# OpenScholar:
# Retrieval-augmented Language Models for Scientific Literature Synthesis

**Akari Asai,** Rulin Shao, Jacqueline He, Weijia Shi, Amanpreet Singh,
Kyle Lo, Dave Wadden, Luca Soldaini, Sergey Feldman,
Joseph Chang, Mike D'arcy, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong,
Bohao Wu, Yanyu Xiong,
Graham Neubig, Luke Zettlemoyer, Dan Weld, Doug Downey,
Wen-tau Yih, Hannaneh Hajishirzi, Pang Wei Koh




Will be released in November 12th!

https://tinyurl.com/akariopenscholar

# LLMs Got Much More Powerful as They Scale

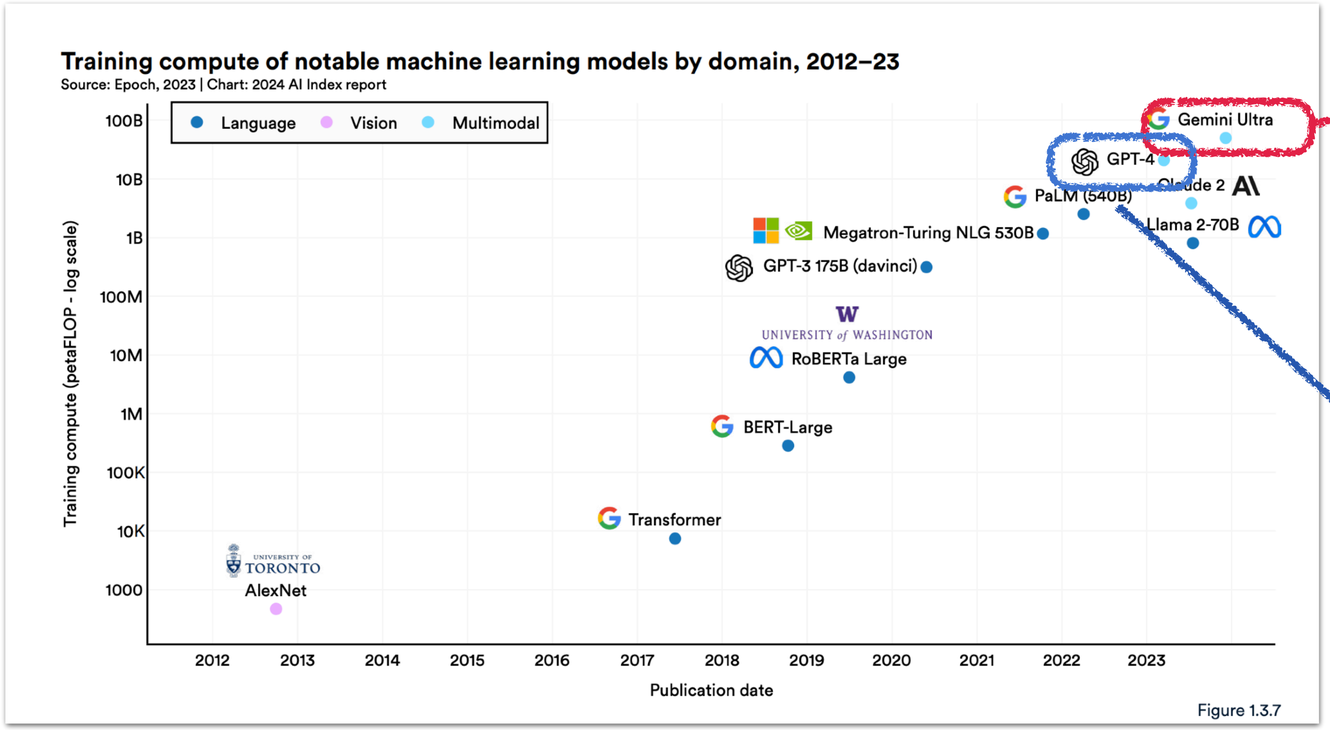
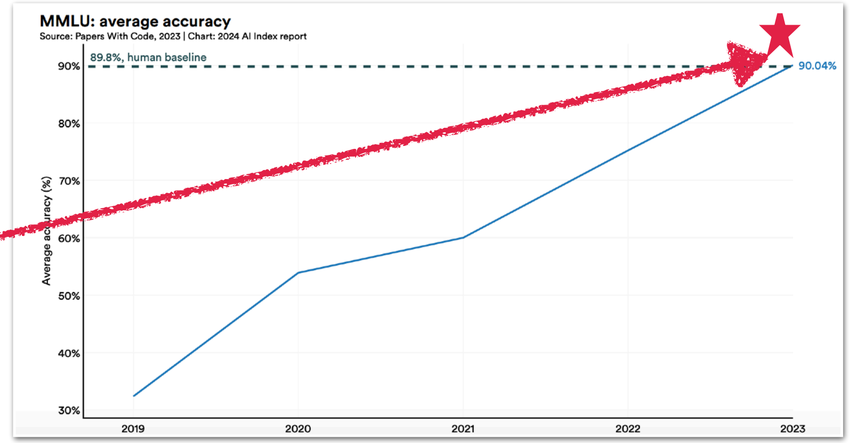
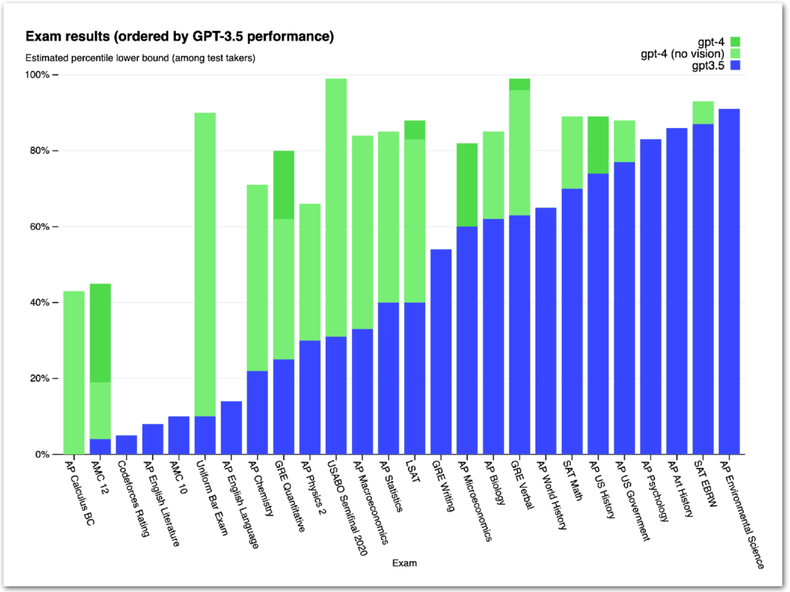



Figure 1.3.7




Stanford University HAI. Artificial Intelligence Index Report 2024.

Open AI. GPT-4 (OpenAI Blog). 2023.

# Hallucinations in LM outputs

Explain Akari Asai, including her representative papers

(*GPT4o, October 29, 2024)

Akari Asai is a rising researcher in the field of **Natural Language Processing (NLP)** and **Machine Learning (ML)**, particularly known for her work on **open-domain question answering**, **multi-hop reasoning**, and **retrieval-based language models**. She completed her PhD at the University of Washington under the guidance of Professor Hannaneh Hajishirzi. Her research aims to develop models and systems that allow machines to answer complex queries by reasoning over large collections of text and retrieving relevant information efficiently.

Here are some of Akari Asai's **representative papers** that highlight her significant contributions:

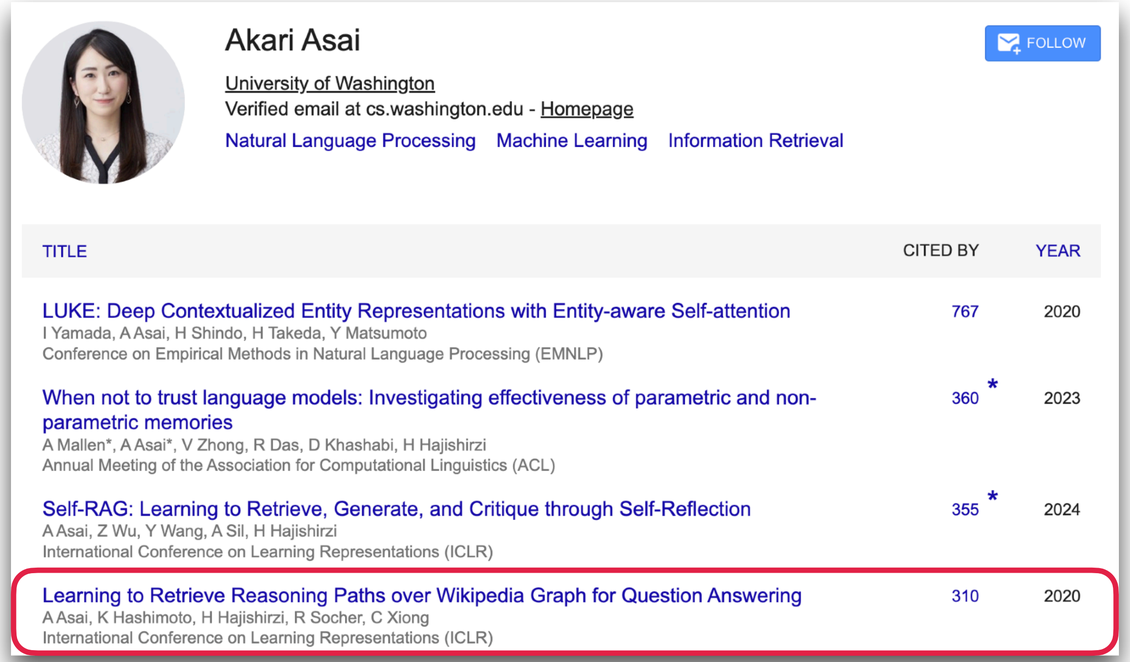
### 2. "Unsupervised Dense Information Retrieval with Contrastive Learning" (2021)

**Conference**: North American Chapter of the Association for Computational Linguistics ~~(NAACL~~ 2021)

TMLR 2022

**Summary**:

In this paper, Asai and her co-authors propose an **unsupervised approach** for training dense retrievers using **contrastive learning**. The method leverages large-scale unlabeled data to improve the performance of information retrieval systems without requiring annotated datasets. Contrastive learning helps the model to distinguish between relevant and irrelevant passages, improving the quality of retrieval in a range of QA tasks.



Akari Asai

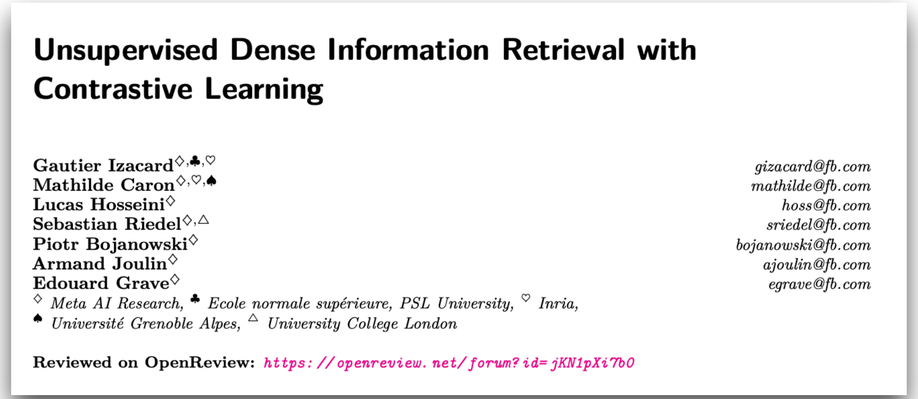
FOLLOW

University of Washington

Verified email at cs.washington.edu - Homepage

Natural Language Processing  Machine Learning  Information Retrieval

| TITLE | CITED BY | YEAR |
|---|---|---|
| LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention<br>I Yamada, A Asai, H Shindo, H Takeda, Y Matsumoto<br>Conference on Empirical Methods in Natural Language Processing (EMNLP) | 767 | 2020 |
| When not to trust language models: Investigating effectiveness of parametric and non-parametric memories<br>A Mallen*, A Asai*, V Zhong, R Das, D Khashabi, H Hajishirzi<br>Annual Meeting of the Association for Computational Linguistics (ACL) | 360 * | 2023 |
| Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection<br>A Asai, Z Wu, Y Wang, A Sil, H Hajishirzi<br>International Conference on Learning Representations (ICLR) | 355 * | 2024 |
| Learning to Retrieve Reasoning Paths over Wikipedia Graph for Question Answering<br>A Asai, K Hashimoto, H Hajishirzi, R Socher, C Xiong<br>International Conference on Learning Representations (ICLR) | 310 | 2020 |



**Unsupervised Dense Information Retrieval with Contrastive Learning**


Gautier Izacard[◇,♣,♡] gizacard@fb.com
Mathilde Caron[◇,♡,♠] mathilde@fb.com
Lucas Hosseini[◇] hoss@fb.com
Sebastian Riedel[◇,△] sriedel@fb.com
Piotr Bojanowski[◇] bojanowski@fb.com
Armand Joulin[◇] ajoulin@fb.com
Edouard Grave[◇] egrave@fb.com
[◇] *Meta AI Research,* [♣] *Ecole normale supérieure, PSL University,* [♡] *Inria,*
[♠] *Université Grenoble Alpes,* [△] *University College London*



**Reviewed on OpenReview:** *https://openreview.net/forum?id=jKN1pXi7b0*

# Catastrophic Errors as Results of LM Hallucinations

TECH · LAW

## Humiliated lawyers fined $5,000 for submitting ChatGPT hallucinations in court: 'I heard about this new site, which I falsely assumed was, like, a super search engine'

BY RACHEL SHIN
June 23, 2023 at 9:41 AM PDT


Lawyers who filed legal documents with false citations generated by ChatGPT have been fine

ERIK MCGREGOR—LIGHTROCKET/GETTY IMAGES

MIT Technology Review

Featured Topics Newsletters Events Podcasts SIGN IN

ARTIFICIAL INTELLIGENCE

## Why Meta's latest large language model survived only three days online

Galactica was supposed to help scientists. Instead, it mindlessly spat out biased and incorrect nonsense.

By Will Douglas Heaven November 18, 2022

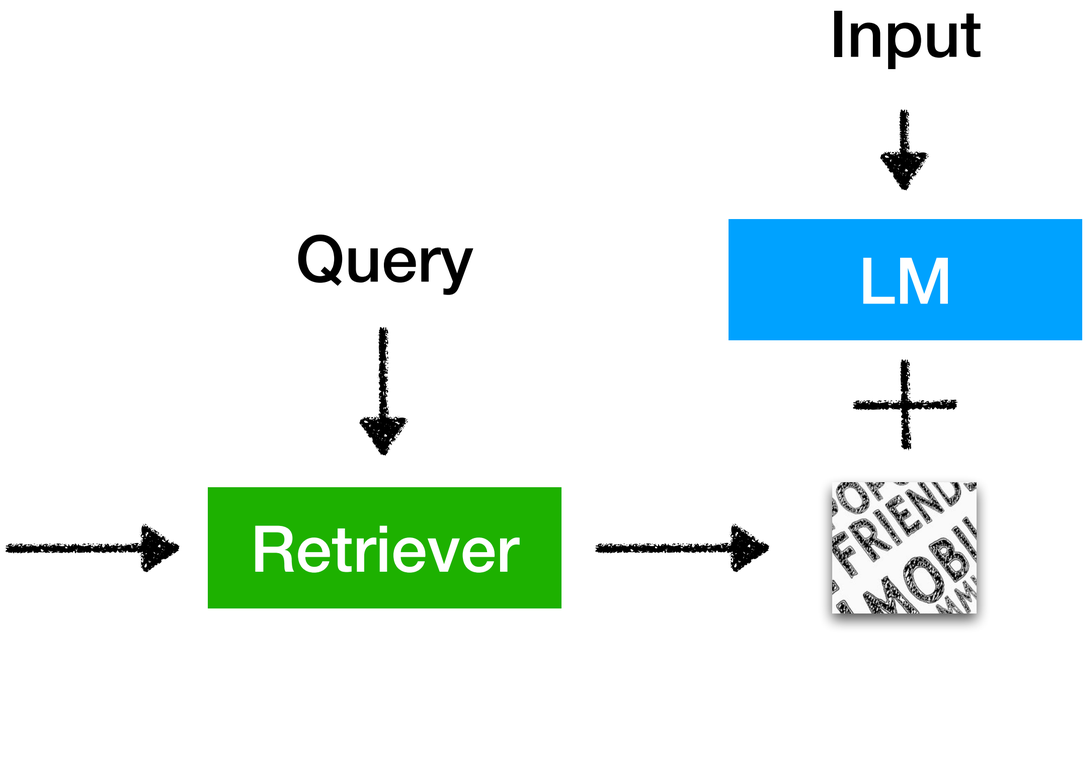
## Air Canada must honor re… invented by airline's chatb…

Air Canada appears to have quietly killed its costly chatbot support.

ASHLEY BELANGER - 2/16/2024, 12:12 PM

# Retrieval-augmented LMs (RALM)

Datastore




Learn more about retrieval-augmented LMs? Check out our ACL 2023 tutorial

https://acl2023-retrieval-lm.github.io/ by Akari Asai, Sewon Min, Zexuan Zhong, Danqi Chen

# Retrieval-augmented generations (RAG)

How did US states get their names?

**Retriever (e.g., Google, BM 25)**

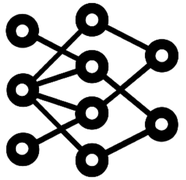
❶ Of the fifty states, eleven including New York, Georgia, Washington named after an individual person.

❷ UTAH: Name taken from the Ute people who inhabited that region

❸ The history of human activity in Michigan began with settlement by Paleo-Indians.

Retrieve

**Answer my question using references.**
**References:** ❶ ❷ ❸
**Question:** How did US states get their names?



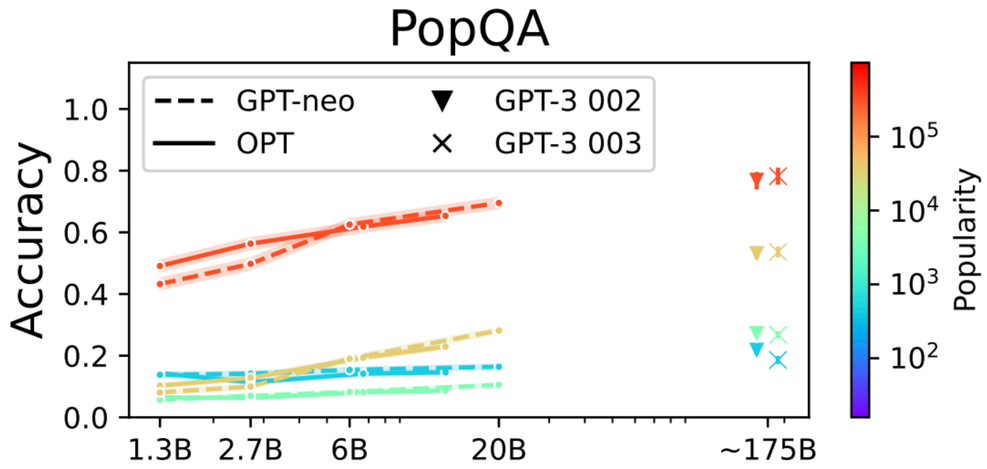
**LMs (e.g., GPT-4, Llama 2)**

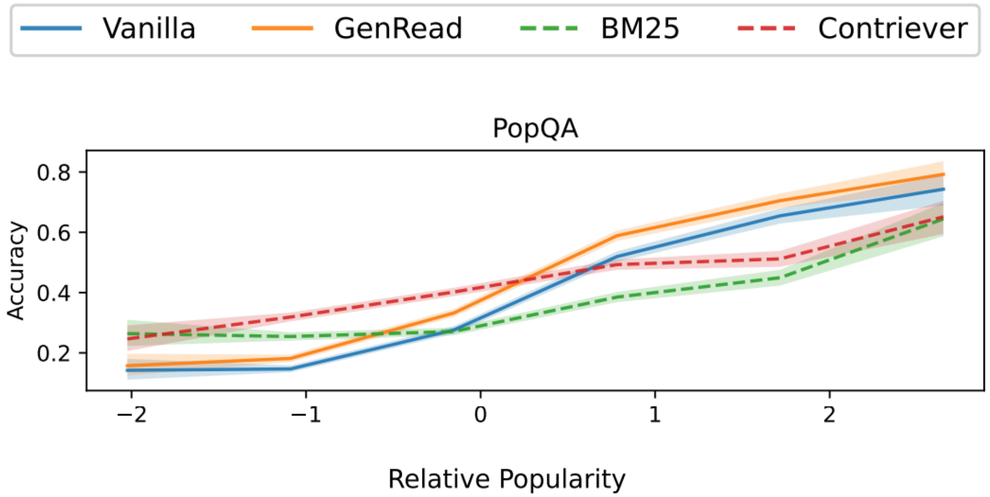
Eleven states are named after an individual person. Some states including Utah are named after native American tribe names.

Read

# Benefits of Retrieval-augmented LMs

Reduced hallucinations in long tail




Asai* and Mallen* et al. "When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories". *In ACL. 2023.*

# Benefits of Retrieval-augmented LMs

Reduced hallucinations in long tail

Knowledge updates w.o re-training

Q: How many home runs has Shohei Ohtani hit?
A: 24

Q: How many home runs has Shohei Ohtani hit?
A: **25**

| Real-time Baselines | | | Multi-choice | | Generation | |
|---|---|---|---|---|---|---|
| | Retrieve | Predict | Orig. | NOTA | EM | F1 |
| Open | DPR | RAG | 27.4 | 24.8 | 2.4 | 4.1 |
| | DPR | GPT-3 | 43.9 | 35.8 | 13.3 | 19.7 |
| | GCS | RAG | 46.9 | 37.9 | 17.5 | 22.1 |
| | GCS | GPT-3 | **66.5** | **58.4** | **34.6** | **45.3** |
| Closed | — | T5 | 39.1 | 35.3 | 9.7 | 14.7 |
| | — | GPT-3 | 44.9 | 34.1 | 15.3 | 22.3 |

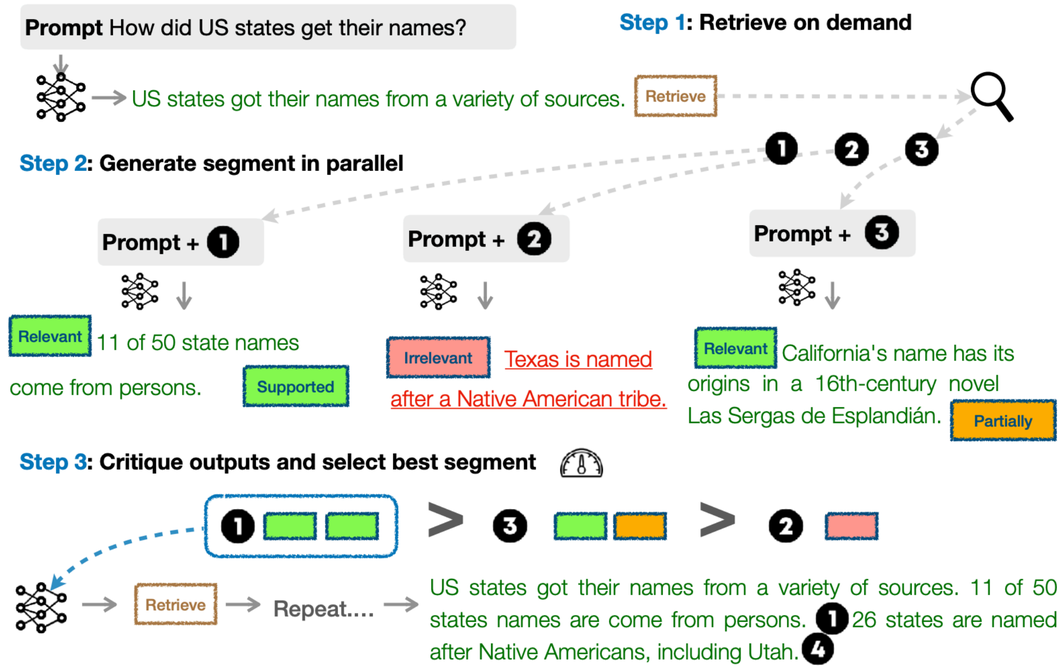
Kasai, ... Asai, et al. "REALTIME QA: What's the Answer Right Now?". *In NeurIPS (Dataset and Benchmark). 2023.*

# Benefits of Retrieval-augmented LMs

Reduced hallucinations in long tail

Knowledge updates w.o re-training

Attributions to evidence text




Asai et al. "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection". *In ICLR. 2024.*

Akari Asai is a rising researcher in the field of **Natural Language Processing (NLP)** and **Machine Learning (ML)**, particularly known for her work on **open-domain question answering**, **multi-hop reasoning**, and **retrieval-based language models**. She completed her PhD at the University of Washington under the guidance of Professor Hannaneh Hajishirzi. Her research [illegible] develop models and systems that allow machines to answer complex queries by reaso[illegible]ge collections of text and retrieving relevant information efficiently.

Here are some of Akari Asai's **representative papers** that highlight her significant contributions:

### 2. "Unsupervised Dense Information Retrieval with Contrastive Learning" (2021)

**Conference**: North American Chapter of the Association for Computational Linguistics (NAACL 2021)

**Summary**:
In this paper, Asai and her co-authors propose an **unsupervised approach** for training dense retrievers using **contrastive learning**. The method leverages large-scale unlabeled data to improve the performance of information retrieval systems without requiring annotated datasets. Contrastive learning helps the model to distinguish between relevant and irrelevant passages, improving the quality of retrieval in a range of QA tasks.

# RALMs are Now Widely Used in Academia and Industries

Reduced hallucinations in long tail

Knowledge updates w.o re-training

Attributions to evidence text

RAG-based search systems


Libraries for customized RAG


# RALMs Are Particularly Useful for Expert Domains

Reduced hallucinations in long tail

Knowledge updates w.o re-training

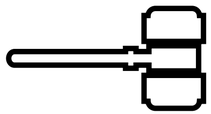
Attributions to evidence text

Medical

Engineering

Finance

Science



Legal

# RALMs Are Particularly Useful for Expert Domains

Reduced hallucinations in long tail

Knowledge updates w.o re-training

Attributions to evidence text



Medical

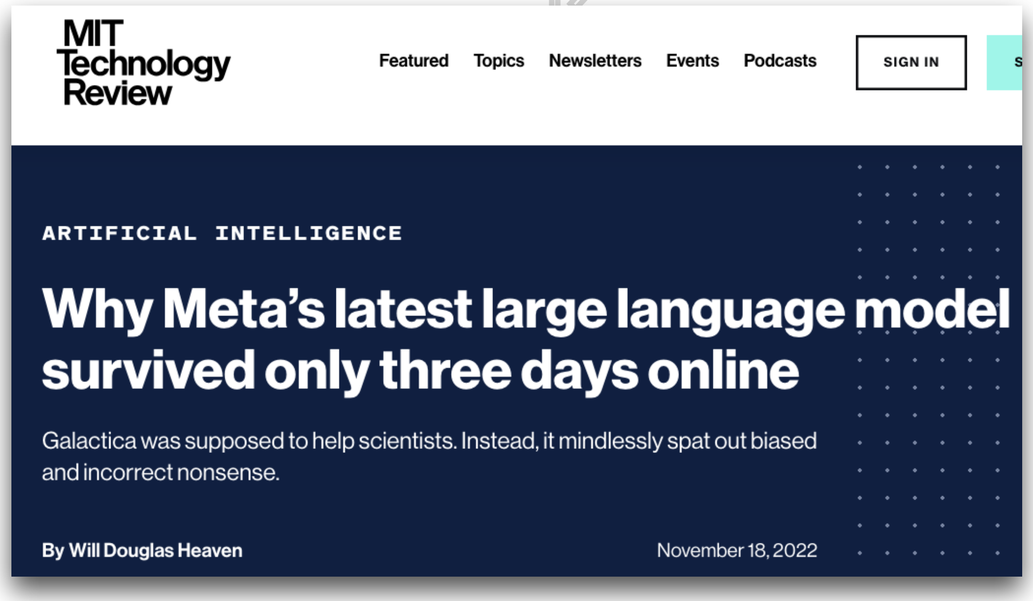


Engineering



Finance

Science



# RALMs Are Particularly Useful for Expert Domains

Reduced hallucinations in long tail

Knowledge updates w.o re-training

Attributions to evidence text

Medical

Engineering

Finance

TECH · LAW

**Humiliated lawyers fined $5,000 for submitting ChatGPT hallucinations in court: 'I heard about this new site, which I falsely assumed was, like, a super search engine'**

BY RACHEL SHIN
June 23, 2023 at 9:41 AM PDT


Lawyers who filed legal documents with false citations generated by ChatGPT have been fined.
ERIK MCGREGOR—LIGHTROCKET/GETTY IMAGES

Legal

# OpenScholar:
# Retrieval-augmented Language Models for Scientific Literature Synthesis

**Akari Asai,** Rulin Shao, Jacqueline He, Weijia Shi, Amanpreet Singh,
Kyle Lo, Dave Wadden, Luca Soldaini, Sergey Feldman,
Joseph Chang, Mike D'arcy, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong,
Bohao Wu, Yanyu Xiong,
Graham Neubig, Luke Zettlemoyer, Dan Weld, Doug Downey,
Wen-tau Yih, Hannaneh Hajishirzi, Pang Wei Koh

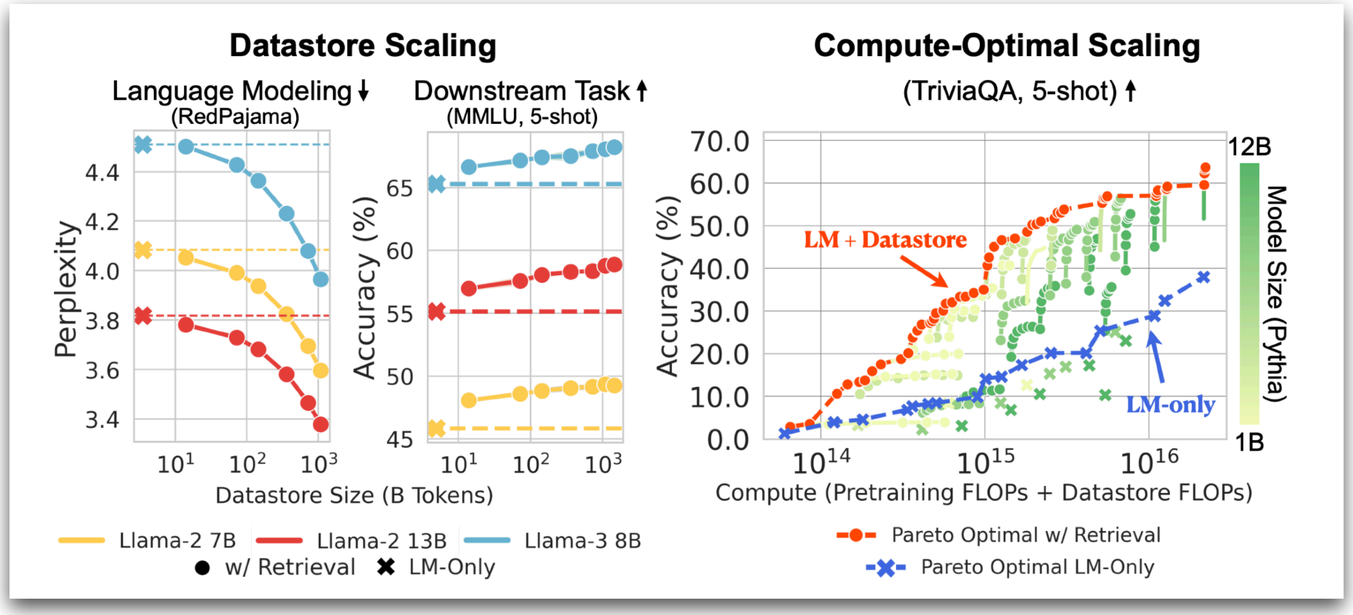



Will be released in November 12th!

https://tinyurl.com/akariopenscholar

# Can LLMs Assist Scientific Literature Synthesis?

Has anyone showed the effectiveness of scaling retrieval-augmented LMs' retrieval tokens to trillion-token scale?



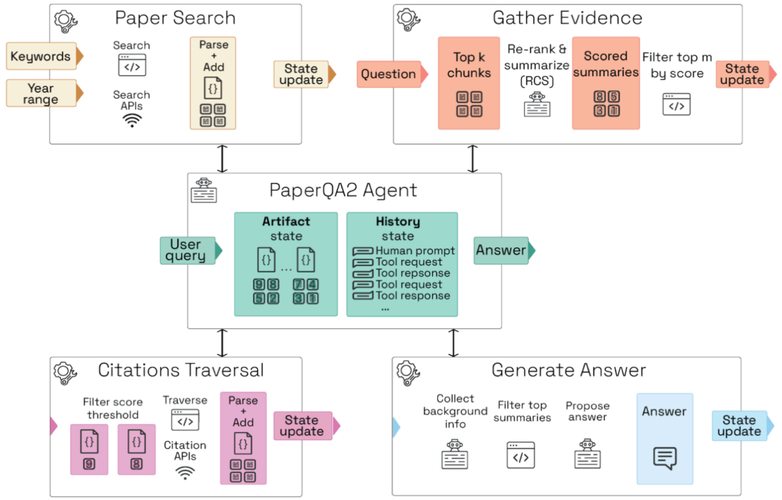

Shao, He, **Asai** et al. "Scaling Retrieval-Based Language Models with a Trillion-Token Datastore". *In NeurIPS. 2024.*

# Can LLMs Assist Scientific Literature Synthesis?

Has anyone showed the effectiveness of scaling retrieval-augmented LMs' retrieval tokens to trillion-token scale?

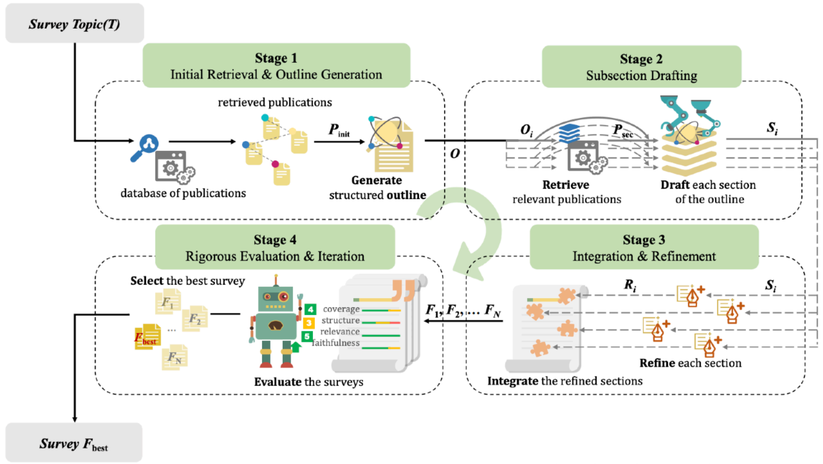
**As of my last update, there hasn't been a published study** explicitly demonstrating the effectiveness of scaling retrieval-augmented language models (LMs) to the trillion-token scale specifically in terms of retrieval tokens. ....

# Related Work




PaperQA2
(Skarlinski et al., 2024)

Depending on proprietary LMs (e.g., GPT4o, Claude)



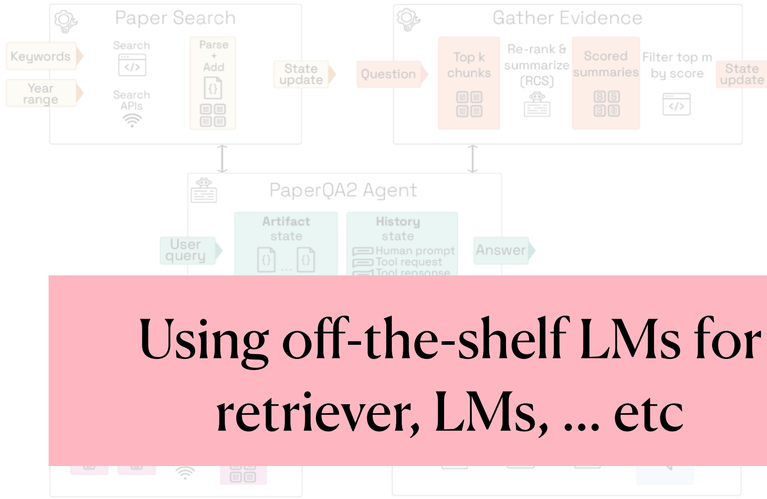

AutoSurvey
(Wang et al., 2024)

Datastores are either small (~500k) or private

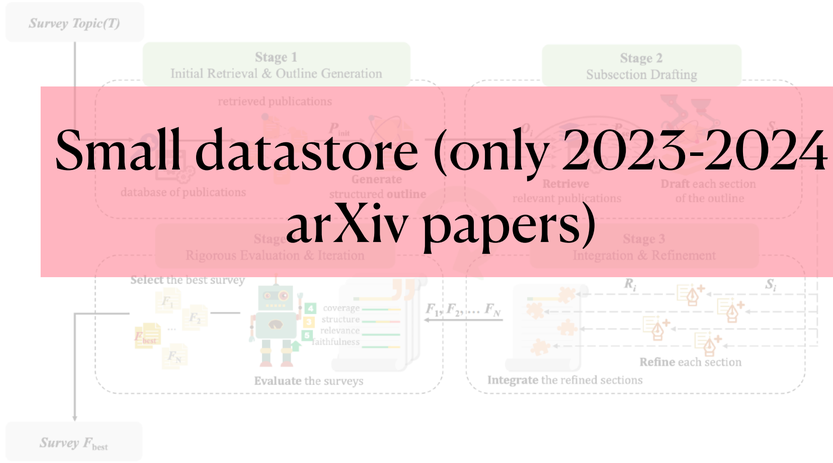

Skarlinski et al. “Language agents achieve superhuman synthesis of scientific knowledge”. *Arxiv*. 2024.

Wang et al. “AutoSurvey: Large Language Models Can Automatically Write Surveys”. In *NuerIPS*. 2024.

# Related Work

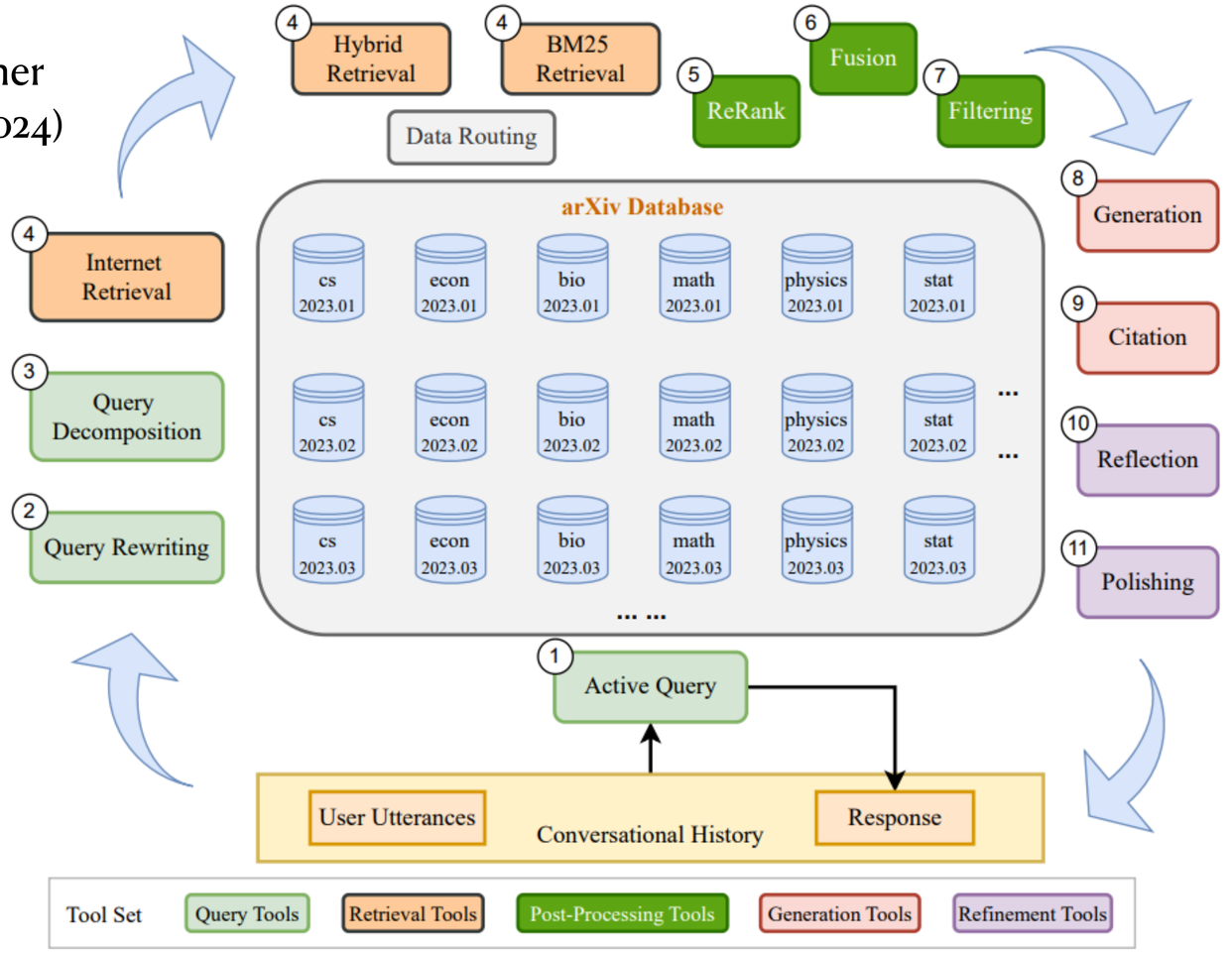



Using off-the-shelf LMs for retriever, LMs, ... etc



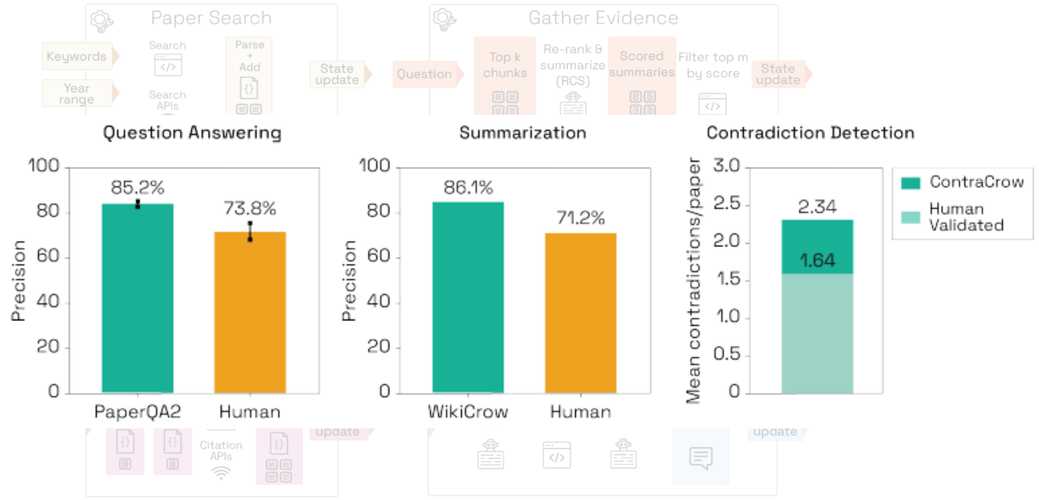

Small datastore (only 2023-2024 arXiv papers)

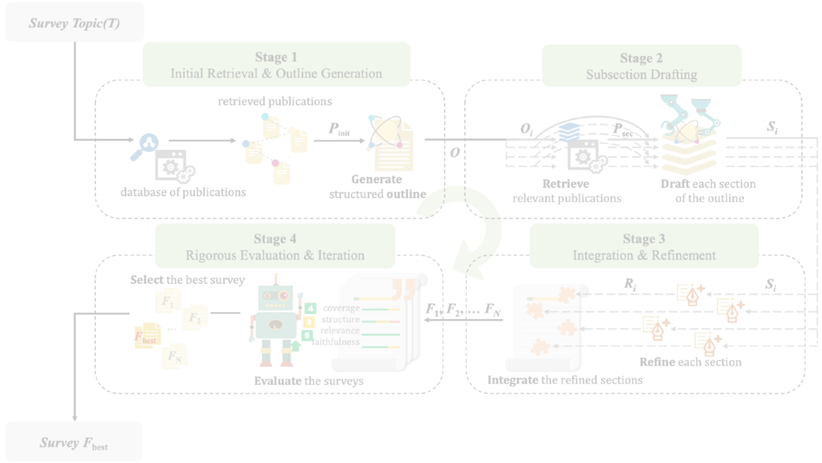
OpenResearcher (Zheng et al., 2024)

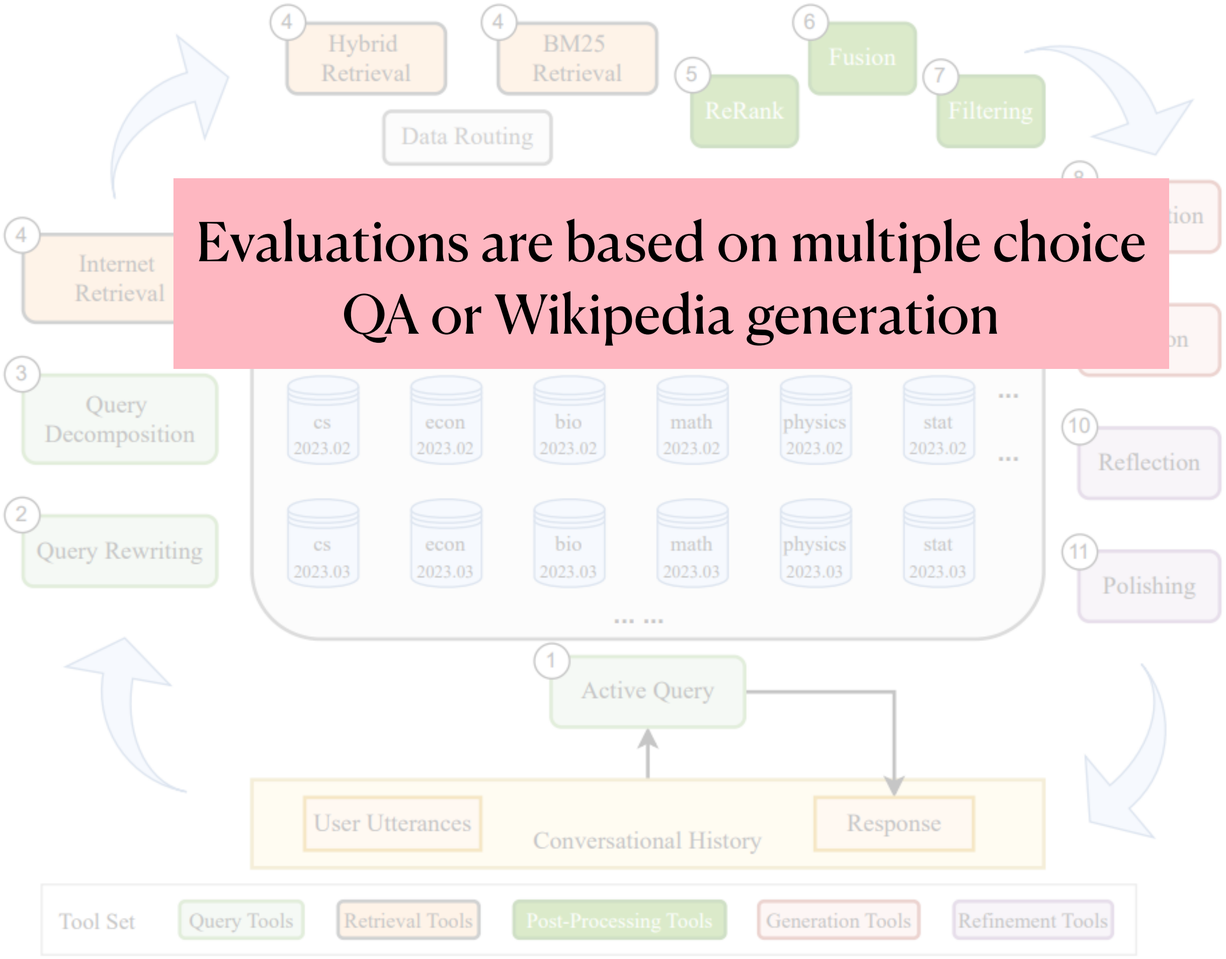



Zheng et al. “OpenResearcher: Unleashing AI for Accelerated Scientific Research”. *Arxiv*. 2024.

# Related Work

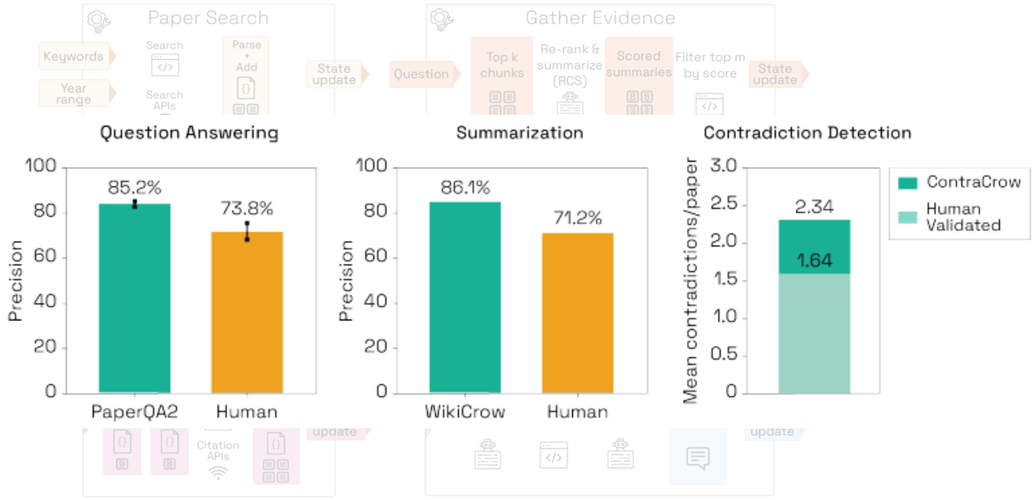


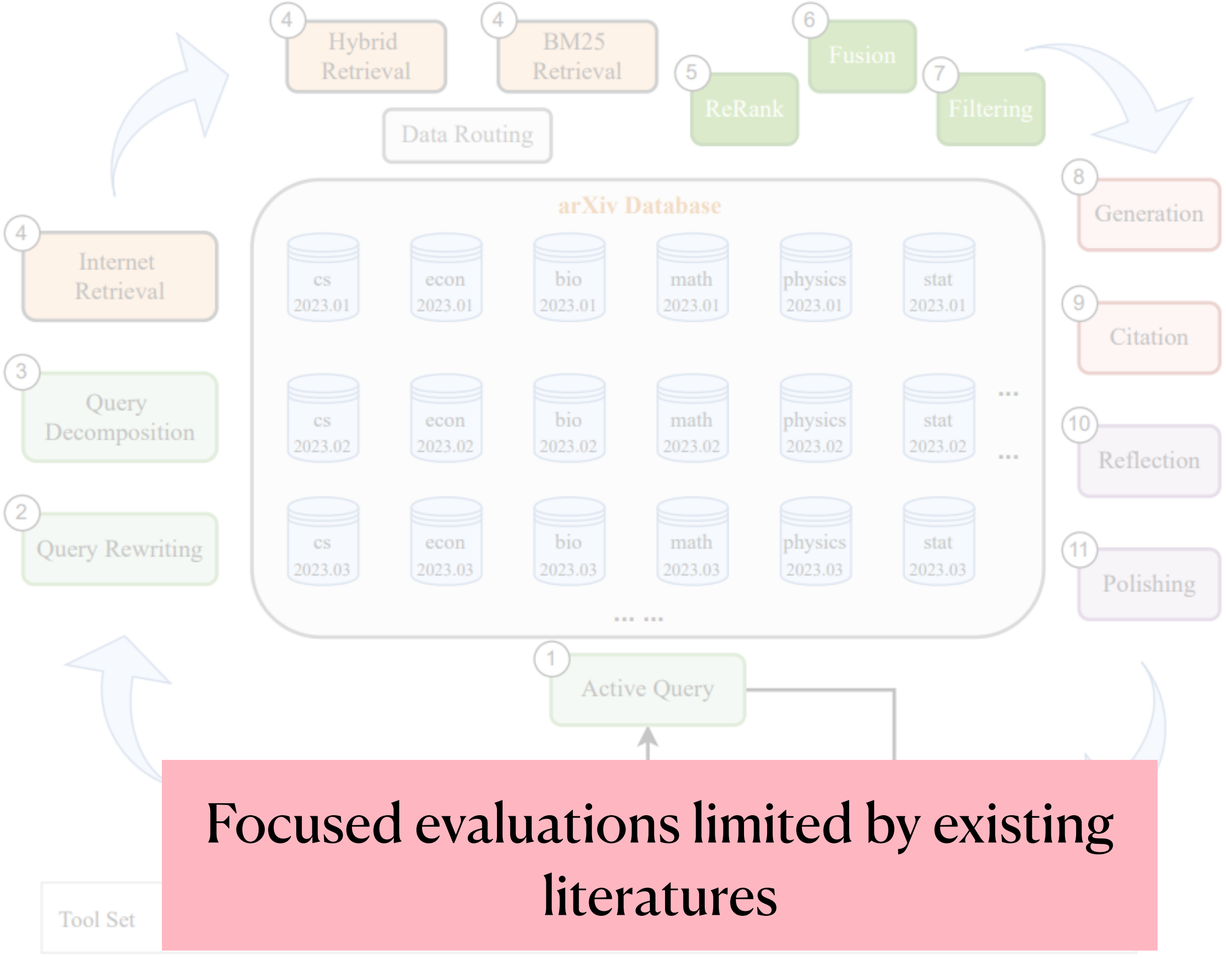




Evaluations are based on multiple choice QA or Wikipedia generation

# Related Work




Focused evaluations limited by existing literatures

*Survey Topic(T)*

Table 2: Results of naive RAG-based LLM generation, Human writing and AutoSurvey. Note that AutoSurvey and naive RAG-based LLM generation both use Claude-haiku as the writer. **Note that human writing surveys used for evaluation are excluded during the retrieval process.**

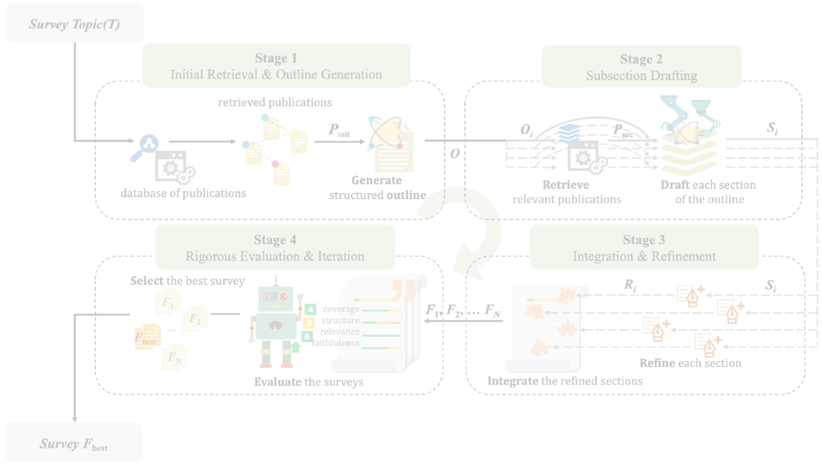
| Survey Length (#tokens) | Methods | Speed | Citation quality | | Content Quality | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Recall | Precision | Coverage | Structure | Relevance | Avg. |
| 8k | Human writing | 0.16 | 80.00 | 87.50 | 4.50 | 4.16 | 5.00 | 4.52 |
| | Naive RAG-based LLM generation | 79.67 | $78.14_{\pm 5.23}$ | $71.92_{\pm 6.83}$ | $4.40_{\pm 0.48}$ | $3.86_{\pm 0.71}$ | $4.86_{\pm 0.33}$ | 4.33 |
| | AutoSurvey | 107.00 | $82.48_{\pm 2.77}$ | $77.42_{\pm 3.28}$ | $4.60_{\pm 0.48}$ | $4.46_{\pm 0.49}$ | $4.8_{\pm 0.39}$ | 4.61 |
| 16k | Human writing | 0.14 | 88.52 | 79.63 | 4.66 | 4.38 | 5.00 | 4.66 |
| | Naive RAG-based LLM generation | 43.41 | $71.48_{\pm 12.50}$ | $65.31_{\pm 15.36}$ | $4.46_{\pm 0.49}$ | $3.66_{\pm 0.69}$ | $4.73_{\pm 0.44}$ | 4.23 |
| | AutoSurvey | 95.51 | $81.34_{\pm 3.65}$ | $76.94_{\pm 1.93}$ | $4.66_{\pm 0.47}$ | $4.33_{\pm 0.59}$ | $4.86_{\pm 0.33}$ | 4.60 |
| 32k | Human writing | 0.10 | 88.57 | 77.14 | 4.66 | 4.50 | 5.00 | 4.71 |
| | Naive RAG-based LLM generation | 22.64 | $79.88_{\pm 4.35}$ | $65.03_{\pm 8.39}$ | $4.41_{\pm 0.64}$ | $3.75_{\pm 0.72}$ | $4.66_{\pm 0.47}$ | 4.23 |
| | AutoSurvey | 91.46 | $83.14_{\pm 2.44}$ | $78.04_{\pm 3.14}$ | $4.73_{\pm 0.44}$ | $4.26_{\pm 0.69}$ | $4.8_{\pm 0.54}$ | 4.58 |
| 64k | Human writing | 0.07 | 86.33 | 77.78 | 5.00 | 4.66 | 5.00 | 4.88 |
| | Naive RAG-based LLM generation | 12.56 | $68.79_{\pm 11.00}$ | $61.97_{\pm 13.45}$ | $4.4_{\pm 0.61}$ | $3.66_{\pm 0.47}$ | $4.66_{\pm 0.47}$ | 4.19 |
| | AutoSurvey | 73.59 | $82.25_{\pm 3.64}$ | $77.41_{\pm 3.84}$ | $4.73_{\pm 0.44}$ | $4.33_{\pm 0.47}$ | $4.86_{\pm 0.33}$ | 4.62 |

*Survey $F_{\text{best}}$*

# Related Work




Really small scale human evaluation (~30 queries in CS)



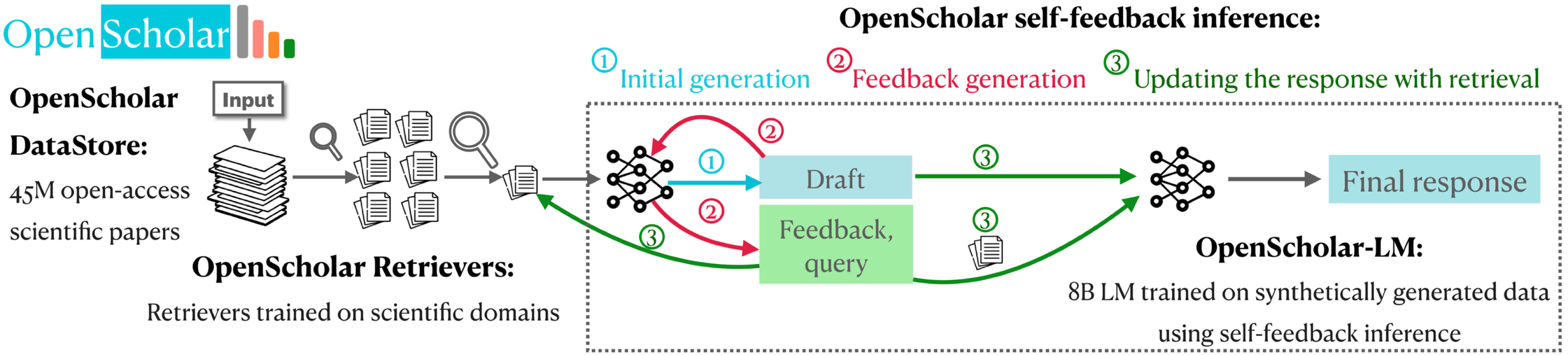

| Models | Correctness | | | Richness | | | Relevance | | |
|---|---|---|---|---|---|---|---|---|---|
| | Win | Tie | Lose | Win | Tie | Lose | Win | Tie | Lose |
| Ask | 2 | 16 | 12 | 12 | 6 | 12 | 2 | 8 | 20 |
| You.com | 3 | 21 | 6 | 9 | 5 | 16 | 4 | 13 | 13 |
| Phind | 2 | 26 | 2 | 15 | 7 | 8 | 5 | 13 | 12 |
| Naive RAG | 1 | 22 | 7 | 14 | 8 | 8 | 5 | 16 | 9 |
| OpenResearcher | **10** | 13 | 7 | **25** | 4 | 1 | **15** | 13 | 2 |

Table 1: Human Preference compared with Perplexity AI outcome. "Win" means that the current method beats Perplexity AI. More "Win" times means a superior application.

| Models | Richness | | | Relevance | | |
|---|---|---|---|---|---|---|
| | Win | Tie | Lose | Win | Tie | Lose |
| iAsk | 42 | 0 | 67 | 38 | 0 | 71 |
| You.com | 15 | 0 | 94 | 16 | 0 | 93 |
| Phind | 52 | 1 | 56 | 54 | 0 | 55 |
| Naive RAG | 41 | 1 | 67 | 57 | 0 | 52 |
| OpenResearcher | **62** | 2 | 45 | **74** | 0 | 35 |

Table 2: GPT-4o Preference Results compared with Perplexity AI outcome.

# OpenScholar: RALMs for Scientific Literature Synthesis




Largest open datastore

Open retriever, reranker and 8B LMs for science

Self-feedback retrieval-augmented generation

# OpenScholar: RALMs for Scientific Literature Synthesis

Largest multi-domain benchmark

Expert-annotated realistic research queries and LF answers

Reliable multi-faced evaluation

ScholarBench

**Data:** 2.2k new expert-written questions & 200 answers in computer science, biomedicine, neuroscience and physics

**Evaluation:** Reproducible multi-faced **model-based** and **human evaluation**

**Input**

What are the recent research advancements in enhancing fluorescence for biosensing using photonic crystal?

**Expert-written answer**

Recent advancements in photonic crystal (PC)-enhanced fluorescence for biosensing have significantly improved ... For instance, a 60-fold increase in fluorescence intensity was achieved using a one-dimensional PC slab with a spatial gradient structure for a surface-attached Cy-5 organic dye layer [1].

**References** [1] Leaky-mode assisted fluorescence extraction: application to fluorescence enhancement biosensors

Evaluated by model | Evaluated by model and humans | Evaluated by humans

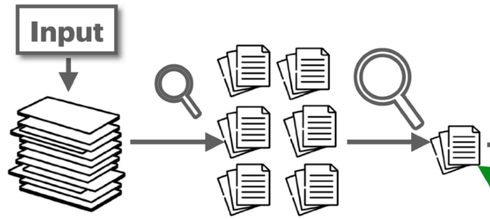
Accuracy | Citations | Coverage | Relevance | Organization | Usefulness

# OpenScholar: RALMs for Scientific Literature Synthesis

OpenScholar

**OpenScholar DataStore:**

45M open-access scientific papers

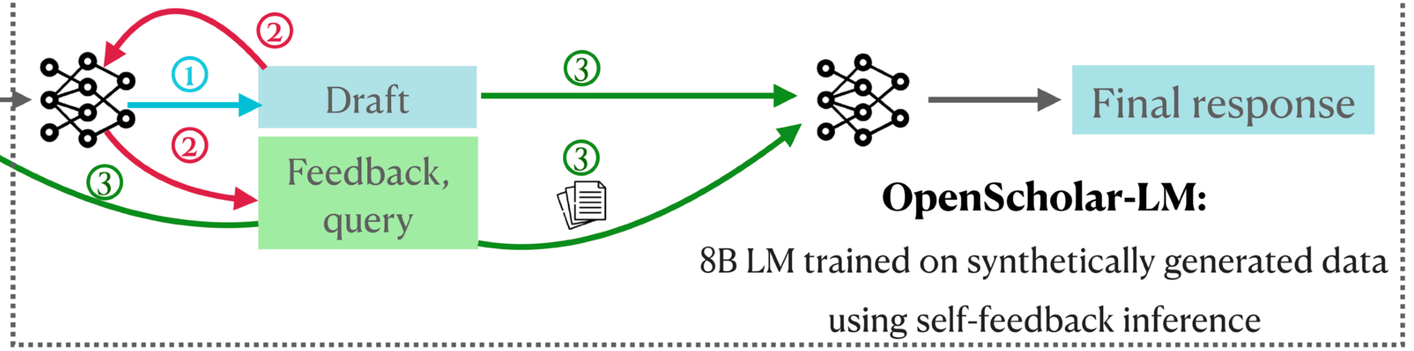



**OpenScholar Retrievers:**

Retrievers trained on scientific domains

**OpenScholar self-feedback inference:**

① Initial generation ② Feedback generation ③ Updating the response with retrieval




**OpenScholar-LM:**

8B LM trained on synthetically generated data using self-feedback inference

ScholarBench

**Data:** 2.2k new expert-written questions & 200 answers in computer science, biomedicine, neuroscience and physics

**Input**

What are the recent research advancements in enhancing fluorescence for biosensing using photonic crystal?

**Expert-written answer**

Recent advancements in photonic crystal (PC)-enhanced fluorescence for biosensing have significantly improved ... For instance, a 60-fold increase in fluorescence intensity was achieved using a one-dimensional PC slab with a spatial gradient structure for a surface-attached Cy-5 organic dye layer [1].

**References** [1] Leaky-mode assisted fluorescence extraction: application to fluorescence enhancement biosensors

**Evaluation:** Reproducible multi-faced **model-based** and **human evaluation**

Evaluated by model | Evaluated by model and humans | Evaluated by humans

Accuracy | Citations | Coverage | Relevance | Organization | Usefulness

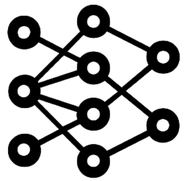
# OpenScholar: Three Main Components

## OpenScholar DataStore:

45M open-access scientific papers; 200+M passages

## OpenScholar Retrievers:

Retriever and ranker trained on scientific domains



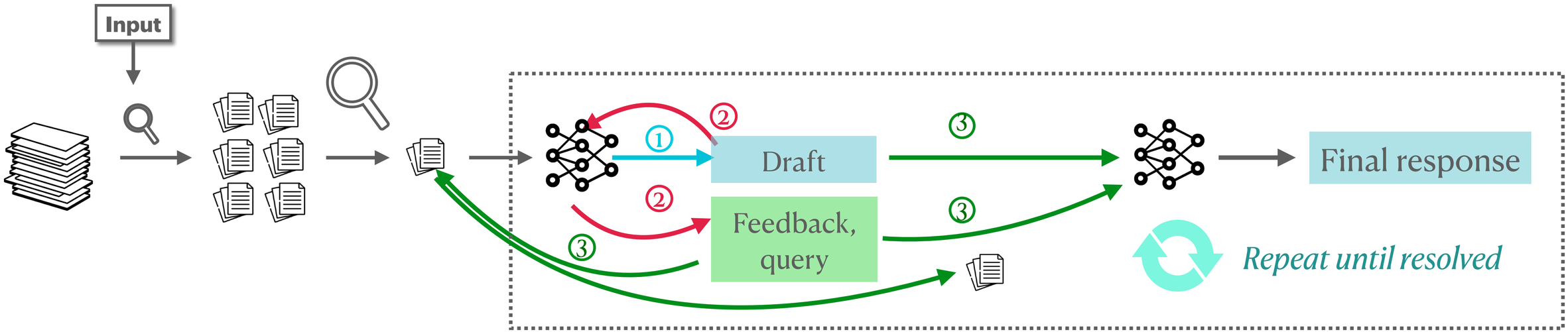
## OpenScholar-LM:

8B LM trained on synthetically generated data using self-feedback inference  (or proprietary LMs)

# OpenScholar: Inference Pipeline

**OpenScholar DataStore** | **OpenScholar Retrievers** | **OpenScholar LM**

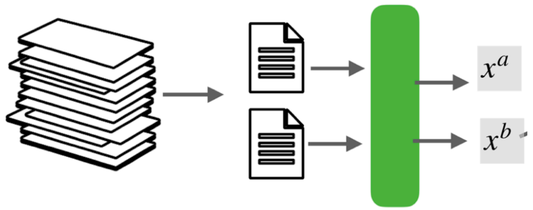
How can we train an efficient & open LM?



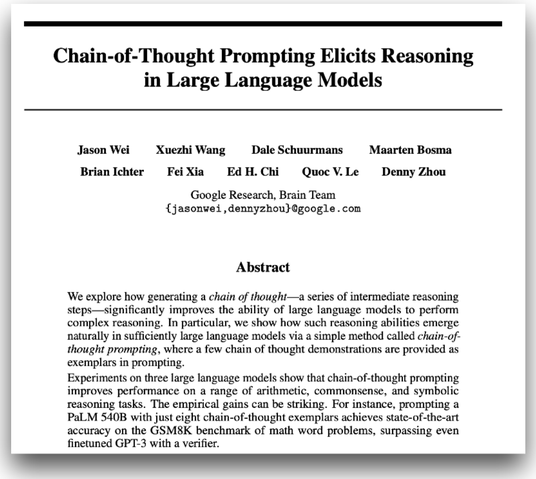

**OpenScholar self-feedback inference:**

① Initial generation ② Feedback generation ③ Updating the response with retrieval

# OpenScholar: Training Data Creation for OpenScholar LM




Sample top cited papers



**Chain-of-Thought Prompting Elicits Reasoning in Large Language Models**

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou
Google Research, Brain Team
{jasonwei,dennyzhou}@google.com

**Abstract**

We explore how generating a *chain of thought*—a series of intermediate reasoning steps—significantly improves the ability of large language models to perform complex reasoning. In particular, we show how such reasoning abilities emerge naturally in sufficiently large language models via a simple method called *chain-of-thought prompting*, where a few chain of thought demonstrations are provided as exemplars in prompting.

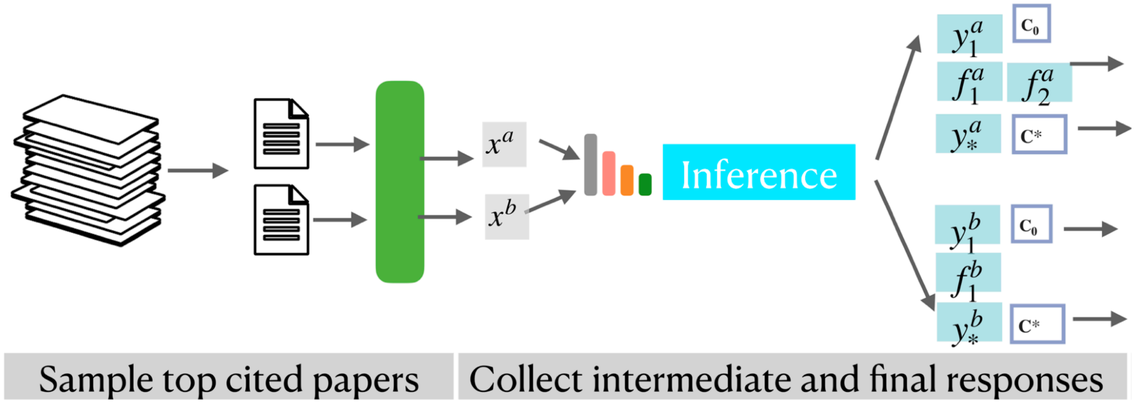
Experiments on three large language models show that chain-of-thought prompting improves performance on a range of arithmetic, commonsense, and symbolic reasoning tasks. The empirical gains can be striking. For instance, prompting a PaLM 540B with just eight chain-of-thought exemplars achieves state-of-the-art accuracy on the GSM8K benchmark of math word problems, surpassing even finetuned GPT-3 with a verifier.

*On which tasks has chain-of-thought prompting been shown to be effective so far?*

*List limitations of chain-of-thought prompting, with previous empirical findings.*

*Is CoT prompting still useful even when generated reasoning steps are incorrect?*

# OpenScholar: Training Data Creation for OpenScholar LM

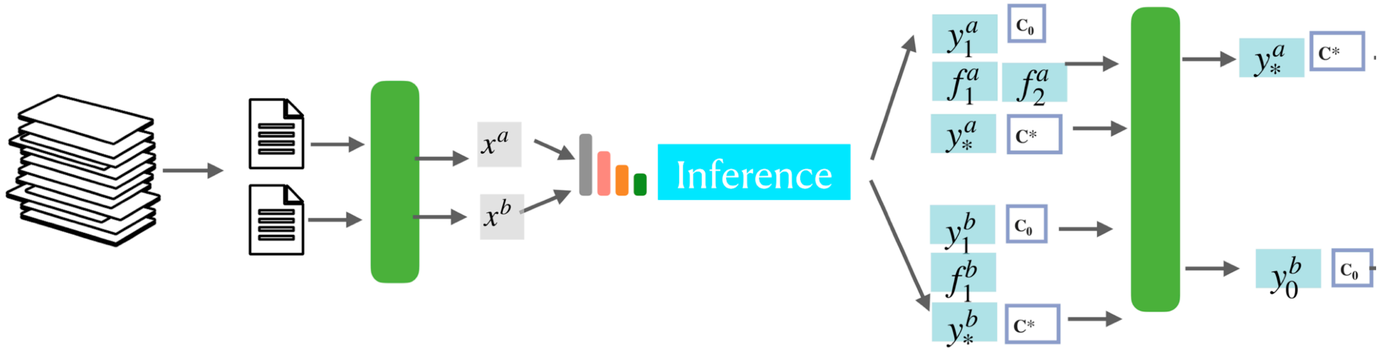



***Initial response without feedback:*** *Chain-of-Thought prompting has shown to be effective on a variety of tasks, particularly in logical and mathematical reasoning tasks, as well as common sense reasoning, as shown in the original paper [1].*

***Final response:*** *Chain-of-Thought prompting has shown to be effective on a variety of tasks, particularly in logical and mathematical reasoning tasks, as well as common sense reasoning, as shown in the original paper [1]. Moreover, many followup studies show that it's effective on other tasks such as translation [2], fact-based QA [3], and summarization [4].*

***Feedback:*** *Your answer should discuss empirical results not only just from the original paper, but also from other followup papers*

# OpenScholar: Training Data Creation for OpenScholar LM



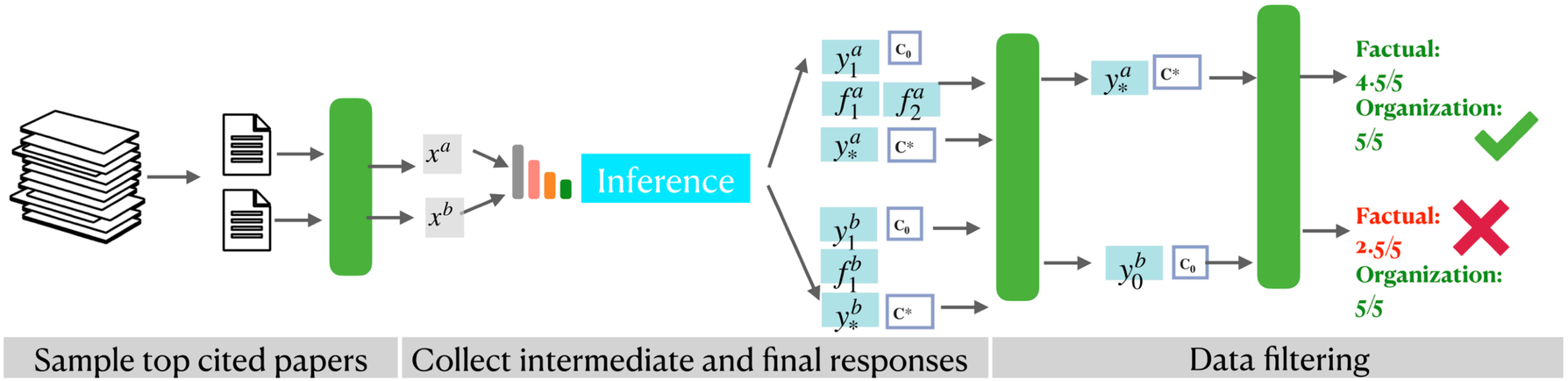

Sample top cited papers | Collect intermediate and final responses | Data filtering | Data mixing and training

***Initial response without feedback:*** *Chain-of-Thought prompting has shown to be effective on a variety of tasks, particularly in logical and mathematical reasoning tasks, as well as common sense reasoning, as shown in the original paper [1].*

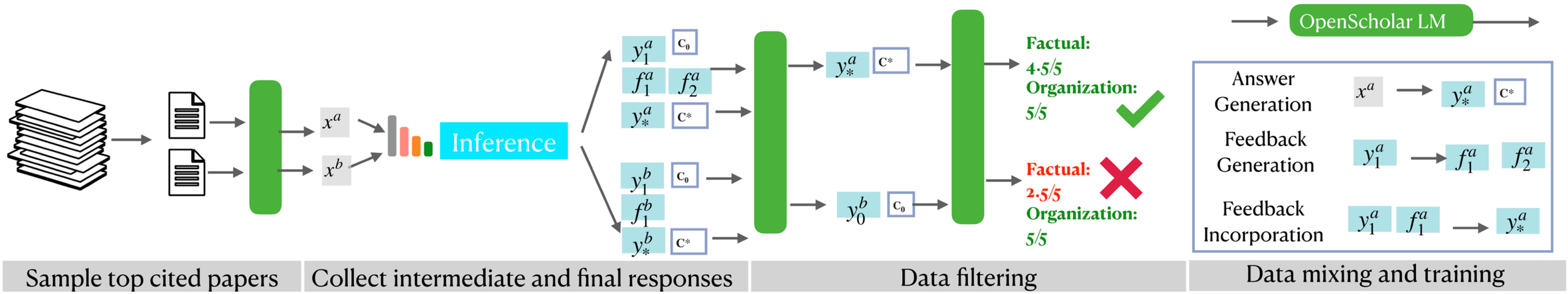
***Final response:*** *Chain-of-Thought prompting has shown to be effective on a variety of tasks, particularly in logical and mathematical reasoning tasks, as well as common sense reasoning, as shown in the original paper [1]. Moreover, many followup studies show that it's effective on other tasks such as translation [2], fact-based QA [3], and summarization [4].*

***Feedback:*** *Your answer should discuss empirical results not only just from the original paper, but also from other followup papers*

# OpenScholar: Training Data Creation for OpenScholar LM



# OpenScholar: Training Data Creation for OpenScholar LM




*On which tasks has chain-of-thought prompting been shown to be effective so far?*

***Final response:*** *Chain-of-Thought prompting has shown to be effective on a variety of tasks, particularly in logical and mathematical reasoning tasks, as well as common sense reasoning, as shown in the original paper [1]. Moreover, many followup studies show that it's effective on other tasks such as translation [2], fact-based QA [3], and summarization [4].*

***Initial response without feedback:*** *Chain-of-Thought prompting has shown to be effective on a variety of tasks, particularly in logical and mathematical reasoning tasks, as well as common sense reasoning, as shown in the original paper [1].*

***Feedback:*** *Your answer should discuss empirical results not only just from the original paper, but also from other followup papers*

# OpenScholar: RALMs for Scientific Literature Synthesis




ScholarBench

**Data:** 2.2k new expert-written questions & 200 answers in computer science, biomedicine, neuroscience and physics

**Input**

What are the recent research advancements in enhancing fluorescence for biosensing using photonic crystal?

**Expert-written answer**

Recent advancements in photonic crystal (PC)-enhanced fluorescence for biosensing have significantly improved ... For instance, a 60-fold increase in fluorescence intensity was achieved using a one-dimensional PC slab with a spatial gradient structure for a surface-attached Cy-5 organic dye layer [1].

**References** [1] Leaky-mode assisted fluorescence extraction: application to fluorescence enhancement biosensors

**Evaluation:** Reproducible multi-faced **model-based** and **human evaluation**

Evaluated by model | Evaluated by model and humans | Evaluated by humans

Accuracy | Citations | Coverage | Relevance | Organization | Usefulness

# Overview of ScholarBench

**Data:** 2.2k new expert-written questions & 200 answers in CS, Bio, Physics, and neuroscience

**Input**

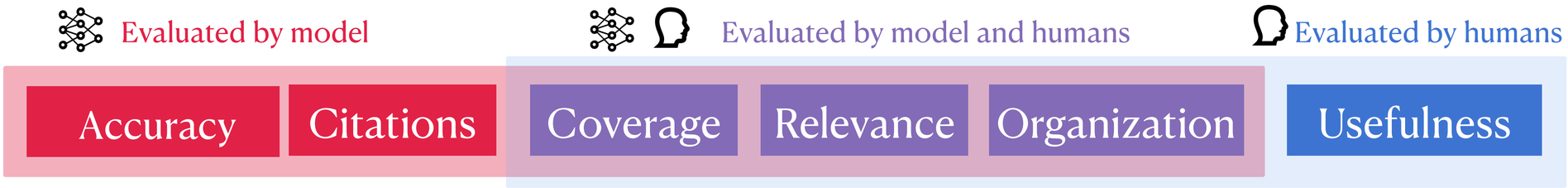
What are the recent research advancements in enhancing fluorescence for biosensing using photonic crystal?

**Expert-written answer**

Recent advancements in photonic crystal (PC)-enhanced fluorescence for biosensing have significantly improved ... For instance, a 60-fold increase in fluorescence intensity was achieved using a one-dimensional PC slab with a spatial gradient structure for a surface-attached Cy-5 organic dye layer [1].

**References** [1] Leaky-mode assisted fluorescence extraction: application to fluorescence enhancement biosensors

**Evaluation:** Reproducible multi-faced **model-based** and **human evaluation**



# ScholarBench Datasets

| | Task | Domain | Output format | Gold reference answer | Multi-paper task? |
|---|---|---|---|---|---|
| SciFact | Fact verification | BioMed | Classification (closed) | yes | |
| Pubmed QA | Boolean QA | BioMed | Classification (closed) | yes | |
| QASA | Paper understanding QA | CS (AI, ML) | Long-form generation | yes | |
| Scholar-CS | Lit. review QA | CS (All) | Long-form generation | yes | yes |
| Scholar-Multi | Lit. review QA | CS, BioMed, Physics | Long-form generation | yes | yes |
| Scholar-Neuro, Multi | Lit. review QA | NeuroScience / Bio Med | Long-form generation | | yes |

# ScholarBench Datasets

| | Task | Domain | Output format | Gold reference answer | Multi-paper task? |
|---|---|---|---|---|---|
| SciFact | Fact verification | BioMed | Classification (closed) | yes | |
| Pubmed QA | Boolean QA | BioMed | Classification (closed) | yes | |
| QASA | Paper understanding QA | CS (AI, ML) | Long-form generation | yes | |
| Scholar-CS | Lit. review QA | CS (All) | Long-form generation | yes | yes |
| Scholar-Multi | Lit. review QA | CS, BioMed, Physics | Long-form generation | yes | yes |
| Scholar-Neuro, Multi | Lit. review QA | NeuroScience / Bio Med | Long-form generation | | yes |

New 2.2k queries and 200+ expert-written answers across multiple areas

# Scholar-CS: a New Approach for Long-form Evaluation

100 queries across CS areas with avg. 4.4 “rubrics”

**Input**

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

**Rubrics**

**Must Have Item-1:** The answer should discuss best practices that can be used to prevent these implications, such as reliable source, update monitoring, code review, minimizing dependencies, sandboxing, security testing, and team

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

# ScholarBench Automatic Evaluation Pipeline

**Input**

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

**Rubric**

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

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

**Output**

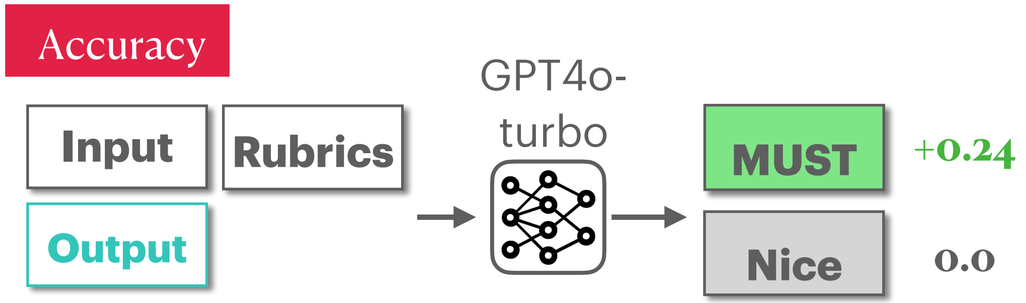
Must have item-1 is included ✓

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

**Citations**

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

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



# ScholarBench Automatic Evaluation Pipeline

**Input**

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

**Rubric**

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

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

**Output**

Must have item-1 is included ✓

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

**Citations**

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

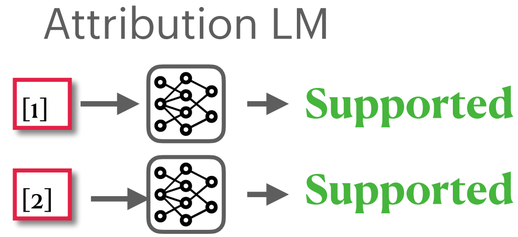
[2] Applying formal verification methods to examine the security properties of ... can also be part of library test suites or continuous integration in order to run




Citation

Automated Vulnerability Detection and Mitigation Tools [1]: Automated

Formal verification and Fuzzing [2]: Applying formal verification



# ScholarBench Automatic Evaluation Pipeline

**Input**

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

**Rubric**

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

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

**Output**

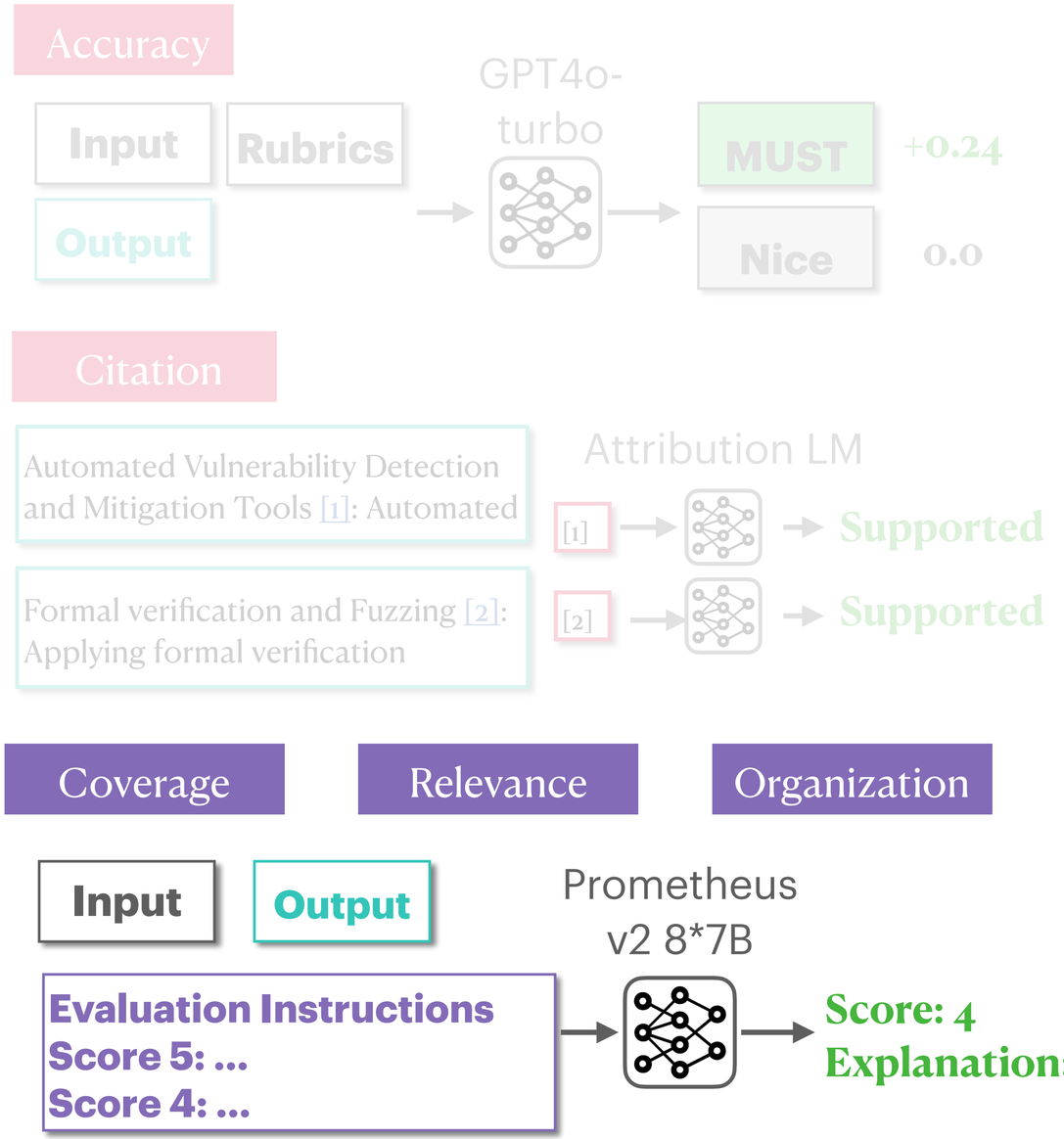
Must have item-1 is included ✓

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

**Citations**

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

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



# Experiments

**Base LMs**

- OS-8B
- Llama 3.1 8B, 70B
- GPT4o

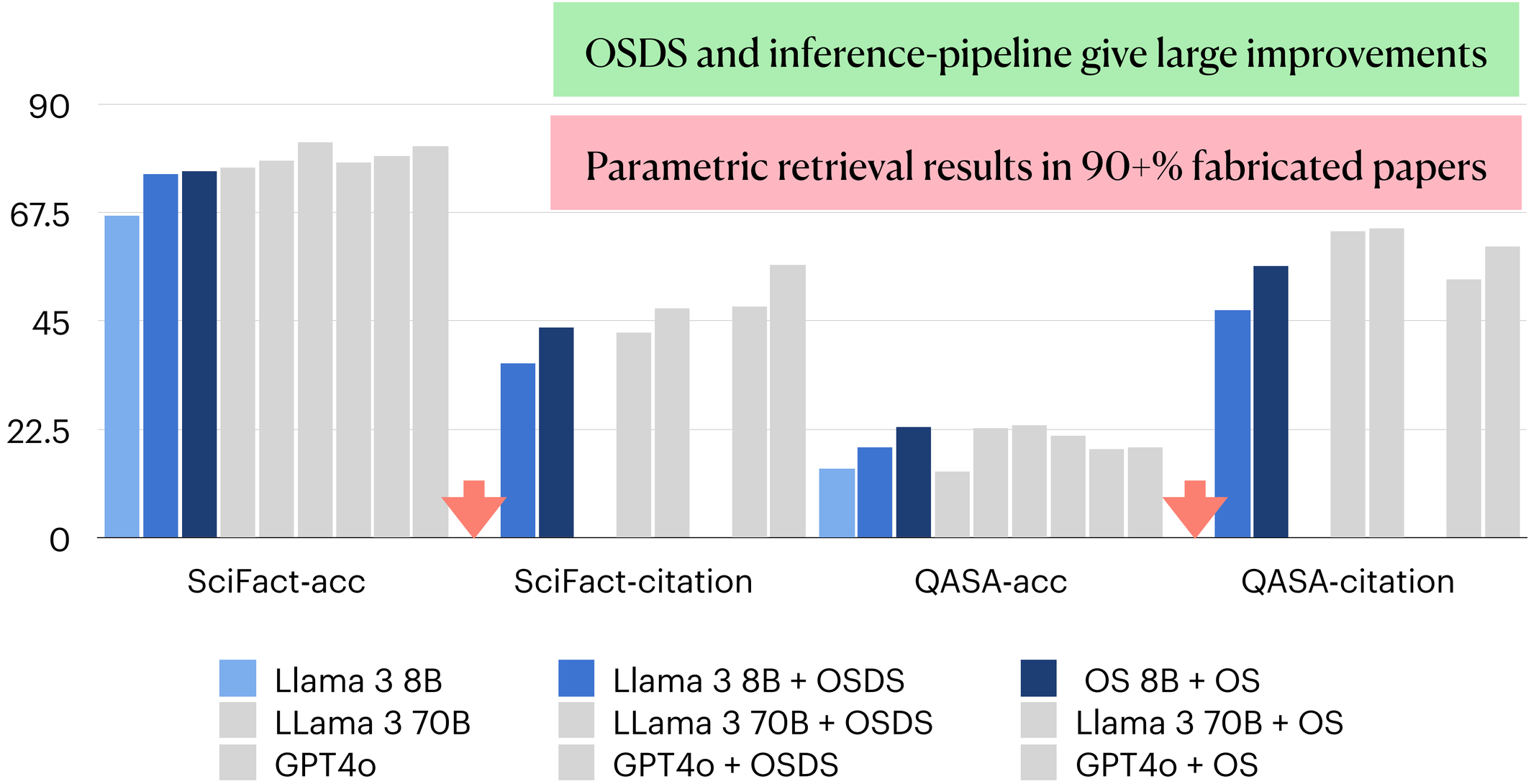
**Inference pipeline**

- No retrieval (generate citations from parametric memories)
- Standard RAG w/ OpenScholar Datastore Top 10
- OpenScholar inference

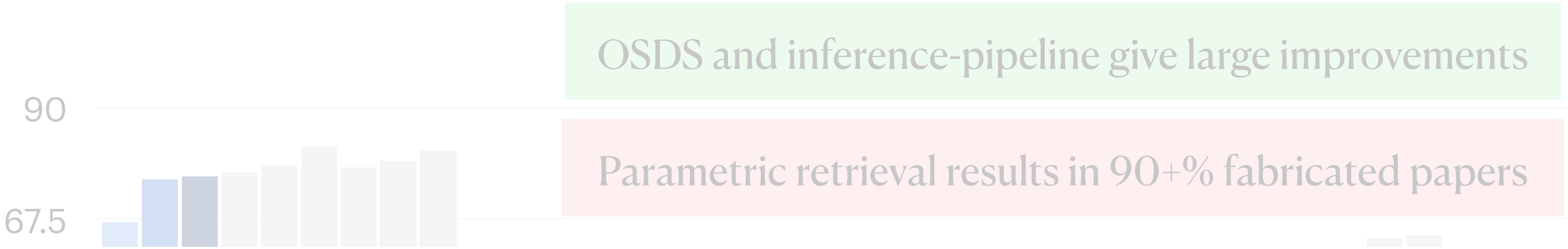
**Proprietary systems**

- PaperQA2 (Skarlinski et al., 2024); using GPT4o for reranking, generation … etc
- Perplexity Pro

# Results on Single-paper Tasks



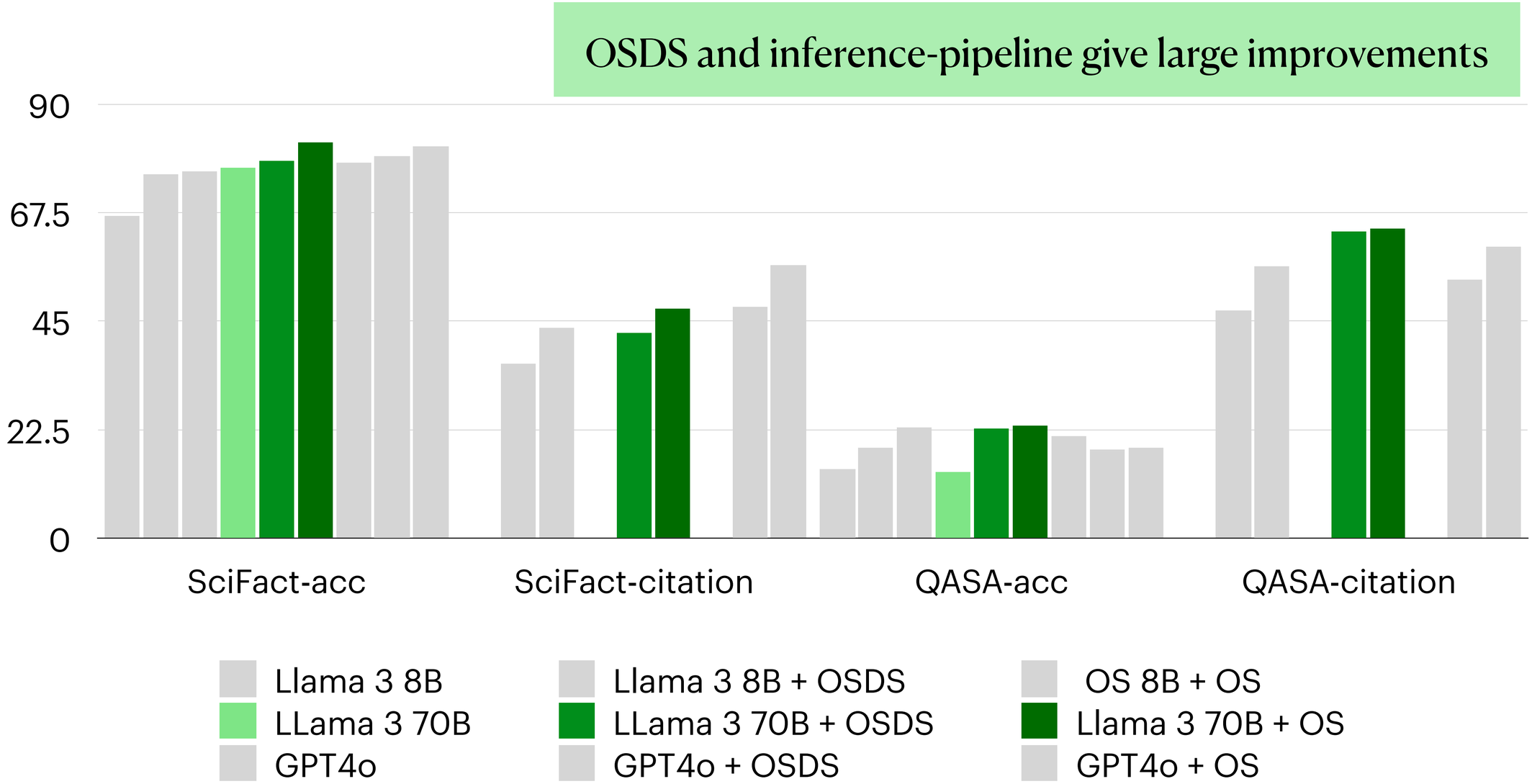
# Results on Single-paper Tasks




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

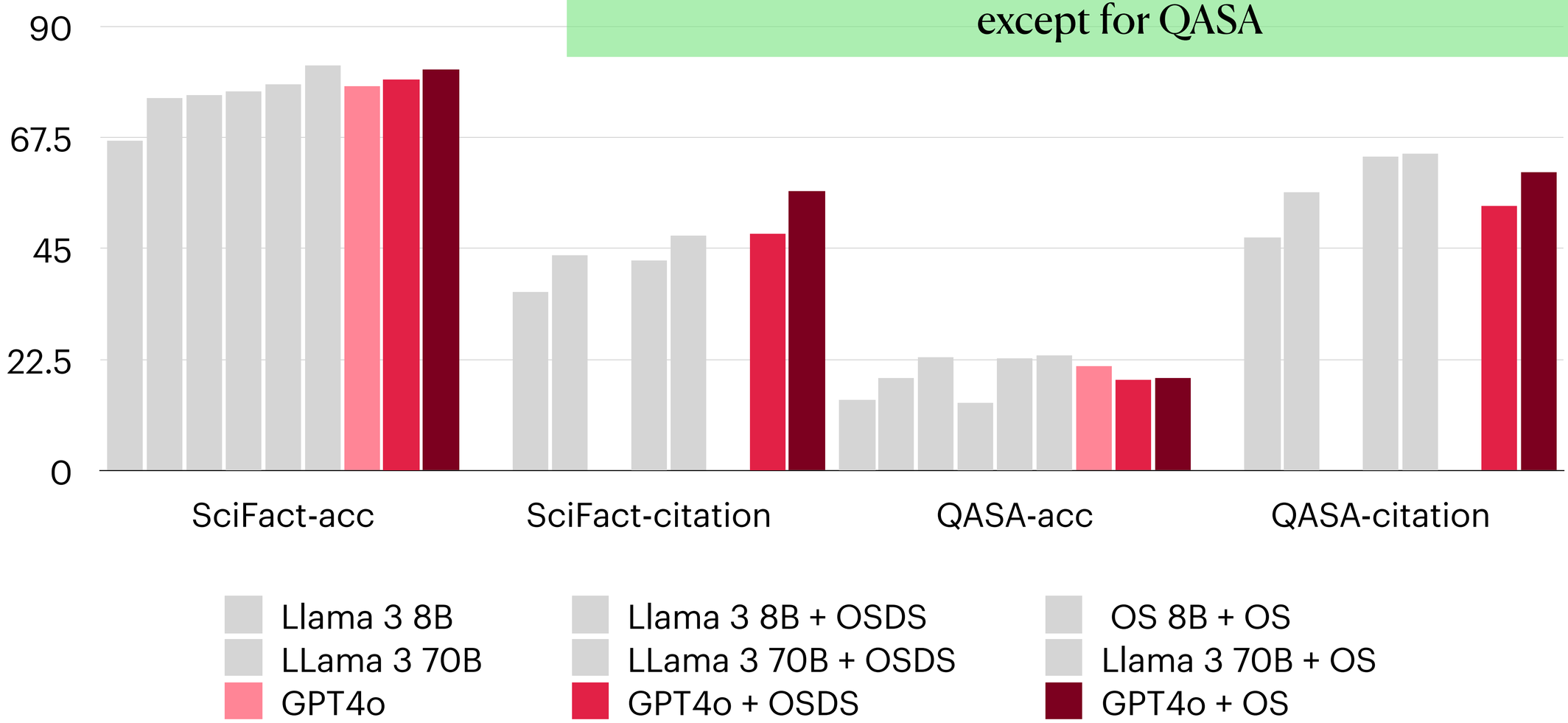
Table 3: **Statistics of hallucinated papers in computer science and biomedicine domains**. Our analysis revealed a significant number of non-existent cited papers in predictions made by LLMs without retrieval, a problem not observed in OPENSCHOLAR.

# Results on Single-paper Tasks



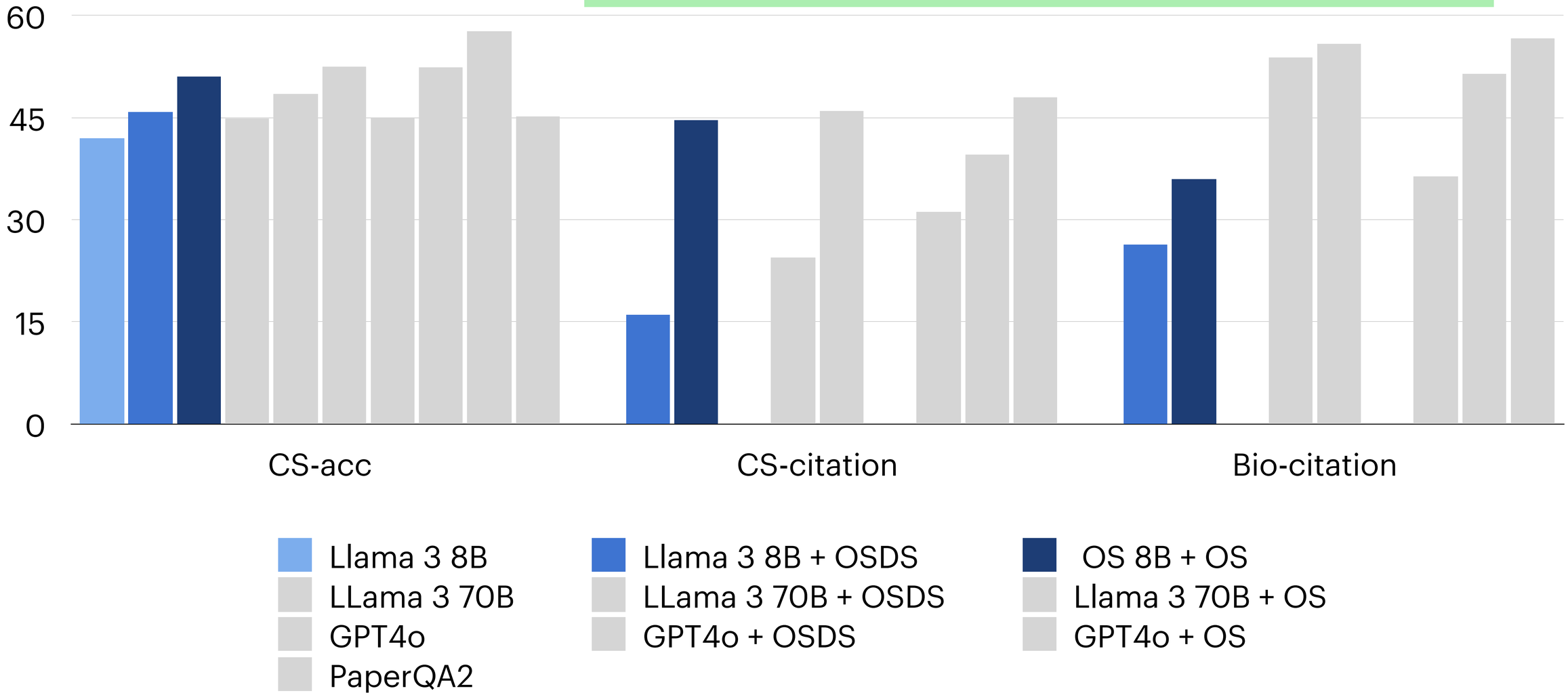
# Results on Single-paper Tasks

Constant improvements for citation accuracy and accuracy except for QASA



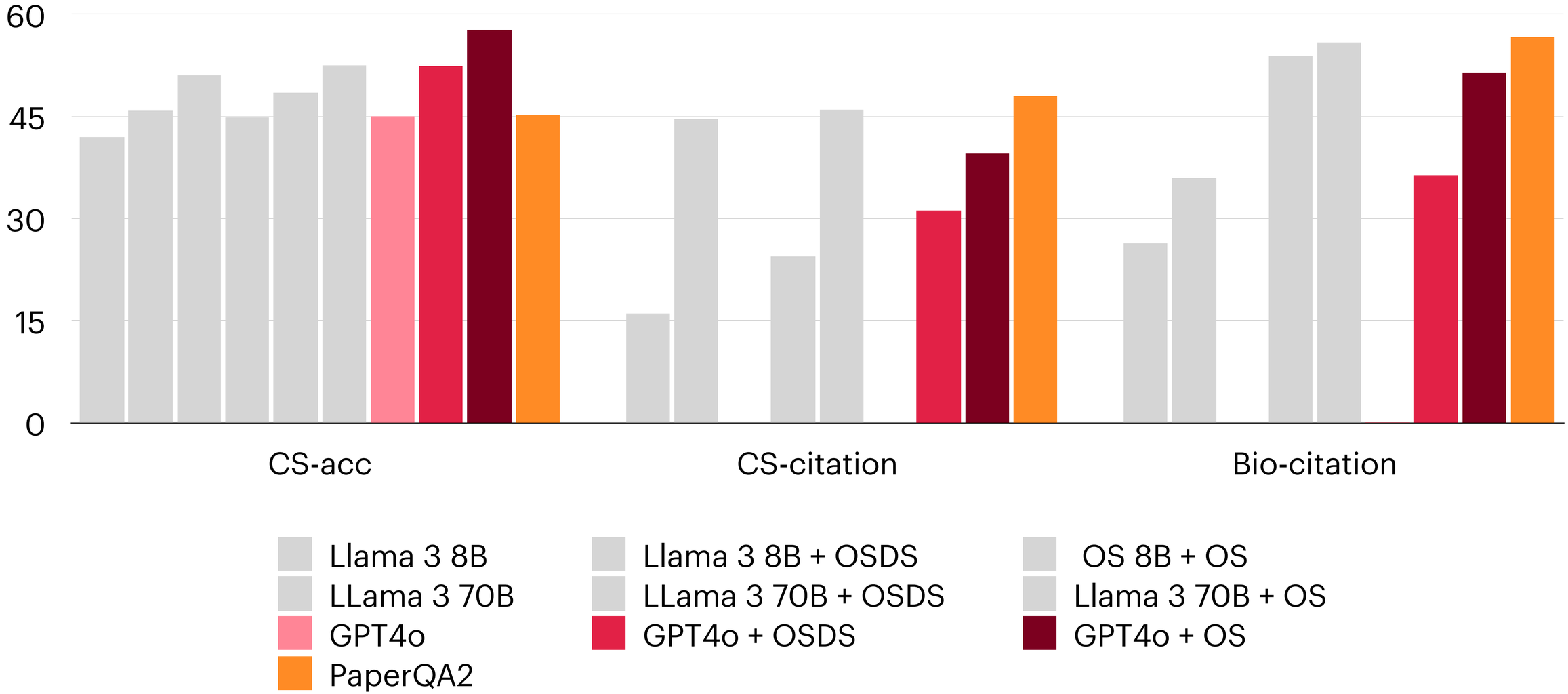
# Results on Multi-paper Tasks

OSDS and inference-pipeline give large improvements

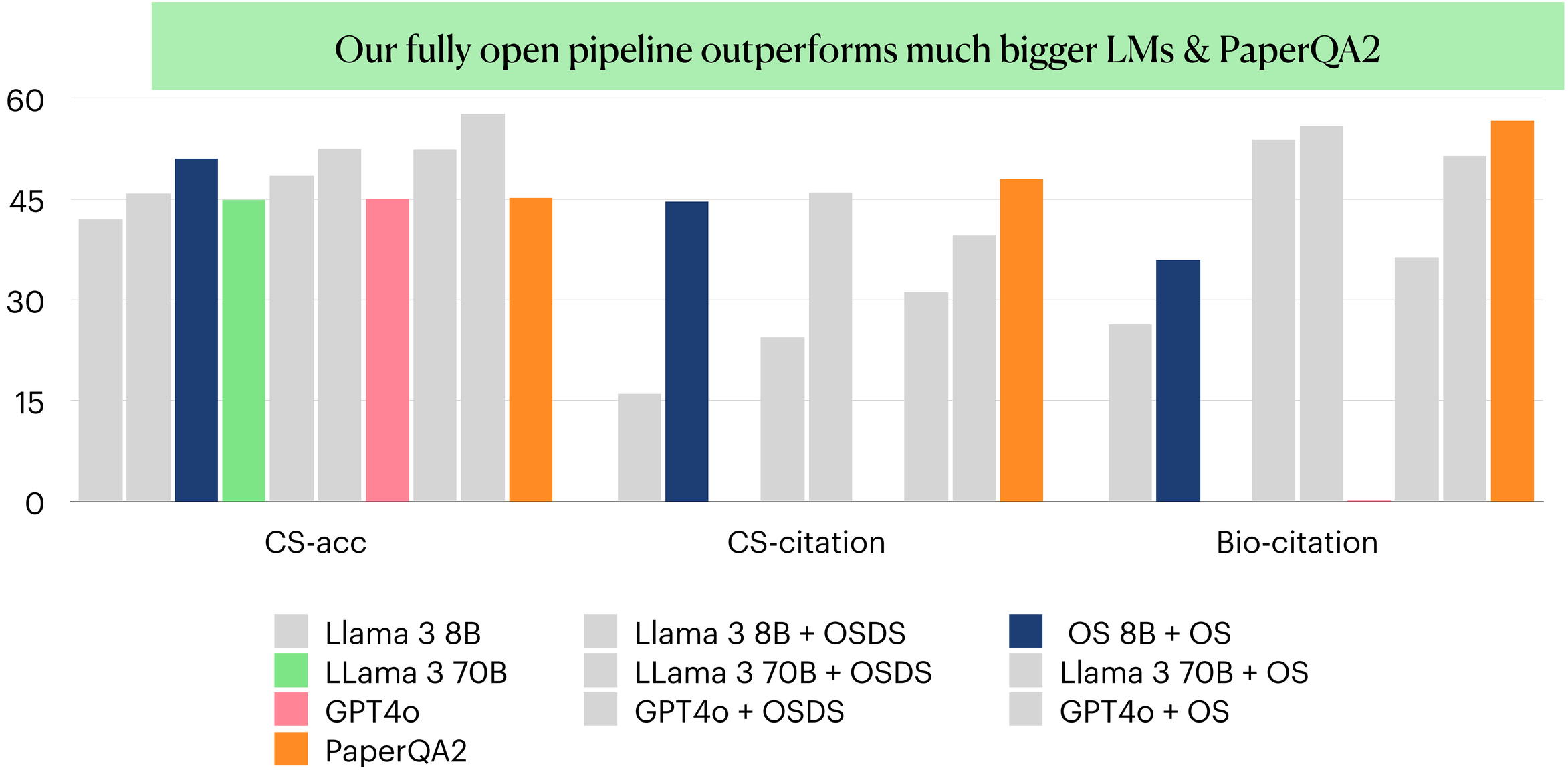


# Results on Multi-paper Tasks

OS-GPT4o achieves SOTA, significantly outperforming PaperQA2



# Results on Multi-paper Tasks



# Qualitative Example of OpenScholar 8B

Has anyone showed the effectiveness of scaling retrieval-augmented LMs' retrieval tokens to trillion-token scale?

But how good are they?

Scaling up the retrieval corpora for Retrieval-Augmented Generation (RAG) to trillion tokens has been explored in several studies [1][2][3][4][5].

Studies have shown that scaling up the retrieval corpora to trillion tokens can bring significant improvements to knowledge-intensive question answering tasks [2][4]. For instance, a Llama-2 7B model that retrieves from fewer than 100B tokens can outperform both its 13B LM-only counterpart and the more capable LM-only Llama-3 8B on tasks such as TQA and NQ [3]. This indicates the effectiveness of storing knowledge in the datastore and retrieving relevant information during inference.

[1] **Improving language models by retrieving from trillions of tokens** In this work, we endeavor to decouple these, by exploring efficient means of augmenting language mod ... (read more details)

[2] **Scaling Retrieval-Based Language Models with a Trillion-Token Datastore** Following prior work (Izacard et al., 2023; Liu et al., 2023; Ram et al., 2023; Shi et al., 2024; Xu ... (read more details)

[3] **Scaling Retrieval-Based Language Models with a Trillion-Token Datastore** Studying datastore scaling requires constructing datastores of varying sizes and varying composition ... (read more details)

[4] **Improving language models by retrieving from trillions of tokens** Fig. 1 (middle) shows how scaling the retrieval database at evaluation improves the language modelli ... (read more details)

[5] **Benchmarking Retrieval-Augmented Generation for Medicine** Figure 3 shows the scaling curves of MedRag on each task in Mirage with different numbers of snippet ... (read more details)

OpenScholar

# Expert Evaluation: *Is OpenScholar as good as expert?*

Step 1:

**Question Writing**

(Group 1) Ph.D. expert

Which motion degrees of freedom of levitated nanospheres have been experimentally studied?

*This data is integrated as **Scholar-Multi**

Step 2:

**Answer Generation**

56.8 min

1 min

Initially, research primarily focused on the translational degrees of freedom of nanoparticles. Through these translational degrees of freedom, researchers

The motion degrees of freedom of levitated nanospheres that have been experimentally studied include translational, rotational, ... Translational Motion:

Step 3:

**Evaluation**

(Group 2) Ph.D. experts

Pair-wise evaluation

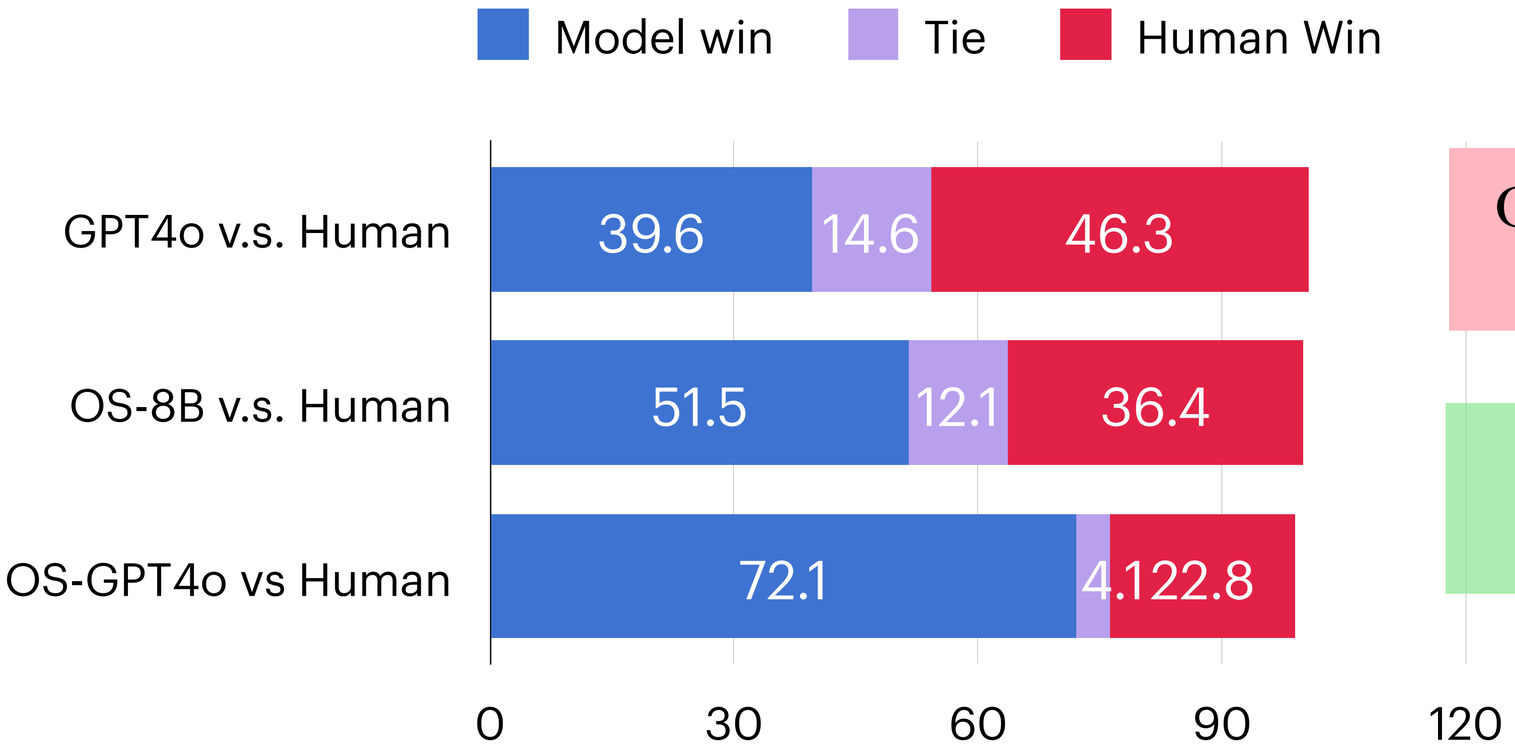
Human < OS

*More organized and providing rich overview with representative set of papers*

Fine-grained evaluation

| | Coverage | Relevance | Organization | Overall usefulness |
|---|---|---|---|---|
| Human | 4 | 4 | 4 | 4 |
| OS | **5** | 4 | **5** | 4 |

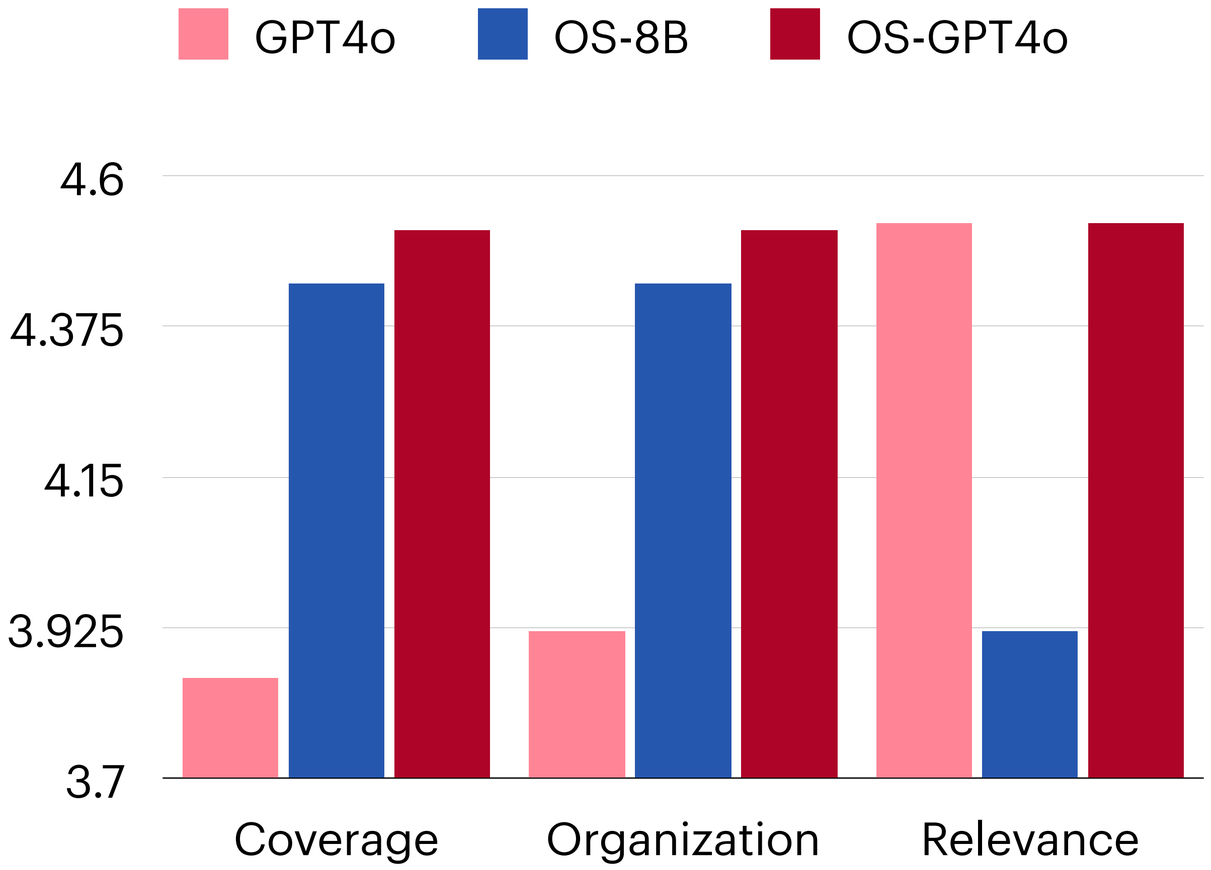
# Expert Evaluation Results - 0.5k Expert Judgements




GPT4o win rate is lower than 50% unlike OS models

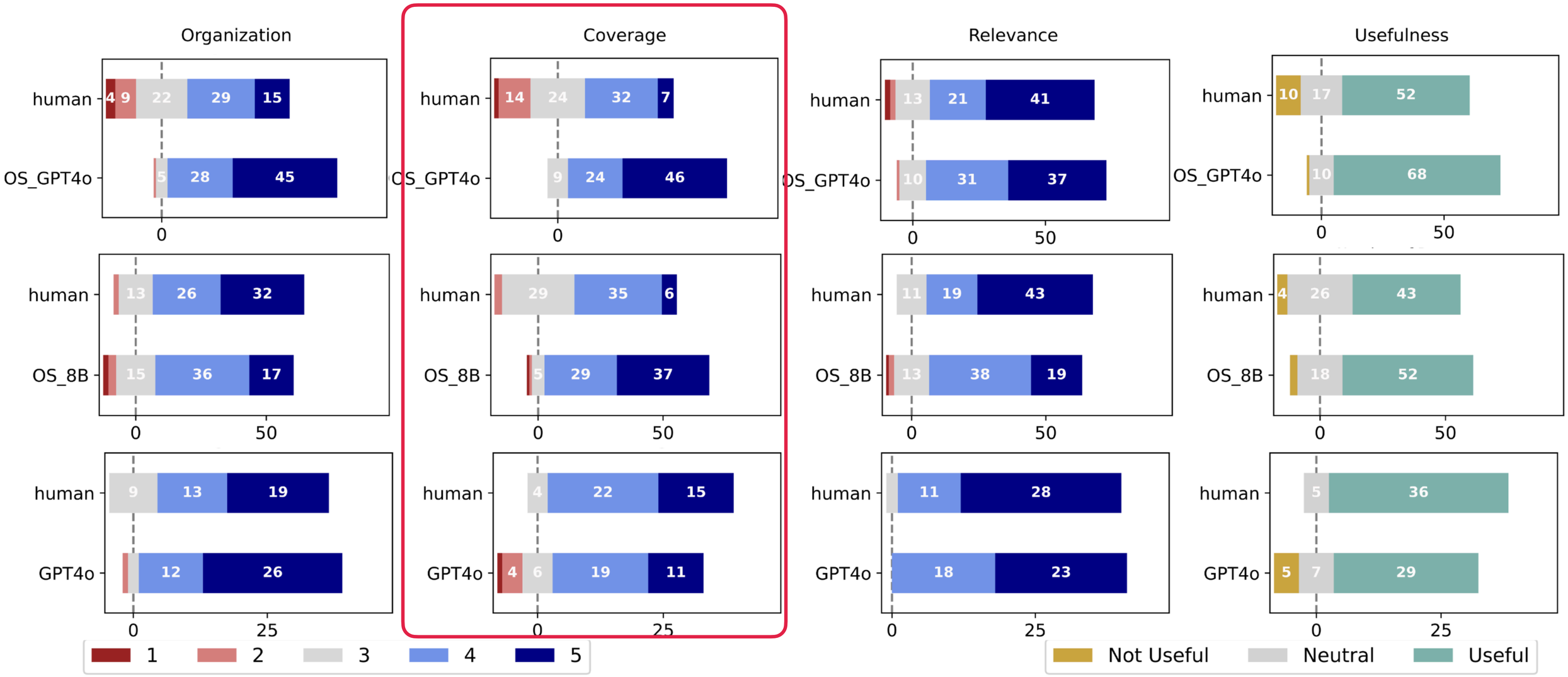
OS-8B and GPT4o answers are preferred over expert answers

# Expert Evaluation Results - 0.5k Expert Judgements

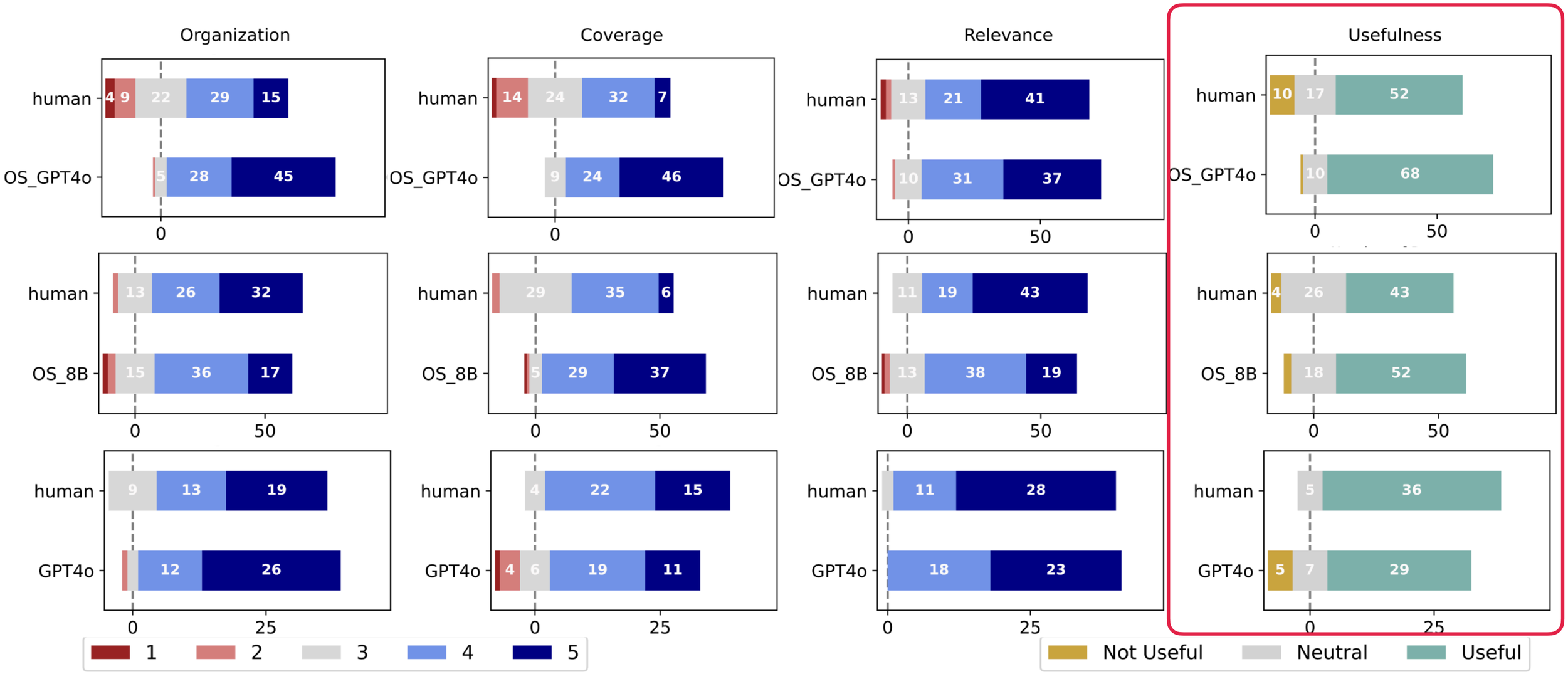



OS models significantly improves **Coverage** and **Organization**

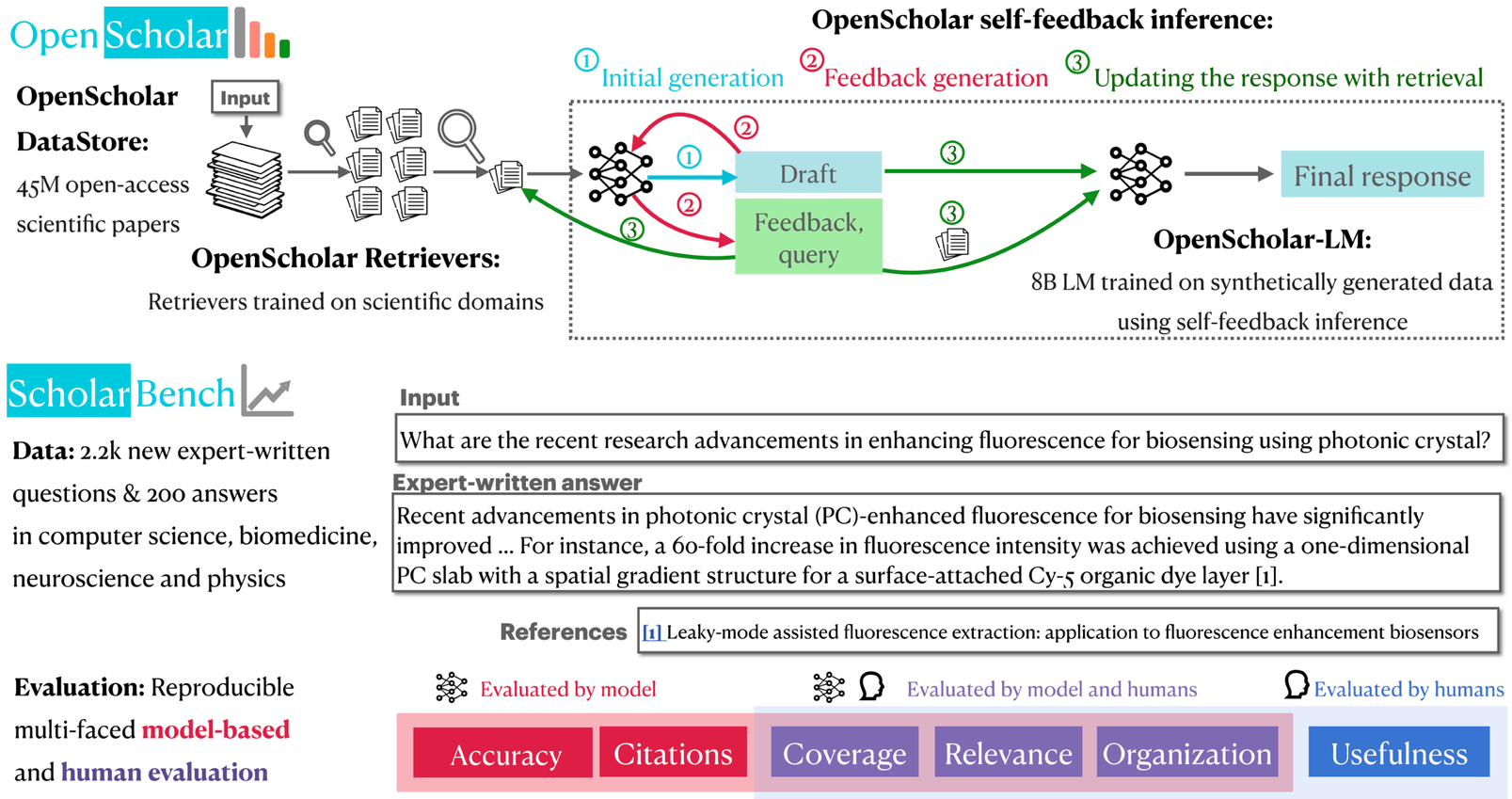
# Expert Evaluation Results - 0.5k Expert Judgements



# Expert Evaluation Results - 0.5k Expert Judgements



# Thanks for listening!




**Contact:** akari@cs.washington.edu
**Website:** https://akariasai.github.io/

**Paper:** https://tinyurl.com/akariopenscholar
**Feedback:** https://tinyurl.com/akarifeedback


54