# RAG in the Wild: On the (In)effectiveness of LLMs with Mixture-of-Knowledge Retrieval Augmentation

**Ran Xu**[†] **Yuchen Zhuang**[‡] **Yue Yu**[‡] **Haoyu Wang**[#] **Wenqi Shi**[◇] **Carl Yang**[†]
[†] Emory University [‡] Georgia Institute of Technology
[#] SUNY Albany [◇] UT Southwestern Medical Center
{ran.xu, j.carlyang}@emory.edu

## Abstract

Retrieval-augmented generation (RAG) enhances large language models (LLMs) by integrating external knowledge retrieved at inference time. While RAG demonstrates strong performance on benchmarks largely derived from general-domain corpora like Wikipedia, its effectiveness under realistic, diverse retrieval scenarios remains underexplored. We evaluated RAG systems using MASSIVEDS, a large-scale datastore with mixture of knowledge, and identified critical limitations: retrieval mainly benefits smaller models, rerankers add minimal value, and no single retrieval source consistently excels. Moreover, current LLMs struggle to route queries across heterogeneous knowledge sources. These findings highlight the need for adaptive retrieval strategies before deploying RAG in real-world settings. Our code and data can be found at https://github.com/ritaranx/RAG_in_the_Wild.

## 1 Introduction

Retrieval-Augmented Generation (RAG) serves as a useful strategy for adapting large language models (LLMs) to tasks that require long-tailed or domain-specific knowledge. By retrieving relevant information from external knowledge sources, RAG provides LLMs with contextual evidence at inference time, enhancing performance on knowledge-intensive NLP tasks including question answering (Lewis et al., 2020; Izacard et al., 2023), fact verification (Lin et al., 2024), and reasoning (Islam et al., 2024; Li et al., 2025).

Despite the strong performance of RAG on a range of benchmarks, these evaluations are predominantly constructed from, or closely aligned with Wikipedia (Kwiatkowski et al., 2019; Joshi et al., 2017; Thorne et al., 2018; Yang et al., 2018; Petroni et al., 2021; Mallen et al., 2023, *inter alia*). Consequently, high accuracy in these settings is somehow not particularly surprising, as many queries are well-covered by the retrieval corpus. In contrast, retrieval in real-world applications is substantially more challenging. The target corpus may be noisy, domain-specific, or misaligned with the query distribution. Although recent efforts have sought to evaluate RAG systems in broader knowledge tasks (Asai et al., 2024; Shi et al., 2024; Huang et al., 2024; Zhang et al., 2024), the underlying corpora in these studies only include Wikipedia or similar sources in the general domain, thus limiting their generalizability. These observations highlight a pressing need to evaluate RAG systems under real-world retrieval conditions.

Motivated by these considerations, our goal is to comprehensively examine how corpus composition and model scale affect the performance of RAG systems under a more realistic setting. To this end, we consider a *mixture-of-knowledge* scenario using MASSIVEDS (Shao et al., 2024) – a large-scale, multi-domain datastore that combines general web sources (e.g., CommonCrawl) with specialized domains (e.g., PubMed). We evaluate tasks spanning both general knowledge and domain-specific QA, where *no prior information* about the underlying knowledge source is available for each question. From the results, we have several findings:

- Under the mixture-of-knowledge setting, the benefits of *retrieval* are largely confined to smaller language models only – once the backbone model becomes sufficiently powerful, these gains diminish significantly, except for factuality-focused QA tasks.
- These diminishing returns aren't solely due to retrieval quality; adding a reranker offers marginal improvements, suggesting deeper integration challenges between retrieval and generation.
- No single source consistently outperforms others, including no-retrieval baselines, emphasizing the need for *adaptive retrieval*. Yet, current LLMs



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 17191–17206

struggle to route queries effectively across heterogeneous sources.

We believe our evaluations provide new insights and motivate future research towards more adaptive and robust retrieval-augmented generation systems under realistic, multi-domain conditions.

## 2 Experiments

### 2.1 Experiment Setup

**Datasets.** To evaluate the accuracy of RAG systems across domains, we identify two key desiderata: (1) *Topic coverage*: the dataset should span a broad range of tasks that necessitate external knowledge for successful resolution; and (2) *Corpus diversity*: questions should not be created solely from a single corpus (e.g., Wikipedia or PubMed).

With this in mind, we select six datasets for our evaluation: (1) **General knowledge-based QA**: we use MMLU (Hendrycks et al., 2021) and MMLU-Pro (Wang et al., 2024b), which contain questions spanning a wide array of subjects such as history, law, and medicine. (2) **Scientific QA**: We evaluate on ARC Challenge (Clark et al., 2018) and SciQ (Welbl et al., 2017) for natural science, and CSBench (Song et al., 2025) for computer science, to test the model's reasoning over domain-specific scientific knowledge. (3) **Factuality**: We adopt SimpleQA (Wei et al., 2024), a benchmark collected from multiple web sources to test factual correctness. Note that we *remove* those questions that were created using Wikipedia pages.

**Backbones.** We consider three families of LLMs: (1) **Llama-series**: Llama-3.2-3B and Llama-3.1-8B (Grattafiori et al., 2024); (2) **Qwen-series**: Qwen3-4B/8B/32B (Yang et al., 2025); (3) **GPT-series**: GPT-4o-mini/-4o (Hurst et al., 2024). Note that we use `instruct` version in our experiments.

**Retrieval Corpora.** We use `MassiveDS`, a massive datastore with 1.4T tokens as the knowledge source, where we consider PubMed, Wikipedia, Pes2o, C4, Github, Math, StackExchange (SE), Book, Arxiv and CommonCrawl (CC) as sources.

**Evaluation Metrics.** Our evaluation covers multiple-choice tasks (MMLU, MMLU-Pro, ARC-C), short-form generation (SimpleQA, SciQ), and mixed formats (CSBench). We report accuracy for multiple-choice, exact match (EM) for generation, and follow Song et al. (2025) for CSBench. Retrieval effectiveness is measured by relative gain: $\Delta(p_s) = \frac{p_s - \rho}{\rho}$, where $p_s$ is the RAG performance using source $s$, and $\rho$ is performance of the LLM baseline without retrieval.

Table 1: Performance Gains of using Retrieval across different Domains in MMLU.

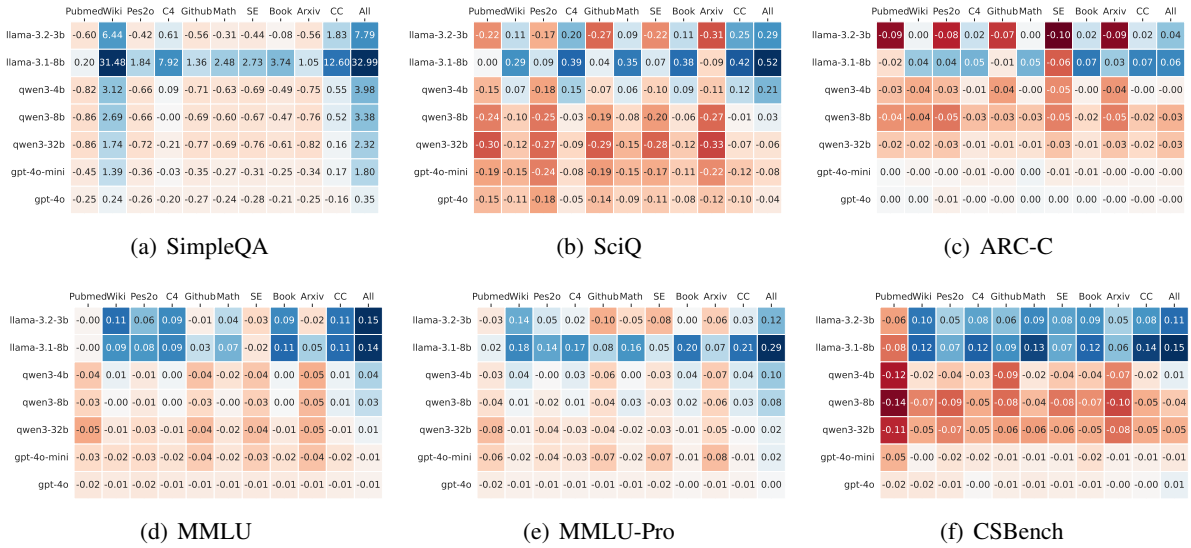
| Model | STEM | Social Sciences | Humanities | Others |
|---|---|---|---|---|
| Llama-3.2-3B | 0.388 | 0.544 | 0.444 | 0.563 |
| w/ retrieval | 0.477 | 0.644 | 0.478 | 0.649 |
| Δ | +22.87% | +18.48% | +7.70% | +15.31% |
| Llama-3.1-8B | 0.472 | 0.598 | 0.492 | 0.578 |
| w/ retrieval | 0.533 | 0.702 | 0.530 | 0.702 |
| Δ | +12.93% | +17.35% | +7.85% | +21.39% |
| Qwen-3-4B | 0.653 | 0.761 | 0.576 | 0.707 |
| w/ retrieval | 0.670 | 0.785 | 0.597 | 0.762 |
| Δ | +2.57% | +3.16% | +3.57% | +7.91% |
| Qwen-3-8B | 0.677 | 0.790 | 0.598 | 0.750 |
| w/ retrieval | 0.699 | 0.807 | 0.618 | 0.790 |
| Δ | +3.19% | +2.13% | +3.24% | +5.29% |
| Qwen3-32B | 0.785 | 0.856 | 0.701 | 0.824 |
| w/ retrieval | 0.803 | 0.851 | 0.705 | 0.827 |
| Δ | +2.28% | -0.50% | +0.61% | +0.41% |
| GPT-4o-mini | 0.686 | 0.853 | 0.671 | 0.816 |
| w/ retrieval | 0.697 | 0.844 | 0.657 | 0.815 |
| Δ | +1.53% | -1.04% | -2.12% | -0.18% |
| GPT-4o | 0.773 | 0.900 | 0.806 | 0.867 |
| w/ retrieval | 0.775 | 0.891 | 0.796 | 0.867 |
| Δ | +0.34% | -1.08% | -1.28% | -0.02% |

**Implementation Details.** For Qwen-3, we use the non-reasoning model in our evaluation. For retrieval, we use *bge-base-en-v1.5* and *bge-reranker-v2-m3* (Chen et al., 2024) as the default retriever and reranker. We evaluate *zero-shot* performance with $k = 5$ top passage. For the reranking setting, we first retrieve $k' = 30$ passages, then use the reranker to obtain $k = 5$ passages with the highest relevance scores. Following Shao et al. (2024), we apply document filtering and deduplication.

### 2.2 RQ1: Effectiveness of RAG under Mixture-of-Knowledge Scenarios.

Figure 1 illustrates consistent benefits of RAG across a variety of real-world, mixture-of-knowledge scenarios. Notably, *smaller models achieve substantial performance gains*, reflecting their limited capacity to store knowledge internally. In contrast, larger models show diminishing returns from retrieval, with improvements mostly limited to factual knowledge tasks such as SimpleQA. For general and scientific knowledge tasks, which are more effectively captured within the models' parametric knowledge through pretraining and finetuning, external retrieval brings less benefit.

As some datasets, such as MMLU, cover a wide range of domains, we further analyze the effect of retrieval augmentation across STEM, Social Sciences, Humanities, and Others, as shown in Table 2. We observe consistent trends across all domains: retrieval brings substantial improvements for smaller




(a) SimpleQA (b) SciQ (c) ARC-C

(d) MMLU (e) MMLU-Pro (f) CSBench

Figure 1: The relevance performance of different LLMs compared to non-retrieval baselines on six datasets. 'All' means the corpus from all domains is considered for retrieval.

and mid-sized models, while the gains diminish as model size increases. For the strongest models, retrieval offers marginal or even slightly negative gains, suggesting that larger models are increasingly able to internalize domain knowledge without the need for external retrieval. These results highlight the generalizability of our findings and demonstrate that the impact of retrieval augmentation is robust across different knowledge domains.

### 2.3 RQ 2: Instance-level Analysis of Multiple Retrieval Sources

To investigate whether different retrieval sources provide unique advantages, we conduct an instance-level study (Figure 2) measuring the proportion of queries that can be solved only by retrieving from a specific corpus—compared to using all sources or no retrieval at all. Our results show that a significant fraction of cases (e.g., 8%–39% for Llama-3.1-8B) depend exclusively on retrieval from particular corpora. Crucially, *no single retrieval source consistently outperforms others across all query types*, highlighting the need for dynamic, query-specific routing to the most relevant corpus to maximize RAG performance in heterogeneous, real-world knowledge environments.

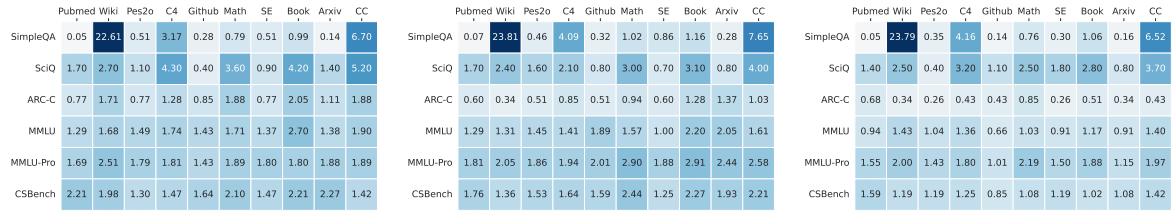
### 2.4 RQ 3: Effectiveness of Reranking

One possible reason for the limited improvement in RAG performance is the inherent limitations of retrievers. To investigate this, inspired by prior work demonstrating benefits from adding reranking to the RAG pipeline (Shao et al., 2024; Yu et al., 2024), we apply reranking to the top-30 retrieved results. As shown in Figure 3, reranking yields only marginal gains across datasets. This suggests that *improving retrieval quality through reranking is insufficient* in mixture-of-knowledge scenarios. The retriever's limited capacity and restricted access to relevant knowledge highlight the need for deeper integration between knowledge sources, retrieval mechanisms, and generative models.

### 2.5 RQ 4: Evaluating LLMs as Routers for Mixture-of-Knowledge Retrieval

Previous analyses highlight the need for *adaptive retrieval* mechanisms that dynamically route queries to the most relevant corpus based on topic context. Here, we investigate whether current LLMs can effectively perform query routing. Figure 4 evaluates LLMs from the `Qwen-3` series (4B, 8B, and 32B parameters) as "routers" that select among heterogeneous knowledge sources at inference time. We compare plain prompting versus chain-of-thought prompting on the MMLU and MMLU-Pro datasets, which cover diverse general and professional domains.

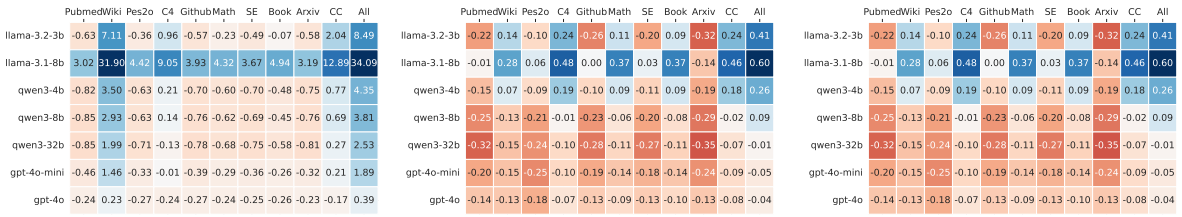
Specifically, we benchmark: (1) no-retrieval baselines; (2) static retrieval from all corpora ("all sources"); (3) LLM-prompted routing variants (plain and chain-of-thought); and (4) an oracle router upper bound. Surprisingly, neither prompting strategy consistently outperforms static retrieval. In fact, both routing approaches often




(a) LLaMA-3.2-3B (b) LLaMA-3.1-8B (c) Qwen-3-4B

(d) Qwen-3-8B (e) Qwen-3-32B (f) GPT-4o

Figure 2: Number of cases (in %) specifically resolved by retrieving from an individual corpus.




(a) SimpleQA (b) SciQ (c) ARC-C

(d) MMLU (e) MMLU-Pro (f) CSBench

Figure 3: Performance with rerank on retrieval effectiveness across different datasets and models.

underperform compared to simply retrieving from all sources, and occasionally fall below no-retrieval baselines. Chain-of-thought prompting provides only marginal improvements, while scaling model size yields negligible or even negative returns.

We attribute this failure to two main factors: (1) **Inaccurate relevance estimation:** Without dedicated training, LLMs struggle to reliably identify which corpus holds the needed information, especially amid overlapping or stylistically diverse corpora. Minor routing errors propagate to poor retrieval quality. (2) **Training-inference mismatch:** LLM training typically lacks explicit multi-source comparison tasks, limiting the effectiveness of prompt-based routing as a meta-reasoning problem. Future directions should focus on learned routing modules trained with supervision or reinforcement learning, or on tightly integrated RAG systems that jointly optimize source selection and generation.

## 3 Related Work

Many of prior works on RAG focus on improving either retrievers (Shi et al., 2024; Shao et al., 2023; Xu et al., 2024b) or language models (Asai et al., 2024; Xu et al., 2024c; Huang et al., 2024), as well as optimizing the whole RAG pipeline (Trivedi et al., 2023; Wang et al., 2024a). To better understand the impact of retrieval corpora, Shao et al. (2024) explores how corpus scale influences RAG performance, while Cuconasu et al. (2024); Niu

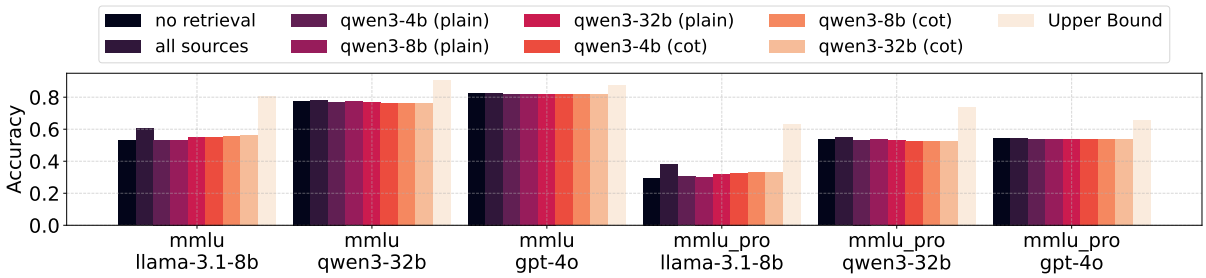



Figure 4: Performance comparison of routing strategies across MMLU and MMLU-Pro datasets.

et al. (2024) examine the effects of noisy or imperfect corpora. Chen et al. (2025); Xu et al. (2024a); Xiong et al. (2024) assess RAG systems across different knowledge sources, though they primarily focus on biomedical tasks. In contrast, our work provides a broader investigation, spanning diverse domains, retrieval corpora, and LLM backbones.

In parallel, data routing strategies have been explored for LLMs. Ong et al. (2025); Frick et al. (2025) routes each prompt to a specialized LLM. Within the RAG setting, Mu et al. (2024, 2025) investigate routing prompts to different search engines, Zhang et al. (2025) routes queries to different LLMs, and Wu et al. (2025); Asai et al. (2024); Yao et al. (2024) studied *adaptive retrieval* to dynamically determine retrieval during inference. Distinct from these efforts, we also focuses on routing but operates at the *corpus level*, studying adaptive selection among heterogeneous knowledge sources.

## 4 Conclusion

Our comprehensive empirical analysis highlights key challenges in deploying RAG systems within realistic, heterogeneous knowledge environments. While retrieval offers benefits for smaller LLMs, larger models exhibit only marginal improvements, except for factual tasks. Additionally, reranking techniques and prompt-based knowledge routing provide limited gains, justifying the need for future research focused on adaptive routing and tighter integration between retrieval and generation.

We identify several promising directions to address these challenges: (1) *Reasoning-enhanced search*: Shao et al. (2025); Zhuang et al. (2025) incorporate reasoning directly into retrieval and ranking stages, improving relevance at the expense of increased inference time. (2) *Agentic RAG systems*: Query rewriting (Ma et al., 2023) and decomposition (Li et al., 2025; Jin et al., 2025) enable multi-turn retrieval and generation workflows to better handle complex queries.

## Limitations

Our study primarily targets question answering tasks with short-form answers, which may limit the applicability of our results to other settings such as open-ended generation or long-form reasoning. While we evaluate several widely used models, our experiments do not comprehensively cover larger open-source models (e.g., DeepSeek-V3 (Liu et al., 2024) or Llama-4 (Meta, 2025)) or alternative retrieval paradigms, mainly due to computational constraints. Consequently, there remains ample opportunity for future work to explore these directions. Lastly, we do not assess computational efficiency or latency trade-offs, which are critical considerations for real-world deployment of RAG systems but are beyond the scope of this study.

## References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Zhe Chen, Yusheng Liao, Shuyang Jiang, Pingjie Wang, Yiqiu Guo, Yanfeng Wang, and Yu Wang. 2025. Towards omni-rag: Comprehensive retrieval-augmented generation for large language models in medical applications. *arXiv preprint arXiv:2501.02460*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.

Evan Frick, Connor Chen, Joseph Tennyson, Tianle Li, Wei-Lin Chiang, Anastasios N Angelopoulos, and Ion Stoica. 2025. Prompt-to-leaderboard. *arXiv preprint arXiv:2502.14855*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jie Huang, Wei Ping, Peng Xu, Mohammad Shoeybi, Kevin Chang, and Bryan Catanzaro. 2024. RAVEN: In-context learning with retrieval-augmented encoder-decoder language models. In *First Conference on Language Modeling*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shayekh Bin Islam, Md Asib Rahman, K S M Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. 2024. Open-RAG: Enhanced retrieval augmented reasoning with open-source large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14231–14244, Miami, Florida, USA. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2024. RA-DIT: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Xinbei Ma, Yeyun Gong, Pengcheng He, hai zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

AI Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation.

Feiteng Mu, Yong Jiang, Liwen Zhang, Liuchu Liuchu, Wenjie Li, Pengjun Xie, and Fei Huang. 2024. Query routing for homogeneous tools: An instantiation in the RAG scenario. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10225–10230, Miami, Florida, USA. Association for Computational Linguistics.

Feiteng Mu, Liwen Zhang, Yong Jiang, Wenjie Li, Zhen Zhang, Pengjun Xie, and Fei Huang. 2025. Unsupervised query routing for retrieval augmented generation. *arXiv preprint arXiv:2501.07793*.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. 2025. RouteLLM: Learning to route LLMs from preference data. In *The Thirteenth International Conference on Learning Representations*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei W Koh. 2024. Scaling retrieval-based language models with a trillion-token datastore. *Advances in Neural Information Processing Systems*, 37:91260–91299.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, et al. 2025. Reasonir: Training retrievers for reasoning tasks. *arXiv preprint arXiv:2504.20595*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Xiaoshuai Song, Muxi Diao, Guanting Dong, Zhengyang Wang, Yujia Fu, Runqi Qiao, Zhexu Wang, Dayuan Fu, Huangxuan Wu, Bin Liang, Weihao Zeng, Yejie Wang, Zhuoma GongQue, Jianing Yu, Qiuna Tan, and Weiran Xu. 2025. CS-bench: A comprehensive benchmark for large language models towards computer science mastery. In *The Thirteenth International Conference on Learning Representations*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Haoyu Wang, Ruirui Li, Haoming Jiang, Jinjin Tian, Zhengyang Wang, Chen Luo, Xianfeng Tang, Monica Xiao Cheng, Tuo Zhao, and Jing Gao. 2024a. BlendFilter: Advancing retrieval-augmented large language models via query generation blending and knowledge filtering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1009–1025, Miami, Florida, USA. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

Di Wu, Jia-Chen Gu, Kai-Wei Chang, and Nanyun Peng. 2025. Self-routing rag: Binding selective retrieval with knowledge verbalization. *arXiv preprint arXiv:2504.01018*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May Dongmei Wang, Joyce Ho, and Carl Yang. 2024a. RAM-EHR: Retrieval augmentation meets clinical predictions on electronic health records. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 754–765, Bangkok, Thailand. Association for Computational Linguistics.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May Dongmei Wang, Joyce C. Ho, Chao Zhang, and Carl Yang. 2024b. Bmretriever: Tuning large language models as better biomedical text retrievers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22234–22254, Miami, Florida, USA. Association for Computational Linguistics.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024c. Unsupervised information refinement training of large language models for retrieval-augmented generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 133–145, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, and Juanzi Li. 2024. Seakr: Self-aware knowledge retrieval for adaptive retrieval augmented generation. *arXiv preprint arXiv:2406.19215*.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jiarui Zhang, Xiangyu Liu, Yong Hu, Chaoyue Niu, Fan Wu, and Guihai Chen. 2025. Query routing for retrieval-augmented language models. *arXiv preprint arXiv:2505.23052*.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. RAFT: Adapting language model to domain specific RAG. In *First Conference on Language Modeling*.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning. *arXiv preprint arXiv:2503.06034*.

Table 2: Performance Gains of using Retrieval across different Domains in MMLU.

| Model | STEM | Social Sciences | Humanities | Others |
|---|---|---|---|---|
| Llama-3.2-3B | 0.388 | 0.544 | 0.444 | 0.563 |
| w/ retrieval | 0.477 | 0.644 | 0.478 | 0.649 |
| Δ | +22.87% | +18.48% | +7.70% | +15.31% |
| Llama-3.1-8B | 0.472 | 0.598 | 0.492 | 0.578 |
| w/ retrieval | 0.533 | 0.702 | 0.530 | 0.702 |
| Δ | +12.93% | +17.35% | +7.85% | +21.39% |
| Qwen-3-4B | 0.653 | 0.761 | 0.576 | 0.707 |
| w/ retrieval | 0.670 | 0.785 | 0.597 | 0.762 |
| Δ | +2.57% | +3.16% | +3.57% | +7.91% |
| Qwen-3-8B | 0.677 | 0.790 | 0.598 | 0.750 |
| w/ retrieval | 0.699 | 0.807 | 0.618 | 0.790 |
| Δ | +3.19% | +2.13% | +3.24% | +5.29% |
| Qwen3-32B | 0.785 | 0.856 | 0.701 | 0.824 |
| w/ retrieval | 0.803 | 0.851 | 0.705 | 0.827 |
| Δ | +2.28% | -0.50% | +0.61% | +0.41% |
| GPT-4o-mini | 0.686 | 0.853 | 0.671 | 0.816 |
| w/ retrieval | 0.697 | 0.844 | 0.657 | 0.815 |
| Δ | +1.53% | -1.04% | -2.12% | -0.18% |
| GPT-4o | 0.773 | 0.900 | 0.806 | 0.867 |
| w/ retrieval | 0.775 | 0.891 | 0.796 | 0.867 |
| Δ | +0.34% | -1.08% | -1.28% | -0.02% |

## A Performance Across Different Domains on MMLU

As some datasets, such as MMLU, cover a wide range of domains, we further analyze the effect of retrieval augmentation across STEM, Social Sciences, Humanities, and Others, as shown in Table 2. We observe consistent trends across all domains: retrieval brings substantial improvements for smaller and mid-sized models, while the gains diminish as model size increases. For the strongest models, retrieval offers marginal or even slightly negative gains, suggesting that larger models are increasingly able to internalize domain knowledge without the need for external retrieval. These results highlight the generalizability of our findings and demonstrate that the impact of retrieval augmentation is robust across different knowledge domains.

## B Performance with Stronger LLM Reranker

While using stronger LLM-based rerankers can often lead to additional gains, the limitations mainly lies in latency. For example, for a dataset with $m$ training examples, reranking top-$k$ documents lead to $m * k$ forward runs ($k = 30$ in our study), which can be very large with large-scale datasets. For example, using RankLlama (castorini/rankllama-v1-7b-lora-passage) as the reranker took more than 5 seconds for each query, which leads to 2x inference overhead. Despite this, we run the experiments using BGE-v2-gemma on MMLU datasets using Qwen3-4b and Llama-3.1-8b as the backbone. The results in Table 3 suggest that, in our setting, even with stronger LLM-based rerankers, the improvements are marginal and do not affect our overall findings. We hypothesize that it is because in many cases, all top-$k$ passages do not contain information that directly leads to the answer, then incorporating a more advanced reranking model does not help much.

Table 3: Performance with Stronger LLM Reranker on the MMLU dataset.

| Method | w/ Qwen3-4B | w/ Llama3.1-8B |
|---|---|---|
| No Retrieval | 0.6595 | 0.5296 |
| All Sources w/ Retrieval | 0.6885 | 0.6055 |
| All Sources w/ Rerank (BGE-Reranker) | 0.6854 | 0.6117 |
| All Sources w/ Rerank (BGE-v2-gemma) | 0.6870 | 0.6109 |

## C Prompt Template

The prompt template for different tasks are listed as follows.

**System Prompt:** You are a useful assistant. I will provide one question, several pieces of passages (which may be related or unrelated to the question). Please answer the question by selecting from one of the choice listed below. Please answer with the capitalized alphabet only, without adding any extra phrase or period.

**User Prompt:** Passages: $\mathcal{P}_q$
Question: $q$

**Assistant Prompt:** {answer $a$}

Figure 5: Prompt for answer generation on multi-choice questions (e.g. MMLU, MMLU-Pro).

**System Prompt:** You are a useful assistant. I will provide one question, several pieces of passages (which may be related or unrelated to the question). Please answer the question with a short span containing one or few keywords.

**User Prompt:** Passages: $\mathcal{P}_q$
Question: $q$

**Assistant Prompt:** {answer $a$}

Figure 6: Prompt for answer generation on span-based questions (e.g. SciQ, SimpleQA).

**System Prompt:** You are a useful assistant.

**User Prompt:** Here are a list of available external corpora:
pubmed: A high-quality biomedical corpus combining PubMed abstracts, focused on formal clinical language.
wikipedia: A cleaned and curated version of English Wikipedia containing factual and encyclopedic content across diverse topics.
c4: A filtered subset of the Colossal Clean Crawled Corpus, representing diverse high-quality web documents in English with broad topical coverage.
pes2o: The peS2o dataset is a collection of approximately 40 million open-access academic papers.
github: A curated collection of open-source GitHub repositories, featuring code, documentation, and discussions centered on software engineering.
math: A dataset focused on mathematical reasoning and problem solving, containing questions, proofs, and solutions.
stackexchange: A cleaned snapshot of Stack Exchange QA threads, spanning technical, academic, and lifestyle topics.
book: A collection of long-form literary and non-fiction texts from public domain and licensed sources (e.g., PG19).
arxiv: A filtered set of LaTeX-based academic papers from arXiv, covering STEM domains with a focus on technical writing.
commoncrawl: A nonprofit organization that provides a free, open repository of web crawl data to support research, analysis.
no: No retrieval. The model have enough knowledge to answer the question.
Given the question: {question}, identify the most appropriate external corpus to retrieve relevant information for answering it. Your response must be one of the source names listed above. If no external source is necessary, respond with no. Please directly output your predicted source with <source> and </source> tags.

Figure 7: Prompt for question routing without chain-of-thought prompting.

**System Prompt:** You are a useful assistant.

**User Prompt:** Here are a list of available external corpora:
pubmed: A high-quality biomedical corpus combining PubMed abstracts, focused on formal clinical language.
wikipedia: A cleaned and curated version of English Wikipedia containing factual and encyclopedic content across diverse topics.
c4: A filtered subset of the Colossal Clean Crawled Corpus, representing diverse high-quality web documents in English with broad topical coverage.
pes2o: The peS2o dataset is a collection of approximately 40 million open-access academic papers.
github: A curated collection of open-source GitHub repositories, featuring code, documentation, and discussions centered on software engineering.
math: A dataset focused on mathematical reasoning and problem solving, containing questions, proofs, and solutions.
stackexchange: A cleaned snapshot of Stack Exchange QA threads, spanning technical, academic, and lifestyle topics.
book: A collection of long-form literary and non-fiction texts from public domain and licensed sources (e.g., PG19).
arxiv: A filtered set of LaTeX-based academic papers from arXiv, covering STEM domains with a focus on technical writing.
commoncrawl: A nonprofit organization that provides a free, open repository of web crawl data to support research, analysis.
no: No retrieval. The model have enough knowledge to answer the question.
Given the question: {question}, identify the most appropriate external corpus to retrieve relevant information for answering it. Your response must be one of the source names listed above. If no external source is necessary, respond with no. Please concise reasoning before output the final source. Please wrap your predicted source with <source> and </source> tags.

Figure 8: Prompt for question routing with chain-of-thought prompting (Wei et al., 2022).

## D Detailed Per-task Performance

The performance of different backbones on these six benchmarks with different knowledge is shown in the Table 4 - 9. The results after adding the reranking stage is shown in Table 10 - 15.

Table 4: Zero-shot Accuracy across different retrieval sources on the SimpleQA task.

| Dataset | Model | plain | Pubmed | Wiki | Pes2o | C4 | Github | Math | SE | Book | Arxiv | CC | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Simpleqa | Llama-3.2-3B | 0.034 | 0.014 | 0.251 | 0.020 | 0.054 | 0.015 | 0.023 | 0.019 | 0.031 | 0.015 | 0.095 | 0.296 |
| Simpleqa | Llama-3.1-8B | 0.009 | 0.011 | 0.286 | 0.025 | 0.079 | 0.021 | 0.031 | 0.033 | 0.042 | 0.018 | 0.120 | 0.299 |
| Simpleqa | Qwen3-4B | 0.068 | 0.013 | 0.281 | 0.023 | 0.075 | 0.020 | 0.026 | 0.021 | 0.035 | 0.017 | 0.106 | 0.339 |
| Simpleqa | Qwen3-8B | 0.079 | 0.011 | 0.293 | 0.027 | 0.079 | 0.025 | 0.032 | 0.026 | 0.042 | 0.019 | 0.121 | 0.345 |
| Simpleqa | Qwen3-32B | 0.105 | 0.015 | 0.296 | 0.030 | 0.082 | 0.024 | 0.032 | 0.025 | 0.041 | 0.019 | 0.121 | 0.347 |
| Simpleqa | GPT-4o-mini | 0.144 | 0.079 | 0.344 | 0.093 | 0.140 | 0.093 | 0.105 | 0.100 | 0.108 | 0.096 | 0.169 | 0.404 |
| Simpleqa | GPT-4o | 0.343 | 0.258 | 0.425 | 0.253 | 0.275 | 0.249 | 0.262 | 0.248 | 0.273 | 0.258 | 0.288 | 0.463 |

Table 5: Zero-shot accuracy of ARC-Challenge for different models and retrieval sources.

| Dataset | Model | plain | Pubmed | Wiki | Pes2o | C4 | Github | Math | SE | Book | Arxiv | CC | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| arc_c | Llama-3.2-3B | 0.626 | 0.572 | 0.629 | 0.579 | 0.639 | 0.583 | 0.627 | 0.564 | 0.639 | 0.570 | 0.637 | 0.650 |
| arc_c | Llama-3.1-8B | 0.693 | 0.682 | 0.718 | 0.720 | 0.724 | 0.687 | 0.724 | 0.652 | 0.742 | 0.717 | 0.738 | 0.736 |
| arc_c | Qwen3-4B | 0.853 | 0.829 | 0.823 | 0.827 | 0.846 | 0.819 | 0.849 | 0.814 | 0.849 | 0.817 | 0.852 | 0.850 |
| arc_c | Qwen3-8B | 0.890 | 0.851 | 0.853 | 0.844 | 0.864 | 0.863 | 0.866 | 0.846 | 0.874 | 0.846 | 0.868 | 0.860 |
| arc_c | Qwen3-32B | 0.928 | 0.905 | 0.906 | 0.904 | 0.919 | 0.915 | 0.916 | 0.903 | 0.917 | 0.904 | 0.911 | 0.903 |
| arc_c | GPT-4o-mini | 0.904 | 0.904 | 0.901 | 0.902 | 0.899 | 0.899 | 0.907 | 0.898 | 0.899 | 0.899 | 0.901 | 0.899 |
| arc_c | GPT-4o | 0.942 | 0.943 | 0.944 | 0.937 | 0.941 | 0.940 | 0.945 | 0.943 | 0.944 | 0.942 | 0.939 | 0.940 |

Table 6: Zero-shot accuracy on SciQ for different models and retrieval sources.

| Dataset | Model | plain | Pubmed | Wiki | Pes2o | C4 | Github | Math | SE | Book | Arxiv | CC | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| sciq_test | Llama-3.2-3B | 0.465 | 0.364 | 0.515 | 0.388 | 0.560 | 0.338 | 0.507 | 0.362 | 0.517 | 0.321 | 0.580 | 0.598 |
| sciq_test | Llama-3.1-8B | 0.423 | 0.425 | 0.545 | 0.461 | 0.589 | 0.440 | 0.571 | 0.451 | 0.583 | 0.386 | 0.602 | 0.645 |
| sciq_test | Qwen3-4B | 0.530 | 0.451 | 0.565 | 0.432 | 0.610 | 0.491 | 0.561 | 0.476 | 0.580 | 0.470 | 0.591 | 0.641 |
| sciq_test | Qwen3-8B | 0.631 | 0.480 | 0.565 | 0.472 | 0.612 | 0.514 | 0.579 | 0.507 | 0.594 | 0.458 | 0.624 | 0.652 |
| sciq_test | Qwen3-32B | 0.705 | 0.494 | 0.621 | 0.512 | 0.641 | 0.500 | 0.599 | 0.507 | 0.622 | 0.471 | 0.658 | 0.660 |
| sciq_test | GPT-4o-mini | 0.690 | 0.562 | 0.586 | 0.525 | 0.636 | 0.556 | 0.586 | 0.574 | 0.613 | 0.537 | 0.608 | 0.634 |
| sciq_test | GPT-4o | 0.720 | 0.615 | 0.641 | 0.587 | 0.686 | 0.619 | 0.653 | 0.638 | 0.660 | 0.632 | 0.650 | 0.690 |

Table 7: Zero-shot accuracy on MMLU for different models and retrieval sources.

| Dataset | Model | plain | Pubmed | Wiki | Pes2o | C4 | Github | Math | SE | Book | Arxiv | CC | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mmlu | Llama-3.2-3B | 0.481 | 0.480 | 0.534 | 0.511 | 0.525 | 0.477 | 0.499 | 0.468 | 0.526 | 0.471 | 0.532 | 0.552 |
| mmlu | Llama-3.1-8B | 0.530 | 0.528 | 0.576 | 0.574 | 0.580 | 0.547 | 0.565 | 0.521 | 0.590 | 0.554 | 0.588 | 0.606 |
| mmlu | Qwen3-4B | 0.660 | 0.635 | 0.667 | 0.650 | 0.663 | 0.633 | 0.647 | 0.635 | 0.663 | 0.627 | 0.665 | 0.689 |
| mmlu | Qwen3-8B | 0.689 | 0.668 | 0.688 | 0.679 | 0.689 | 0.670 | 0.686 | 0.666 | 0.689 | 0.658 | 0.697 | 0.713 |
| mmlu | Qwen3-32B | 0.778 | 0.736 | 0.767 | 0.759 | 0.767 | 0.744 | 0.764 | 0.745 | 0.768 | 0.742 | 0.773 | 0.783 |
| mmlu | GPT-4o-mini | 0.746 | 0.726 | 0.728 | 0.727 | 0.731 | 0.719 | 0.733 | 0.720 | 0.730 | 0.716 | 0.732 | 0.741 |
| mmlu | GPT-4o | 0.833 | 0.821 | 0.823 | 0.820 | 0.823 | 0.824 | 0.824 | 0.823 | 0.823 | 0.823 | 0.826 | 0.828 |

Table 8: Zero-shot accuracy on MMLU-Pro across different models and retrieval sources.

| Dataset | Model | plain | Pubmed | Wiki | Pes2o | C4 | Github | Math | SE | Book | Arxiv | CC | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mmlu_pro | Llama-3.2-3B | 0.261 | 0.254 | 0.297 | 0.273 | 0.265 | 0.236 | 0.248 | 0.239 | 0.261 | 0.244 | 0.269 | 0.293 |
| mmlu_pro | Llama-3.1-8B | 0.301 | 0.308 | 0.356 | 0.344 | 0.352 | 0.325 | 0.350 | 0.317 | 0.363 | 0.324 | 0.365 | 0.389 |
| mmlu_pro | Qwen3-4B | 0.411 | 0.399 | 0.428 | 0.411 | 0.423 | 0.388 | 0.413 | 0.397 | 0.426 | 0.384 | 0.429 | 0.452 |
| mmlu_pro | Qwen3-8B | 0.446 | 0.428 | 0.450 | 0.438 | 0.451 | 0.427 | 0.458 | 0.435 | 0.456 | 0.420 | 0.461 | 0.482 |
| mmlu_pro | Qwen3-32B | 0.540 | 0.496 | 0.533 | 0.518 | 0.525 | 0.517 | 0.532 | 0.523 | 0.534 | 0.513 | 0.540 | 0.552 |
| mmlu_pro | GPT-4o-mini | 0.438 | 0.412 | 0.430 | 0.421 | 0.424 | 0.408 | 0.430 | 0.408 | 0.432 | 0.402 | 0.434 | 0.448 |
| mmlu_pro | GPT-4o | 0.554 | 0.544 | 0.547 | 0.547 | 0.548 | 0.548 | 0.547 | 0.546 | 0.552 | 0.547 | 0.549 | 0.556 |

Table 9: Zero-shot accuracy on CSBench across different models and retrieval sources.

| Dataset | Model | plain | Pubmed | Wiki | Pes2o | C4 | Github | Math | SE | Book | Arxiv | CC | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| csbench | Llama-3.2-3B | 0.440 | 0.414 | 0.485 | 0.460 | 0.474 | 0.466 | 0.480 | 0.474 | 0.478 | 0.460 | 0.477 | 0.489 |
| csbench | Llama-3.1-8B | 0.472 | 0.435 | 0.528 | 0.506 | 0.529 | 0.514 | 0.536 | 0.507 | 0.527 | 0.500 | 0.537 | 0.542 |
| csbench | Qwen3-4B | 0.653 | 0.576 | 0.639 | 0.629 | 0.634 | 0.597 | 0.642 | 0.629 | 0.630 | 0.607 | 0.641 | 0.661 |
| csbench | Qwen3-8B | 0.698 | 0.602 | 0.649 | 0.638 | 0.661 | 0.642 | 0.667 | 0.645 | 0.652 | 0.631 | 0.663 | 0.673 |
| csbench | Qwen3-32B | 0.760 | 0.676 | 0.719 | 0.706 | 0.723 | 0.715 | 0.715 | 0.718 | 0.721 | 0.700 | 0.721 | 0.722 |
| csbench | GPT-4o-mini | 0.691 | 0.657 | 0.687 | 0.678 | 0.680 | 0.682 | 0.685 | 0.685 | 0.676 | 0.686 | 0.686 | 0.682 |
| csbench | GPT-4o | 0.749 | 0.737 | 0.737 | 0.740 | 0.748 | 0.742 | 0.741 | 0.742 | 0.743 | 0.747 | 0.748 | 0.757 |

Table 10: Zero-shot accuracy on SimpleQA with reranked retrieval sources.

| Dataset | Model | Pubmed | Wiki | Pes2o | C4 | Github | Math | SE | Book | Arxiv | CC | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| simpleqa | Llama-3.2-3B | 0.013 | 0.273 | 0.022 | 0.066 | 0.015 | 0.026 | 0.017 | 0.031 | 0.014 | 0.098 | 0.320 |
| simpleqa | Llama-3.1-8B | 0.035 | 0.290 | 0.048 | 0.088 | 0.043 | 0.047 | 0.041 | 0.052 | 0.037 | 0.122 | 0.309 |
| simpleqa | Qwen3-4B | 0.012 | 0.306 | 0.025 | 0.082 | 0.021 | 0.027 | 0.021 | 0.035 | 0.017 | 0.120 | 0.364 |
| simpleqa | Qwen3-8B | 0.012 | 0.312 | 0.030 | 0.091 | 0.019 | 0.030 | 0.025 | 0.044 | 0.019 | 0.134 | 0.382 |
| simpleqa | Qwen3-32B | 0.015 | 0.313 | 0.030 | 0.090 | 0.023 | 0.034 | 0.026 | 0.044 | 0.020 | 0.133 | 0.369 |
| simpleqa | GPT-4o-mini | 0.078 | 0.354 | 0.096 | 0.143 | 0.088 | 0.103 | 0.093 | 0.107 | 0.098 | 0.175 | 0.417 |
| simpleqa | GPT-4o | 0.260 | 0.420 | 0.252 | 0.261 | 0.250 | 0.262 | 0.259 | 0.253 | 0.265 | 0.285 | 0.477 |

Table 11: Zero-shot accuracy on ARC-Challenge with reranked retrieval sources.

| Dataset | Model | Pubmed | Wiki | Pes2o | C4 | Github | Math | SE | Book | Arxiv | CC | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| arc_c | Llama-3.2-3B | 0.586 | 0.642 | 0.600 | 0.641 | 0.570 | 0.612 | 0.568 | 0.621 | 0.564 | 0.639 | 0.658 |
| arc_c | Llama-3.1-8B | 0.692 | 0.718 | 0.709 | 0.730 | 0.676 | 0.715 | 0.664 | 0.729 | 0.691 | 0.732 | 0.755 |
| arc_c | Qwen3-4B | 0.817 | 0.824 | 0.829 | 0.842 | 0.816 | 0.848 | 0.811 | 0.837 | 0.812 | 0.848 | 0.848 |
| arc_c | Qwen3-8B | 0.851 | 0.856 | 0.846 | 0.864 | 0.852 | 0.878 | 0.846 | 0.868 | 0.835 | 0.859 | 0.860 |
| arc_c | Qwen3-32B | 0.902 | 0.907 | 0.897 | 0.916 | 0.911 | 0.917 | 0.905 | 0.911 | 0.888 | 0.913 | 0.914 |
| arc_c | GPT-4o-mini | 0.903 | 0.904 | 0.900 | 0.901 | 0.903 | 0.903 | 0.902 | 0.904 | 0.899 | 0.904 | 0.902 |
| arc_c | GPT-4o | 0.942 | 0.933 | 0.936 | 0.939 | 0.943 | 0.943 | 0.945 | 0.945 | 0.938 | 0.942 | 0.944 |

Table 12: Zero-shot accuracy on SciQ with reranked retrieval sources.

| Dataset | Model | Pubmed | Wiki | Pes2o | C4 | Github | Math | SE | Book | Arxiv | CC | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| sciq_test | Llama-3.2-3B | 0.363 | 0.531 | 0.420 | 0.577 | 0.342 | 0.514 | 0.372 | 0.505 | 0.315 | 0.575 | 0.654 |
| sciq_test | Llama-3.1-8B | 0.418 | 0.543 | 0.450 | 0.625 | 0.425 | 0.581 | 0.436 | 0.581 | 0.364 | 0.616 | 0.678 |
| sciq_test | Qwen3-4B | 0.449 | 0.566 | 0.480 | 0.629 | 0.477 | 0.576 | 0.470 | 0.578 | 0.428 | 0.627 | 0.669 |
| sciq_test | Qwen3-8B | 0.473 | 0.551 | 0.497 | 0.623 | 0.487 | 0.593 | 0.502 | 0.579 | 0.450 | 0.620 | 0.689 |
| sciq_test | Qwen3-32B | 0.482 | 0.602 | 0.536 | 0.638 | 0.507 | 0.630 | 0.512 | 0.626 | 0.460 | 0.655 | 0.698 |
| sciq_test | GPT-4o-mini | 0.551 | 0.589 | 0.515 | 0.620 | 0.557 | 0.593 | 0.567 | 0.593 | 0.523 | 0.628 | 0.657 |
| sciq_test | GPT-4o | 0.619 | 0.625 | 0.587 | 0.672 | 0.629 | 0.656 | 0.628 | 0.651 | 0.624 | 0.665 | 0.695 |

Table 13: Zero-shot accuracy on MMLU with reranked retrieval sources.

| Dataset | Model | Pubmed | Wiki | Pes2o | C4 | Github | Math | SE | Book | Arxiv | CC | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mmlu | Llama-3.2-3B | 0.480 | 0.535 | 0.505 | 0.523 | 0.469 | 0.499 | 0.468 | 0.523 | 0.459 | 0.528 | 0.559 |
| mmlu | Llama-3.1-8B | 0.526 | 0.582 | 0.567 | 0.576 | 0.535 | 0.557 | 0.523 | 0.577 | 0.540 | 0.579 | 0.612 |
| mmlu | Qwen3-4B | 0.634 | 0.658 | 0.648 | 0.665 | 0.631 | 0.646 | 0.627 | 0.661 | 0.623 | 0.670 | 0.685 |
| mmlu | Qwen3-8B | 0.658 | 0.685 | 0.673 | 0.689 | 0.665 | 0.679 | 0.663 | 0.686 | 0.652 | 0.697 | 0.710 |
| mmlu | Qwen3-32B | 0.733 | 0.763 | 0.755 | 0.759 | 0.743 | 0.759 | 0.743 | 0.762 | 0.738 | 0.768 | 0.780 |
| mmlu | GPT-4o-mini | 0.727 | 0.733 | 0.728 | 0.729 | 0.720 | 0.730 | 0.724 | 0.730 | 0.717 | 0.734 | 0.741 |
| mmlu | GPT-4o | 0.822 | 0.820 | 0.820 | 0.823 | 0.822 | 0.826 | 0.823 | 0.821 | 0.821 | 0.824 | 0.827 |

Table 14: Zero-shot accuracy on MMLU-Pro with reranked retrieval sources.

| Dataset | Model | Pubmed | Wiki | Pes2o | C4 | Github | Math | SE | Book | Arxiv | CC | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mmlu_pro | Llama-3.2-3B | 0.260 | 0.300 | 0.279 | 0.279 | 0.245 | 0.261 | 0.251 | 0.273 | 0.244 | 0.280 | 0.297 |
| mmlu_pro | Llama-3.1-8B | 0.310 | 0.362 | 0.337 | 0.351 | 0.318 | 0.345 | 0.315 | 0.358 | 0.316 | 0.361 | 0.395 |
| mmlu_pro | Qwen3-4B | 0.398 | 0.424 | 0.408 | 0.423 | 0.388 | 0.419 | 0.396 | 0.423 | 0.381 | 0.425 | 0.454 |
| mmlu_pro | Qwen3-8B | 0.420 | 0.449 | 0.435 | 0.451 | 0.424 | 0.454 | 0.431 | 0.457 | 0.418 | 0.458 | 0.485 |
| mmlu_pro | Qwen3-32B | 0.499 | 0.527 | 0.518 | 0.527 | 0.514 | 0.529 | 0.520 | 0.535 | 0.506 | 0.535 | 0.553 |
| mmlu_pro | GPT-4o-mini | 0.412 | 0.427 | 0.417 | 0.423 | 0.403 | 0.429 | 0.408 | 0.430 | 0.400 | 0.433 | 0.459 |
| mmlu_pro | GPT-4o | 0.544 | 0.547 | 0.544 | 0.546 | 0.551 | 0.552 | 0.548 | 0.551 | 0.546 | 0.549 | 0.558 |

Table 15: Zero-shot accuracy on CSBench with reranked retrieval sources.

| Dataset | Model | Pubmed | Wiki | Pes2o | C4 | Github | Math | SE | Book | Arxiv | CC | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| csbench | Llama-3.2-3B | 0.412 | 0.478 | 0.458 | 0.480 | 0.462 | 0.471 | 0.463 | 0.466 | 0.447 | 0.473 | 0.501 |
| csbench | Llama-3.1-8B | 0.427 | 0.528 | 0.497 | 0.523 | 0.511 | 0.525 | 0.520 | 0.533 | 0.507 | 0.541 | 0.547 |
| csbench | Qwen3-4B | 0.571 | 0.644 | 0.624 | 0.643 | 0.613 | 0.655 | 0.623 | 0.643 | 0.596 | 0.634 | 0.658 |
| csbench | Qwen3-8B | 0.602 | 0.658 | 0.638 | 0.661 | 0.650 | 0.664 | 0.643 | 0.651 | 0.619 | 0.669 | 0.676 |
| csbench | Qwen3-32B | 0.684 | 0.713 | 0.694 | 0.714 | 0.712 | 0.726 | 0.728 | 0.727 | 0.688 | 0.728 | 0.722 |
| csbench | GPT-4o-mini | 0.668 | 0.683 | 0.679 | 0.679 | 0.680 | 0.681 | 0.691 | 0.690 | 0.682 | 0.686 | 0.691 |
| csbench | GPT-4o | 0.733 | 0.744 | 0.736 | 0.750 | 0.737 | 0.744 | 0.739 | 0.746 | 0.741 | 0.748 | 0.746 |