrieval task, outperforming a range of bi-directional encoders and other retrieval-based models [5].
Overall, the latest works on finetuning an auto-regressive LM for dense passage retrieval have shown promising results, and further research is needed to explore the potential of this approach in various NLP tasks.
**Citations**
[1] Dense Passage Retrieval with Auto-regressive Language Models
[2] Named Entity Recognition with Auto-regressive Language Models
[3] Passage Retrieval with Auto-regressive Language Models
[4] A Study on the Effectiveness of Auto-regressive Language Models in Dense Passage Retrieval
[5] State-of-the-Art Passage Retrieval with Auto-regressive Language Models

**Question:** What are the current state of the art in which photonic metamaterial surfaces are utilized to enhance Label-Free Imaging in Phase Contrast Microscopy and Refractometric Microscopy?
**Response:** The current state of the art in utilizing photonic metamaterial surfaces to enhance Label-Free Imaging in Phase Contrast Microscopy and Refractometric Microscopy is the development of metasurfaces that can manipulate light in a highly controlled and tailored manner. These metasurfaces can be designed to have specific properties such as negative refractive index, anomalous refraction, and polarization conversion, which can be used to enhance the imaging capabilities of Phase Contrast Microscopy and Refractometric Microscopy [1]. For example, a recent study demonstrated the use of a metasurface-based Phase Contrast Microscopy system that achieved a contrast enhancement factor of 10, allowing for the imaging of sub-wavelength features with high resolution [2]. Another study showed that a metasurface-based Refractometric Microscopy system can achieve a refractive index sensitivity of $10^-4$, enabling the detection of small changes in refractive index with high accuracy [3]. These advances have the potential to revolutionize the field of Label-Free Imaging, enabling the imaging of complex samples with unprecedented resolution and sensitivity.
**Citations**
[1] Metasurface-based Label-Free Imaging: A Review
[2] High-contrast Phase Contrast Microscopy with Metasurfaces
[3] High-sensitivity Refractometric Microscopy with Metasurfaces

Supplementary Table SI.19: **Parametric baseline output examples**. We show example outputs of Llama 3.1 8B Instruct parametric baselines. The paper title with red text shows the examples of hallucinated titles.

**Question:** How can user modeling techniques be improved to capture and leverage individual preferences more effectively in personalized recommendation systems?
**Response:** User modeling techniques play a crucial role in personalized recommendation systems by tailoring content to individual preferences. To improve these techniques, several strategies can be considered, based on recent academic literature.
Firstly, incorporating deep learning approaches such as neural collaborative filtering can enhance the representation of user preferences by capturing complex, non-linear user-item interactions [1]. These models can learn a richer representation of user preferences through multiple layers of abstraction, offering better personalization compared to traditional matrix factorization techniques.
Secondly, context-aware recommendation systems that leverage additional data such as time, location, and device information can provide more accurate recommendations by understanding the contextual nuances of user preferences [2]. Such systems can dynamically adjust recommendations based on the current context, leading to a more personalized user experience.
Another important aspect is the integration of psychological and behavioral theories into user modeling. Understanding the underlying motivations and cognitive biases that drive user choices can inform the development of more sophisticated models that predict not just what users will choose, but why they make those choices [3].
Moreover, the use of transfer learning can be beneficial in improving user modeling techniques. By transferring knowledge from related domains or tasks, models can be quickly adapted to new users or items, reducing the cold-start problem and improving the system's ability to capture individual preferences effectively [4].
Finally, privacy-preserving techniques such as federated learning and differential privacy can help in capturing user preferences while maintaining user trust [5]. By ensuring that user data is kept secure and private, these techniques can encourage more users to engage with recommendation systems, thereby providing richer data for personalization.
In summary, by leveraging deep learning, context-awareness, psychological insights, transfer learning, and privacy-preserving methods, user modeling techniques can be significantly improved to capture and leverage individual preferences more effectively in personalized recommendation systems.
**Citations:**
[1] Neural Collaborative Filtering
[2] Context-Aware Recommender Systems: A Survey of the State of the Art
[3] A Psychological Approach to the Design of Recommender Systems
[4] Transfer Learning in Collaborative Filtering for Sparsity Reduction
[5] Privacy-Preserving Personalized Recommendation Systems: Challenges and Opportunities

Supplementary Table SI.20: **Parametric baseline output examples**. We show example outputs of GPT4o baseline. The paper title with red text shows the examples of hallucinated titles.

Provide a detailed, informative answer to the following research-related question. Your answer should be more than one paragraph, offering a comprehensive overview.
Base your answer on multiple pieces of evidence and references, rather than relying on a single reference for a short response.
Aim to give a holistic view of the topic. Ensure the answer is well-structured, coherent, and informative so that real-world scientists can gain a clear understanding of the subject. Rather than simply summarizing multiple papers one by one, try to organize your answers based on similarities and differences between papers.
Make sure to add citations to all citation-worthy statements using the provided references (References), by indicating the citation numbers of the corresponding passages. More specifically, add the citation number at the end of each relevant sentence e.g., "This work shows the effectiveness of problem X [1]." when the passage [1] in References provides full support for the statement.
You do not need to add the author names, title or publication year as in the ordinal paper writing, and just mention the citation numbers with your generation.
Not all references may be relevant, so only cite those that directly support the statement.
You only need to indicate the reference number, and you do not need to add the Reference list by yourself. If multiple references support a statement, cite them together (e.g., [1][2]). Yet, for each citation-worthy statement, you only need to add at least one citation, so if multiple evidence support the statement, just add the most relevant citation to the sentence.
Your answer should be marked as [Response_Start] and [Response_End].

Here's an example:
**References:** *A set of references*
**Question:** *A question*
[Response_Start] *An answer* [Response_End]

Now, please answer this question:

Supplementary Table SI.21: **Instruction for standard RAG baseline (OSDS RAG).** We provide a one-shot example.

| Bi-encoder | Cross-Encoder | NDCG@1 | NDCG @3 | NDCG@5 | NDCG@10 | NDCG@100 |
|---|---|---|---|---|---|---|
| BGE-base | – | 93.9 | 96.1 | 96.6 | 96.7 | 96.9 |
| GIST | – | 94.2 | 96.2 | 96.5 | 96.7 | 96.9 |
| Contriever | – | 92.2 | 94.9 | 95.3 | 95.5 | 95.8 |
| OS-Retriever | – | <u>94.9</u> | <u>96.8</u> | <u>97.1</u> | <u>97.2</u> | <u>97.4</u> |
| OS-Retriever | BGE reranker | 95.5 | 97.3 | 97.4 | 97.6 | 97.7 |
| OS-Retriever | OS-Reranker | **96.0** | **97.8** | **97.9** | **97.9** | **98.1** |

Supplementary Table SI.22: **Retrieval performance of bi-encoder and cross-encoder models.** We evaluate several bi-encoder (embedding) and cross-encoder models on our synthetically generated retrieval benchmark. For the cross-encoder experiments, we use our SCHOLARQABENCH retriever as the base bi-encoder, as it achieved the best performance among retrievers of the same scale. We report NDCG scores, with the best bi-encoder–only result indicated by underlining and the best overall result indicated in bold.

| | OSDS-Emb | S2 API | Web Search | Rub | Cite (Precision) | Cite (Recall) | Cite (F1) |
|---|---|---|---|---|---|---|---|
| All | ✓ | ✓ | ✓ | 49.6 | 48.2 | 347.1 | 47.6 |
| *w/o* OSDS | | ✓ | ✓ | 47.6 | 36.4 | 38.4 | 37.4 |
| *w/o* S2 | ✓ | | ✓ | 48.8 | 35.4 | 37.1 | 34.1 |
| *w/o* Web | ✓ | ✓ | | 48.7 | 34.7 | 33.4 | 36.3 |

Supplementary Table SI.23: **Retrieval Pipeline Ablations**. ✓indicates that the retrieval pipeline is used. We evaluated the performance on data-CS.

**Question:** What are the leading approaches for evaluating complex scientific question answering systems in NLP?

**8B + RAG$^{\text{OSDS}}$:** Evaluating complex scientific question answering systems in NLP presents several challenges, particularly when the questions are open-ended, require reasoning, or involve multiple steps [0]. Common evaluation metrics for factual question answering, such as ROUGE and BLEU, are insufficient to assess these systems.

One approach is to use a ranking-based evaluation, where the goal is to rank the provided answers in order of their relevance to the question [1]. Following this method, human evaluators assess the relevance of each answer based on their perception of the answer's usefulness to the user, considering factors like accuracy, coherence, and relevance to the question. Another approach is to use a ranking-based evaluation with a specific metric, such as Mean Average Precision (MAP), to assess the performance of a ranking model. This metric calculates the average precision of the relevant documents at different recall levels, providing a more comprehensive evaluation of the ranking quality [5].

Moreover, researchers have proposed using machine learning-based approaches for answer ranking. For instance, a study on learning to rank answers to non-factoid questions from web collections used several linguistic features, such as word overlap, positional similarity, and semantic relatedness, to evaluate answer ranking models [5].

Question generation is another crucial aspect of NLI, and recent studies have focused on using machine learning techniques, such as sequence-to-sequence models, to generate questions from user input [0]. These models have proven effective in generating fluent and coherent questions, but their performance is highly dependent on the quality of the training data.

To evaluate question generation systems, researchers have used various metrics, such as BLEU score, to assess the fluency and grammar of generated questions and to compare them to human-generated questions [2]. Some studies have also explored using human evaluation, where human evaluators assess the relevance, coherence, and naturalness of generated questions [4]. ...

Furthermore, a study on collaborative and AI-aided exam question generation using Wikidata highlights the importance of considering the contributions of individual system modules to overall performance [2]. This study suggests that a multi-step approach, involving both machine learning and linguistic methods, can effectively generate high-quality questions.

In summary, evaluating complex scientific question answering systems in NLP requires a combination of ranking-based evaluation, machine learning-based approaches, and human evaluation. The choice of evaluation metric depends on the specific goals and characteristics of the system being evaluated. Recent studies have shown the effectiveness of machine learning-based approaches, such as sequence-to-sequence models, and the importance of considering the contributions of individual system modules to overall performance.

**OS-8B:** To evaluate complex scientific question answering systems in NLP, researchers and practitioners employ various approaches. Here, we'll outline some of the leading methods:

1. **Holistic evaluation methods**: These approaches use end-to-end models, such as transformers, to evaluate the quality of generated responses in a holistic manner. Examples include BERTScore, RUSE, RUBER, BLEURT, ADEM, and MaUde. These models learn an alignment score between context and response to approximate human judgments [0].

2. **Human Evaluation**: Human evaluation is considered the most trustworthy evaluation method, quality, and reliability of NLP systems depend highly on the quality of the evaluation itself. Human-driven evaluations, such as Holistic Cognitive Evaluation Scores (HCES), provide a comprehensive assessment of the output based on criteria like adequacy, fluency, and coherence [1].

...

7. **Rationale-based evaluation**: Rationale-based evaluation assesses the quality of the rationales provided by the system for its answers [7]. This approach has been shown to be effective in evaluating the performance of question-answering systems and can provide insights into the reasoning process of the system.

These evaluation methods can be used effectively in combination to assess the performance of complex scientific question answering systems in NLP, providing a holistic understanding of their strengths and weaknesses [4]. Such integrated evaluation methods can provide a more comprehensive picture of the system's performance, including its ability to handle multiple types of questions, its ability to retrieve relevant information from various sources, and its ability to reason and make connections between different pieces of information [7].

Overall, a combination of these evaluation methods can provide a holistic understanding of a complex scientific question answering system's strengths and weaknesses, and help identify areas for improvement [0].

Supplementary Table SI.24: **An comparison of model responses of RAG$^{\text{OSDS}}$ and OPENSCHOLAR-8B.** The OPENSCHOLAR-8B enhanced responses in terms of coverage and organization.

**Question:** How can question generation be used to mitigate hallucination in LLMs?

**8B + RAG$^{\text{OSDS}}$:** Question generation can be used to mitigate hallucination in LLMs by leveraging the model's ability to reason and generate responses to questions.

One approach is to use Retrieval-Augmented Generation (RAG) methods, which condition the model's output on relevant documents retrieved from an external knowledge source [0]. Another approach is to employ Chain-of-Thought prompting, which requires the model to provide intermediate reasoning steps to generate responses [2].

Additionally, question generation can be used to mitigate hallucination by repeatedly generating sentences through the model and actively detecting and mitigating hallucinations [4].

Furthermore, self-reflection methods can be employed, which involve iteratively generating pertinent background knowledge for a given question, followed by a factuality evaluation, and repeating the process until a satisfactory level of factuality is achieved [8][9].

These methods can enhance the model's ability to provide accurate, reliable, and factually grounded responses.

**OS-8B:** Mitigating hallucination in large language models (LLMs) is a critical problem that has garnered significant attention in recent years. One potential method to address this issue is through the use of question generation. Retrieval-Augmented Generation (RAG) is a promising strategy that has shown effectiveness in mitigating hallucinations in LLMs by conditioning the model's output on relevant documents retrieved from external knowledge sources [0]. This approach enables the model to accurately recall knowledge and generate responses that are more faithful to the input [2].

Another approach is to employ Chain-of-Thought prompting, which requires the model to provide intermediate reasoning steps to generate responses. This method has been shown to increase knowledge recall, which can boost performance in editing facts under multi-hop settings [2]. Self-reflection methods, such as iterative self-reflection, have also been proposed, which involve generating, scoring, and refining knowledge and answers until they reach satisfactory levels [8][9]. These methods have been demonstrated to be effective in mitigating hallucination in LLMs, with accuracy, consistency, and entailment improvements reported in medical GQA tasks [7][9].

In addition, question generation can be used to mitigate hallucination by repeatedly iteratively generating sentences through the model and actively detecting and mitigating hallucinations [4]. This approach has been shown to reduce hallucination rates, with a recall of 88% and a mitigation success rate of 57.6% reported in a study using a retriever-generative framework [3]. (cont.)

Additionally, other approaches, such as knowledge graph-based retrofitting [10], real-time verification and rectification [11], and fine-tuning with caption rewrites [13], have also been proposed to mitigate hallucination in LLMs. Real-time Verification and Rectification (EVER) is another approach that employs real-time validation to identify both intrinsic and extrinsic hallucinations during the generation process [11]. This method can prevent error propagation and ensure the trustworthiness of the generated content. (cont)

Furthermore, fine-tuning large vision-language models with caption rewrites can also be effective in reducing fine-grained hallucination [13].

Experimental results from these methods have shown promising results in reducing hallucination in LLMs, with some achieving accuracy, consistency, and entailment improvements [7][9]. However, further research is needed to explore the effectiveness of these methods in various domains and tasks. (cont.)

**Feedback**:

✗ Only one specific method, Chain-of-Thought prompting, is mentioned as a way to mitigate hallucination through question generation

✗ The answer could be improved by providing more concrete examples and evaluation results of these methods.

Supplementary Table SI.25: **Another comparison of model responses of RAG$^{\text{OSDS}}$ and OPEN-SCHOLAR-8B.** In this example, while OS-8B generates helpful feedback, the final edited response includes repetitive discussions and less relevant discussions.

| |
|---|
| **Question:** How do large language models like ChatGPT impact the diversity of published scientific ideas? |
| **PaperQA2:** Large language models (LLMs) like ChatGPT have a complex impact on the diversity of published scientific ideas. These models, by automating certain aspects of the research and writing process, have the potential to both enhance and diminish the diversity of scientific discourse. On one hand, LLMs can democratize the research process by assisting with tasks such as grammar correction and translation, potentially enabling a broader range of researchers to contribute their ideas. This could lead to an increase in the diversity of scientific publications by incorporating perspectives from non-native English speakers and researchers from underrepresented regions **(Fecher2023 pages 6-7)**. Additionally, LLMs can generate a wide array of ideas, which might help in diversifying the topics and concepts explored in scientific literature **(Fecher2023 pages 5-6)**. However, the impact of LLMs on the diversity of scientific ideas is not solely positive. The training data of these models often contain biases, which can lead to the perpetuation of existing stereotypes and trivialities in scientific discourse **(Fecher2023 pages 6-6)**. Moreover, the tendency of LLMs to generate generic and vague content could result in a homogenization of scientific literature, as these models often produce similar outputs based on the most common patterns seen in their training data **(Binza2312 pages 5-5)**. This could undermine the novelty and variety of scientific ideas, potentially leading to a reduction in epistemic diversity **(Binza2312 pages 5-5)**. Furthermore, the reliance on LLMs for generating scientific content raises concerns about the integrity and originality of scientific discourse. The models' limitations in understanding complex scientific content can lead to inaccuracies and the propagation of errors, which might mislead researchers and affect the quality of scientific outputs **(Fecher2023 pages 5-6)**. In summary, while LLMs like ChatGPT offer tools that could potentially expand the diversity of scientific ideas by supporting a wider range of contributors and ideas, they also pose significant risks that could lead to a decrease in the diversity and integrity of scientific discourse **(Fecher2023 pages 6-7; Binza2312 pages 5-5; Fecher2023 pages 5-6)**. |

Supplementary Table SI.26: **An example response from PaperQA2**. We found that PaperQA2 answers tends to cite the same few papers.

| Category | % | Examples |
|---|---|---|
| Coverage | 29% | A is clearly more helpful as it provides a clean organization and cover more papers.<br>Although response A shows a smaller coverage, it is slightly better than B due to detailed description of the techniques in generating reasoning data.<br>both answers are really organized and provides a nice overview! I slightly prefer A as it's more detailed.<br>B is too concise and the organization could be improved.<br>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.<br>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.<br>B contains more comprehensive results that addresses both phenomena and theories. |
| 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".<br>A contains some unnecessary details that don't help understand the topic.<br>Response A has a poor organization and a little deviation from the topic when it talks about privacy and safety of RLHF.<br>Although response A shows a smaller coverage, it is slightly better than B due to detailed description of the techniques in generating reasoning data.<br>B focuses on numerical methods and gives more approachable models. |
| Organization | 12% | Lots of redundancy and repeat in B<br>Although B has lots of overlaps, but B covers more aspects in grating behavior |
| Citations | 9% | Some information provided by B is irrelevant to the quantum effect (like dark matter) and some citations are not representative (biological). |

Supplementary Table SI.27: **Explanations on preferences**. We conducted manual evaluations on the optional explanations for the pair-wise preferences. Our evaluation shows that many expert annotators mention answer coverage.

**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:** Organization: 5.0 Relevance: 5.0 Coverage: 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 ( $3 \times 10^{-7}$, mbar) [Gerard P. Conangla et al. Optimal Feedback Cooling of a Charged Levitated Nanoparticle with Adaptive Control.][Gerard P. Conangla et al. 2018]. (cont.)

**Fine-Grained Assessments:** Organization: 5.0 Relevance: 4.0 Coverage: 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.

Supplementary Table SI.28: 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 $\mu$m 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 $\mu$m, 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 $\mu$m, and they are typically administered via bolus intravenous injection.

**Fine-Grained Assessments:** Organization: 4.0 Relevance: 4.0 Coverage: 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 $\mu m$, 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:** Organization: 3.0 Relevance: 5.0 Coverage: 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.)

Supplementary Table SI.29: Comparison of human and model answers (Biomedicine).

**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:** Organization: 5.0 Relevance: 4.0 Coverage: 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 modelśbroad 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. (cont.)

**Fine-Grained Assessments:** Organization: 5.0 Relevance: 4.0 Coverage: 4.0

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

Supplementary Table SI.30: Comparison of human and model answers (Computer Science).

**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 OpenAIś 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:** Organization: 5.0 Relevance: 4.0 Coverage: 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, itś 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:** Organization: 5.0 Relevance: 5.0 Coverage: 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.

Supplementary Table SI.31: Comparison of human and model answers (Computer Science).

**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:** Organization: 5.0 Relevance: 5.0 Coverage: 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 $\mu$m 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 siliconś 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 $\mu$m, producing results that agreed well with theoretical predictions [G. Bimonte et al. Measurement of the Casimir Force between 0.2 and 8 $\mu$m: 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 $\mu$m: 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:** Organization: 5.0 Relevance: 4.0 Coverage: 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.

Supplementary Table SI.32: Comparison of human and model answers (Photonics).

| Category | Examples |
|---|---|
| Literature Understanding | what is Secure Channel protocol in the context of the SIM and GSM in particular SCP81? |
| | Does the implementation of transparent artificial intelligence algorithms lead to increased user trust and compliance rates? |
| | Show how matrix math is used to represent a slot machine result grid, paytable, and lines. Then give an example of calculating a total award |
| Paper Findings | Paper Finding: Give me a list of important papers on the topic of robot errors in HRI |
| | Find papers on self supervised audio embedding models. |
| | What are the papers in which authors used fine-tuning to jailbreak LLMs? |
| | Are there any papers showing whether scaling up retrieval corpora of RAG can improve performance? |
| Ideation | Give me novel ideas to work on like AI based audit system for vulnerability analysis and management in cybersecurity |
| | help me with ideas on how computer vision can be used to help sales reps increase sales |

Supplementary Table SI.33: **OPENSCHOLAR demo query examples grouped by query intent.** We present a couple of examples of real-world user queries posed in the OPENSCHOLAR public demo of Computer Science domains.

| Dataset | Focus of Questions | Output Format | Evaluation Pipeline |
|---|---|---|---|
| GPQA | Domain knowledge & reasoning | MC | String matching |
| HLE | Domain knowledge & reasoning | MC, SF | Single-faced LM judge |
| SciBench | Domain knowledge & reasoning | MC, SF | String matching |
| LITQA | Literature Synthesis | MC | String matching |
| KIWI | Literature Synthesis | LF* (no human ref) | - |
| Ours | Literature Synthesis | MC, LF | String matching, Multi-faced LM judge |

Supplementary Table SI.34: **Comparison with other datasets.** We compare SCHOLARQABENCH with related datasets, where "MC", "SF", and "LF" denote multiple-choice answers, short-form generation, and long-form generation, respectively. The datasets considered include GPQA (Rein et al., 2024), Humanity's Last Exam (HLE) (Phan et al., 2025), KIWI (Xu et al., 2024), SciBench (Wang et al., 2024), LITQA (Skarlinski et al., 2024), and KIQI (Xu et al., 2024). Datasets that primarily focus on domain knowledge and reasoning are highlighted in gray.

**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 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 ManyTypes4Py 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.**

Supplementary Table SI.35: **Annotation example from SCHOLARQA-CS with associated key ingredients and supporting quotes.** We use the ingredients and quotes as rubrics when evaluating on SCHOLARQABENCH.

**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.

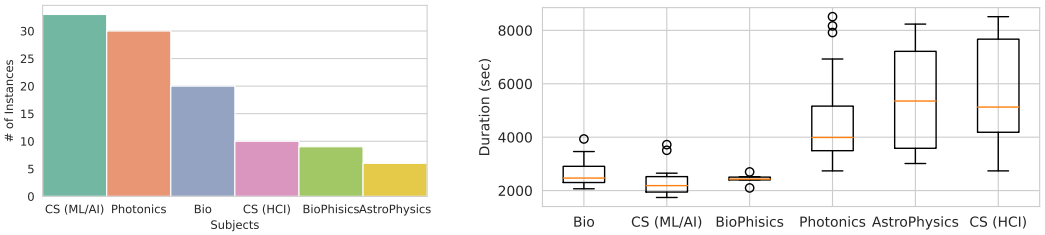
**Question:** How does CRISPR/Cas9 compare to other gene-editing technologies in terms of efficiency and precision?

**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.

Supplementary Table SI.36: **Scholar-Bio query examples**. An "Inspiring Paper" is provided to help expert annotators formulate questions.

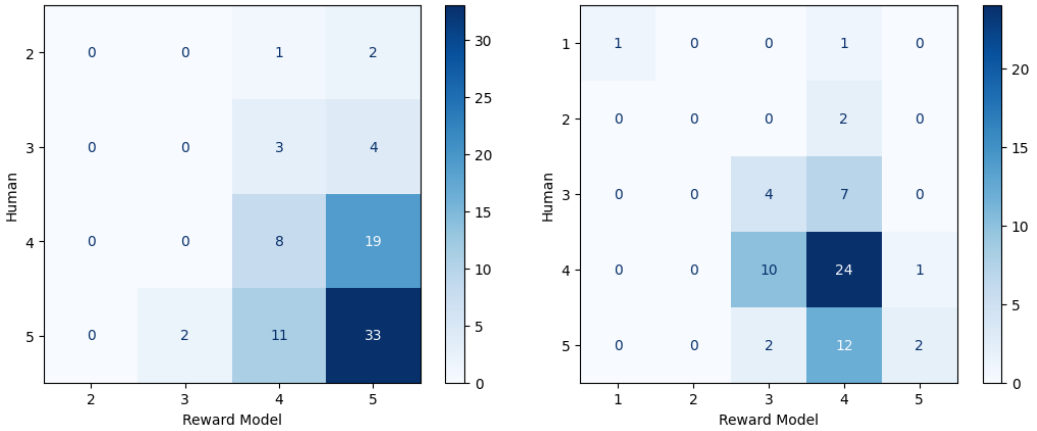
# Supplementary Figures




(a) Subject distributions of expert-annotated answers.

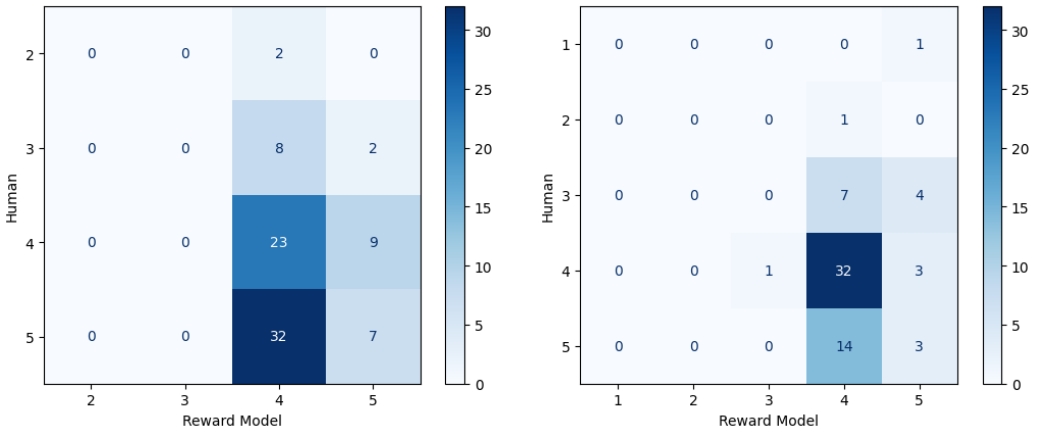
(b) Average per-instance annotation time for each subject (seconds).

Supplementary Figure SI.1: **Analysis of human-written answers:** (a) shows the distribution of instances per subject, and (b) shows the average annotation time per instance per subject.



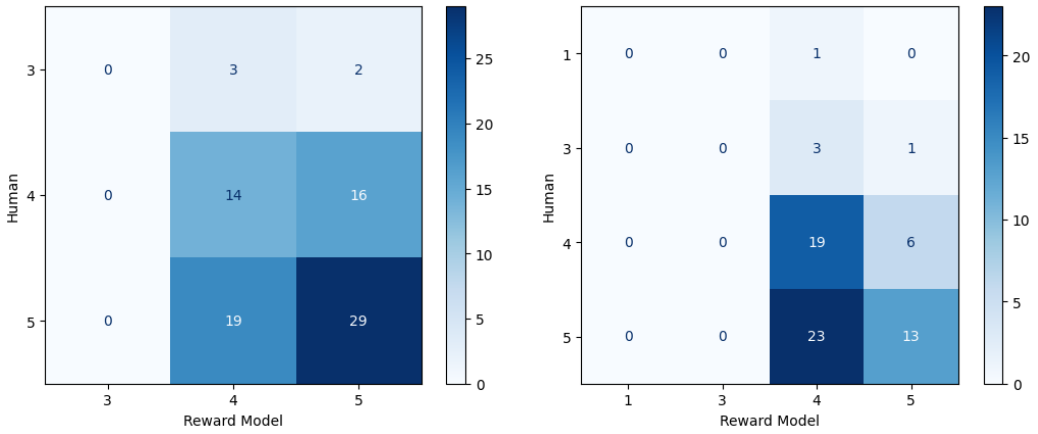

(a) OS-GPT4o answer. (b) OS-8B answer.

Supplementary Figure SI.2: **Agreement between Prometheus and humans on organization.** We plot the confusion matrix between human evaluations and Prometheus (reward model) evaluations for OS-GPT4o and OS-8B responses.




(a) OS-GPT4o answer. (b) OS-8B.

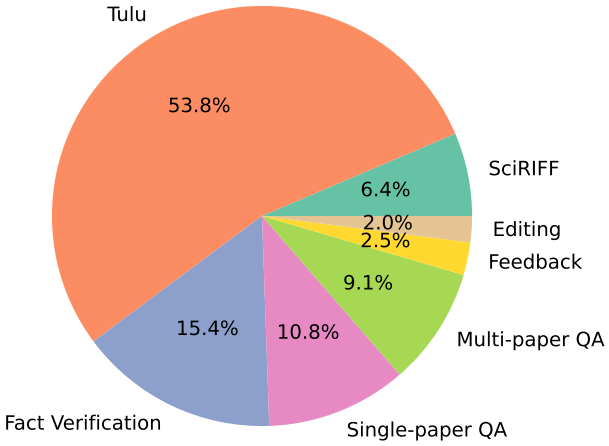
Supplementary Figure SI.3: **Agreement between Prometheus and humans on relevance.** We plot the confusion matrix between human evaluations and Prometheus (reward model) evaluations for OS-GPT4o and OS-8B responses.



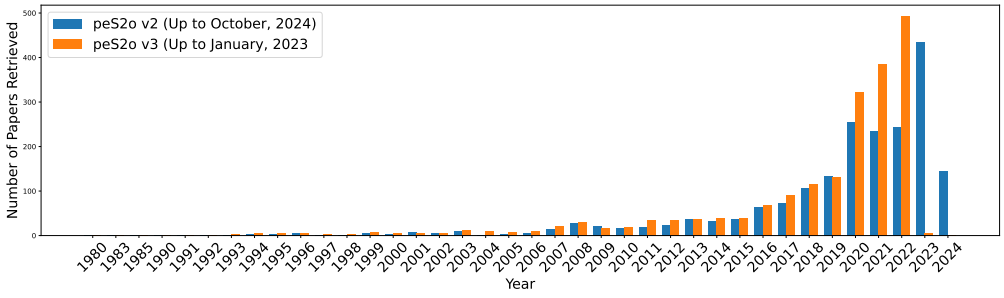

(a) OS-GPT4o answer.

(b) OS-8B answer.

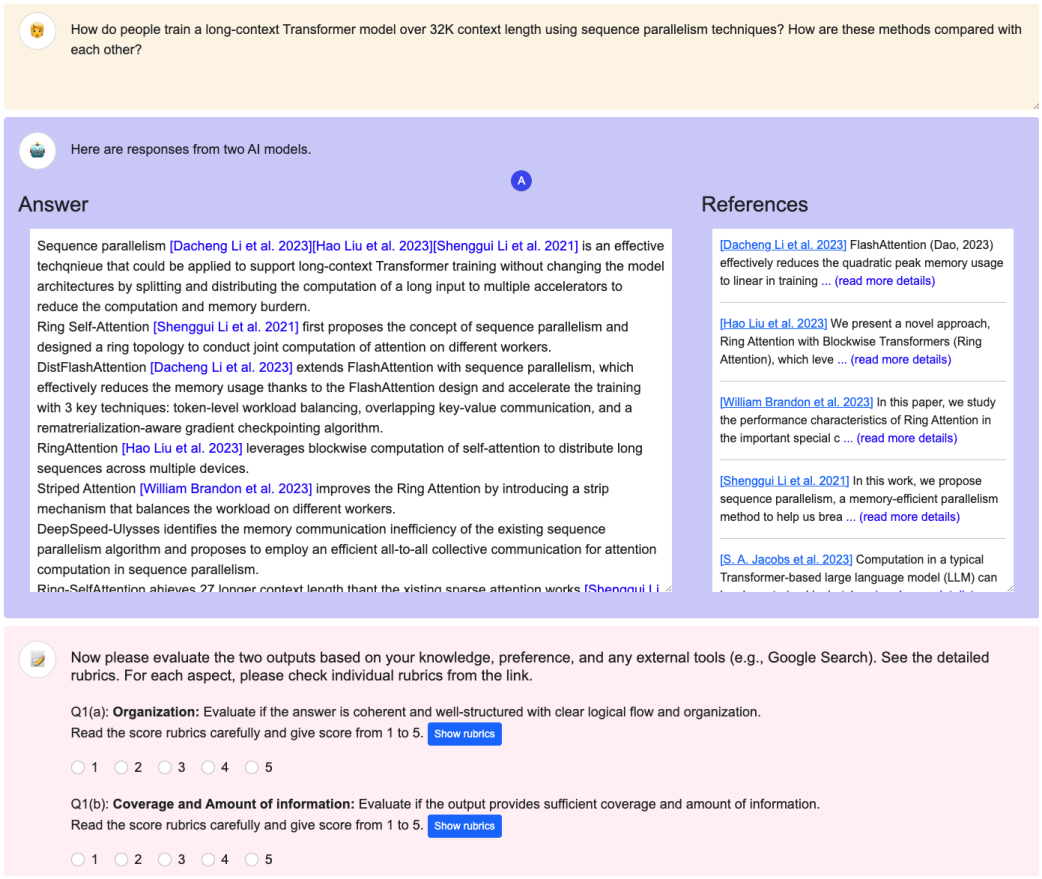
Supplementary Figure SI.4: **Agreement between Prometheus and humans on coverage.** We plot the confusion matrix between human evaluations and Prometheus (reward model) evaluations for OS-GPT4o and OS-8B responses.



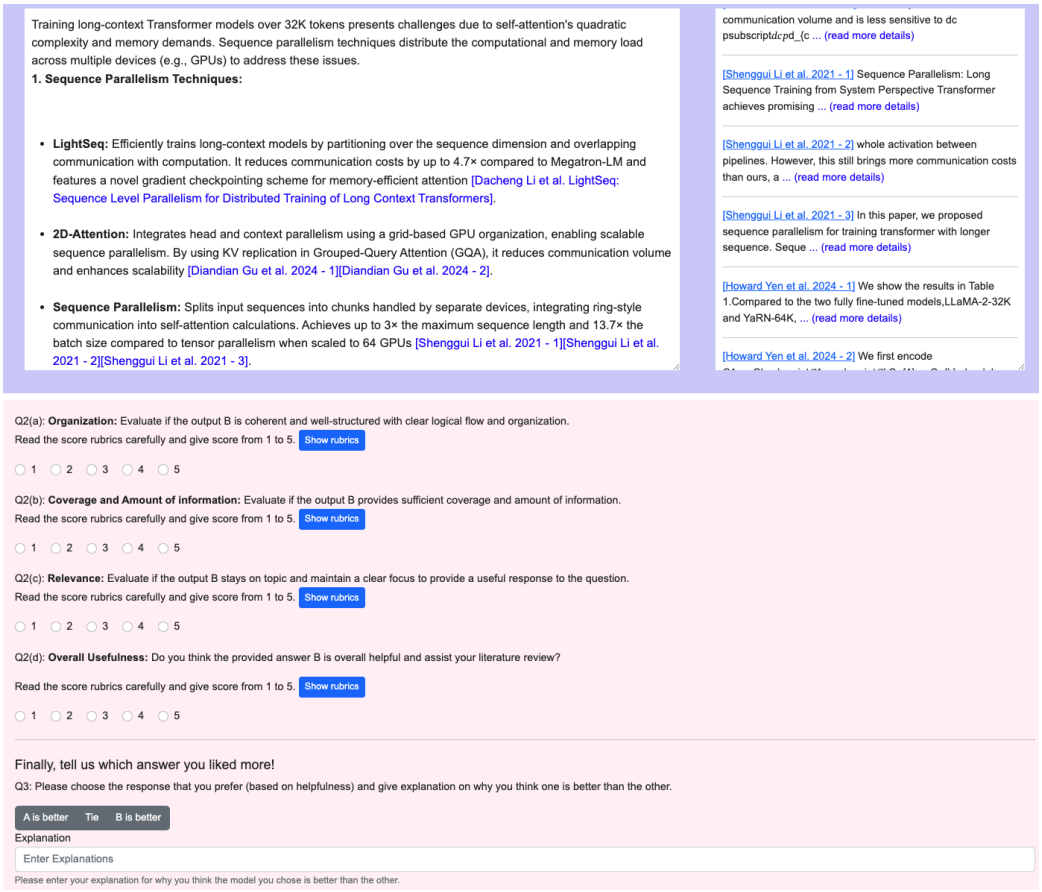

Supplementary Figure SI.5: **Generator LM training data distribution.** We combine training data from diverse sources in addition to synthetically generated data.




Supplementary Figure SI.6: **Publication year distribution of the top 20 retrieved papers for SCHOLARQA-CS.** By updating the datastore from peS2o v2 to the more recent peS2o v3, which includes papers until October 2024, our dense retriever can successfully retrieve more recent papers.




Supplementary Figure SI.7: **Human evaluation annotation interface (top).** We show two model responses in our interface. For each response, we also show the cited papers (text chunks) so that our expert annotators can verify the response.




Supplementary Figure SI.8: **Human evaluation annotation interface (bottom).** Our expert annotators first give five-scale scores for each aspect, and then, after evaluating each response carefully, they give pairwise preference of the two systems.


Supplementary Figure SI.9: **OpenScholar Public Demo:** We released the first of it's kind public demo of OpenScholar, where users can ask questions freely and OpenScholar provides responses accompanied with extensive citations instantly.

## 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

Supplementary Figure SI.10: **Single-task examples in SCHOLARQABENCH.**

## 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.

Supplementary Figure SI.11: 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**

Supplementary Figure SI.12: 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**

Supplementary Figure SI.13: 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**

Supplementary Figure SI.14: 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, m1 and m2, 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 ν ≡ (m1*m2)/(m1 + m2)^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**

Supplementary Figure SI.15: An example of SCHOLARQA-MULTI.

# Supplementary References

## REFERENCES


Ron Artstein and Massimo Poesio. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008. doi: 10.1162/coli.07-034-R2. URL https://aclanthology.org/J08-4004/.

Akari Asai and Eunsol Choi. Challenges in information-seeking QA: Unanswerable questions and paragraph retrieval. In *ACL*, 2021. URL https://aclanthology.org/2021.acl-long.118.

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 *ICLR*, 2024. URL https://openreview.net/forum?id=4WnqRR915j.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In *EMNLP*, Brussels, Belgium, 2018. Association for Computational Linguistics. URL https://aclanthology.org/D18-1241.

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.

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

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

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. URL https://arxiv.org/abs/2311.10702.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *EMNLP-IJCNLP*, 2019. URL https://aclanthology.org/D19-1259.

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 *ICLR*, 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.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. QASA: advanced question answering on scientific articles. In *ICML*, 2023. URL https://proceedings.mlr.press/v202/lee23n.html.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2819–2834, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.251/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

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. URL https://arxiv.org/abs/2409.04109.

Amanpreet Singh, Joseph Chee Chang, Dany Haddad, Aakanksha Naik, Jena D. Hwang, Rodney Kinney, Daniel S. Weld, Doug Downey, and Sergey Feldman. Ai2 scholar qa: Organized literature synthesis with attribution. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2025. URL https://api.semanticscholar.org/CorpusID:280529654.

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*, 2023. URL https://www.nature.com/articles/s41586-023-06291-2.

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.

Aivin V Solatorio. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*, 2024.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS (Datasets and Benchmarks)*, 2021. URL https://openreview.net/forum?id=wCu6T5xFjeJ.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *EMNLP*, 2020. 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. In *NeurIPS (Datasets and Benchmarks Track)*, 2024. URL https://arxiv.org/abs/2406.07835.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models, 2024. URL https://openreview.net/forum?id=u6jbcaCHqO.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packaged resources to advance general chinese embedding, 2023. URL https://arxiv.org/abs/2309.07597.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering. In *ACL*, 2023. URL https://aclanthology.org/2023.acl-long.181.

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 *Findings of ACL*, 2024. URL https://aclanthology.org/2024.findings-acl.770.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. Automatic evaluation of attribution by large language models. In *Findings of EMNLP*, 2023. URL https://aclanthology.org/2023.findings-emnlp.307.

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.