Evaluating the Effectiveness of Retrieval-Augmented Large Language Models in Scientific Document Reasoning

Sai Munikoti* **Anurag Acharya** **Sridevi Wagle** **Sameera Horawalavithana***

Pacific Northwest National Laboratory, Richland, WA, USA
firstname.lastname@pnnl.gov

Abstract

Despite the dramatic progress in Large Language Model (LLM) development, LLMs often provide seemingly plausible but not factual information, often referred to as hallucinations. Retrieval-augmented LLMs provide a non-parametric approach to solve these issues by retrieving relevant information from external data sources and augment the training process. These models help to trace evidence from an externally provided knowledge base allowing the model predictions to be better interpreted and verified. In this work, we critically evaluate these models in their ability to perform in scientific document reasoning tasks. To this end, we tuned multiple such model variants with science-focused instructions and evaluated them on a scientific document reasoning benchmark for the usefulness of the retrieved document passages. Our findings suggest that models justify predictions in science tasks with fabricated evidence and leveraging scientific corpus as pretraining data does not alleviate the risk of evidence fabrication.

1 Introduction

Large Language Models (LLMs) perform competitively in a majority of Natural Language Processing (NLP) tasks, but tend to hallucinate with seemingly plausible but misleading predictions (Mallen *et al.*, 2023; Jiang *et al.*, 2023) with no clear explanations or justifications for their predictions (Mialon *et al.*, 2023). Recently, retrieval-augmented LMs have helped address these issues by augmenting the LLMs with non-parametric memory by using neural retriever to extract relevant information from external knowledge resources like document corpora (Jiang *et al.*, 2023). Retrieving external knowledge helps models update with new knowledge, inject domain specific data, and memorize long-tail knowledge. These models are also relatively small in the number of parameters and require less training and inference costs (Borgeaud *et al.*, 2022).

While retrieval-augmented LMs (RALM) are shown to perform well on knowledge-intensive tasks (Izacard *et al.*, 2022), we have very limited understanding on their ability to perform on the science-focused downstream tasks. For example, we can provide scientific documents as external knowledge at test time, and test the ability of the model to perform on science question and answering (QA) task. In this setup, the model retrieves scientific documents relevant to the question, and then generates an answer based on the retrieved documents. Retrieved documents help the model predictions to be better interpreted and verified. At the same time, we can assess the trustworthiness of these models to understand whether they justify the model predictions with accurate and relevant evidence (Mallen *et al.*, 2023). Recognizing any failure modes is essential to ensuring the safe deployment, and avoiding potential risks or negative consequences of these models, specially across multiple scientific use cases and applications.

This work conducts an evaluation on RALMs to improve our understanding of these models to perform on science tasks. To this end, we used the ATLAS model architecture (Izacard *et al.*, 2022) as an instance of the retrieval-based language model family to drive our experiments. To adopt the models for science tasks, we provide a collection of scientific documents as external corpus during the pretraining, instruction tuning, and evaluation stages. We evaluate the model performance on SciRepEval (Singh *et al.*, 2022) benchmark to test whether model recognizes different scientific domains and disciplines from the given scientific documents. Our hypothesis is that these models will be able to retrieve relevant information from scientific documents, integrate knowledge from diverse scientific domains, and reason over complex scientific concepts. In particular, we evaluate the usefulness of

*These authors contributed equally to this work.

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 84–89

Table 1: Summary of different pretraining, instruction tuning and benchmark datasets used across T5 and ATLAS models. We report the performance of the standalone LLM i) T5 (pretrained with C4), retrieval-augmented LLMs, ii) ATLAS model (pretrained with CC and Wikipedia) and iii) ATLAS-Science model (pretrained with S2ORC) text datasets. We used the S2ORC dataset as the external retrieval corpus in the instruction tuning and evaluation stages to make a fair comparison.

Model	Pretraining		Instruction Tuning		Evaluation	
	Data	Retrieval Corpus	Data	Retrieval Corpus	Data	Retrieval Corpus
T5	C4	N/A	FOS	N/A	FOS	N/A
					MAG	
ATLAS	CC + Wiki	Wiki		S2ORC	FOS	S2ORC
					MAG	
ATLAS-Science	S2ORC	S2ORC			FOS	
					MAG	

retrieved passages in terms of their relevancy and diversity that support the model predictions.

2 Problem Formulation and Eval Setup

2.1 Problem Formulation

Previous research on Retrieval Augmented LLMs focused on solving three major research questions: i) what to retrieve (e.g., chunks, tokens), ii) how to retrieve (e.g., input, intermediary and output layers), and iii) when to retrieve (e.g., once every n$\geq$1 tokens). A majority of proposed models such as REALM (Guu *et al.*, 2020), DPR (Karpukhin *et al.*, 2020), RAG (Lewis *et al.*, 2020), and ATLAS (Izacard *et al.*, 2022) retrieve text chunks and concatenate them in the input layer of the language model. For example, ATLAS combines autoregressive text generation with retrieval-based language model pre-training based on the encoder-decoder architecture and fine-tuned on open-domain QA.

In this work, we aim to improve our understanding on the development of retrieval-based LMs for evidence extraction. We focus on the following related research questions to drive our experiments.

(RQ1) How useful are the evidences generated from retrieval-augmented LLMs to justify model predictions in science tasks?

(RQ2) How do the retrieval-augmented LLMs behave when provided with the scientific knowledge as the external document store?

2.2 Evaluation Setup

In this section, we outline the datasets, models, benchmarks and metrics used in our experiments.

Scientific Text Datasets Retrieval Augmented LLMs provide ideal test bed for scientific applications since they can handle dynamic knowledge updates and different scientific domains and disciplines than what the models see during the pretraining. We focus on evaluating the Retrieval Augmented LLMs on their ability to understand scientific language and retrieve from multiple scientific knowledge sources. We preprocess the S2ORC (Lo *et al.*, 2019) dataset to create a collection of 354M text passages. Each passage has a maximum of 512 tokens, or 100 words that are concatenated with the corresponding title of the document the passage belongs to. We record 19 different scientific domains in the S2ORC collection.[1]

Models Our experiments are based on ATLAS (220M) (Izacard *et al.*, 2022) model architecture unless explicitly mentioned. ATLAS uses the Fusion-in-decoder architecture to fuse the retrieved text chunks with the input queries during the pretraining. In addition to the ATLAS model pretrained with common crawl (CC) and Wikipedia, we also train ATLAS-Science (220M) model from scratch with the S2ORC scientific text datasets. For a fair comparison with ATLAS, we initialize the ATLAS-Science model with the *T5-lm-adapt* (Raffel *et al.*, 2020) model and trained jointly with retrieval model, *Contriever* (Izacard *et al.*, 2021). Figure 1 shows the overview of different components used in the ATLAS-Science model. We provide the collection of scientific text passages as external retrieval corpus. First, we encode the scientific text passages (354M) with the *Contriever* model, and construct a document index in the FLAT (Izacard *et al.*, 2022) mode for faster retrieval. Second, we use the same passages for model pretraining and

[1] S2ORC dataset covers 19 scientific domains; Art, Philosophy, Political-Science, Sociology, Psychology, Geography, History, Business, Economics, Geology, Physics, Chemistry, Biology, Mathematics, Computer Science, Engineering, Environmental science, Material science, Medicine

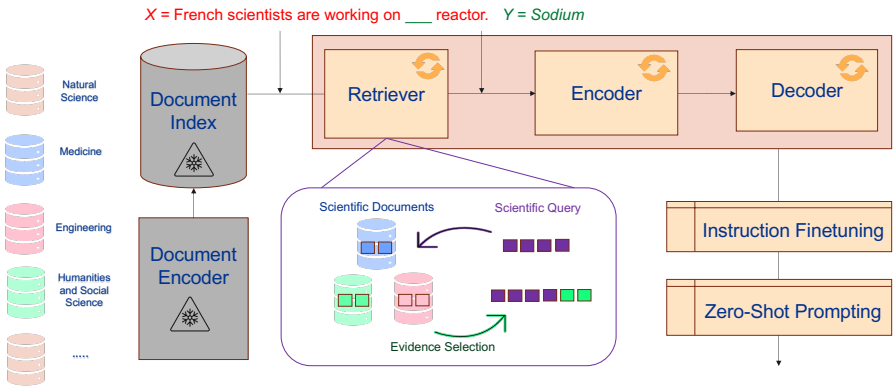

Figure 1: The main experimental setup of this research work, with all the major components displayed.

ensure that the passages used for pretraining are distinct from passages used to build the document index. Third, we train the retriever with the *query side finetuning* approach that originally introduced in the ATLAS model. This approach is very efficient in model training since it keeps the document encoder frozen while training the parameters corresponding to the query encoder (Figure 1). All the models are trained for the same number of tokens for a fair comparison.

Instruction Tuning We use the SciRepEval (Singh *et al.*, 2022) benchmark for training and evaluating the models for the scientific evidence extraction. SciRepEval provides 25 challenging tasks across four formats: classification, regression, ranking, and search. In this work, we focus on the classification tasks, *Fields of study (FoS)*[2] and *MAG* due to two main reasons. First, we need benchmark tasks that test the ability of the models to understand diverse scientific domains and disciplines. For example, FoS task tests the ability of the model to recognize which domain the given text passage belongs to. Second, we want to evaluate on specific instruction template to avoid any prompting bias. So we used the following instruction template:

```
"## Below is an input containing a title-abstract pair. Classify this input into one or more possible Field of Study categories. ### Possible Categories: [..] ### Input: ## Title: .. ## Response:"
```

[2]FoS tasks include instructions from following domains: Materials science, Economics, Chemistry, Medicine, Psychology, Geography, Geology, Political science, Engineering, Philosophy, Sociology, Physics, Computer science, Law, History, Biology, Agricultural and Food sciences, Environmental science, Business, Education, Art, Linguistics, Mathematics

Previous research (Izacard *et al.*, 2022) has shown that ATLAS model is able to learn knowledge intensive tasks with very few training examples (aka few shot learning). To allow the model to perform on the downstream tasks, we tune the model with scientific instructions. We design an instruction template to guide the model to generate the scientific domain that each passage belongs to. We tune the model with *Fields of study (FoS)* training data after converting them to instructions. This process resulted 541,218 training instructions that used to perform instruction tuning. For a fair comparison, we tune the T5 and ATLAS models with these instructions. In comparison to the T5 model, ATLAS models retrieve the top-k relevant passages from the S2ORC document store to augment the instruction tuning process. There are 68,147 and 3,751 test instructions in the FoS and MAG tasks, respectively. We use MAG instructions to test the out-of-distribution task performance. Table 1 summarizes the pretraining, instruction tuning and evaluation data used for the ATLAS and ATLAS-Science models.

Metrics We use the Exact Match (EM) and F1-Score to evaluate the task accuracy. EM metric evaluates the exact token overlap between the ground truth and generated answers. In addition, we design two metrics to evaluate the relevance and diversity of the extracted evidence: the *relevance* and *diversity* metrics. The relevance metric calculates the ratio of the domains present in Top-k evidences matching with the scientific domain corresponding to the query. The diversity metric calculates the ratio of the unique evidences in comparison to the total evidences. Both metrics are in the range of zero and one, with higher the metric scores, higher the quality in the generated evidences.

Table 2: Model ablation study to evaluate performance on in-distribution (SciDocs-FoS) and out-of-distribution (SciDocs-MAG) field of study instruction tuning datasets.

Model	In-distribution Performance				Out-of-distribution Performance			
	Accuracy		Evidence Generation		Accuracy		Evidence Generation	
	EM	F1	Relevance	Diversity	EM	F1	Relevance	Diversity
T5	83.33	0.87	N/A	N/A	57.90	0.72	N/A	N/A
ATLAS	84.42	0.92	0.06	5E-5	59.10	0.75	0.07	60E-5
ATLAS-Science	84.70	0.92	0.05	8E-5	57.80	0.73	0.05	100E-5

3 Measuring the Effectiveness of the Scientific Evidence Extraction

In this section, we address (RQ1) and (RQ2) by evaluating the usefulness of the evidence extracted from RALMs in performing science tasks. We test how retrieval-augmented LLMs behave when provided the scientific knowledge as external memory.

Retrieval-augmented LLMs justify model predictions in science tasks with fabricated evidence We evaluate the pretrained ATLAS model (Izacard *et al.*, 2022) on the benchmark tasks *Fields of study (FoS)* and *MAG*. Additionally, we tune ATLAS model with the *Fields of study (FoS)* instructions (as described in Section 2.2). We use the S2ORC document index to evaluate the instructions tuned ATLAS model in the zero-shot prompting stage. We report the performance of ATLAS model in Table 2. First, we observe that the accuracy of the ATLAS model is better than that of T5 in both the tasks, demonstrating the importance of retrieval augmentation. Second, we observe that although the ATLAS model has achieved a acceptable accuracy of 84.42% in *FOS* and 59.10% in *MAG*, the retrieved evidences are extremely poor in terms of relevance to the query topic. The model only achieves 0.06 relevance score suggesting that the passages returned by the model as evidences do not align with the domain of the query. For example, ATLAS model returns the passages in Geology, History and Chemistry as evidence for a Biology query as shown in Figure 2. The retrieved passages are not at all related to the corresponding query topics, rendering them useless. Finally, we also evaluate the faithfulness of the retrieved docs in terms of diversity score. This score is very low, suggesting that the evidences remain similar across all test queries. Our qualitative check suggests that the top-20 passages returned by the model for different queries are exactly same, while the generated answers are different and mostly accurate in comparison to the ground truth. These observations suggest that the ATLAS model fabricates the evidence to justify the model predictions.

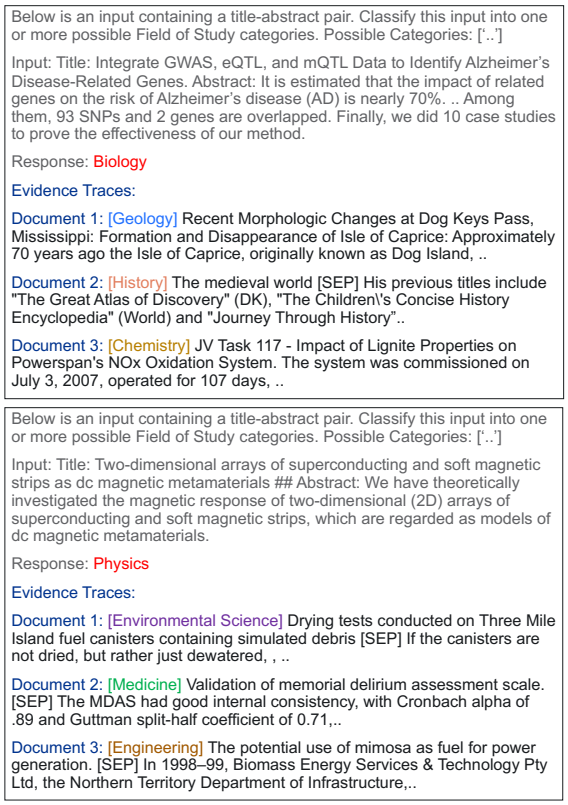

Figure 2: Example generations of the ATLAS instruction tuned model in the SciRepEval-FoS (Singh *et al.*, 2022) task. We color the input query in gray, and the generated answer in red. We list three documents returned by the model as evidence to support the answer. We annotate each document by the corresponding scientific domain. For example, the model accurately generates the Biology domain that the passage belongs to, but justifies the answer with fabricated evidence as retrieved passages are in Geology, History and Chemistry.

Scientific knowledge provided as pretraining data does not alleviate the evidence fabrication To explore the impact of the pretraining data on downstream scientific tasks, we repeat the evaluation with the *ATLAS-Science* model (as described in Section 2.2). Note that the ATLAS-Science model

is pretrained from scratch with S2ORC scientific text data provided as both pretraining and external document store. We evaluate the ATLAS-Science on two benchmark tasks. The results are tabulated in third row of Table 2. In comparison to ATLAS, the accuracy of ATLAS-Science model has a slight improvement in *FOS*, whereas it depreciates in *MAG*. Thus we see that scientific corpus does not help much in improving the performance of the model. More importantly, the relevance and diversity of the retrieved passages only slightly improves over ATLAS. This indicates that leveraging scientific corpus as pretraining data is not an effective approach to address the challenge of evidence fabrication.

4 Related Work

LLMs can be augmented with various external modules such as document corpus (Gao *et al.*, 2020), vector databases[3], etc. Typically, the augmentation is accomplished in two ways, namely sparse (such as Bag of words) and dense where Neural network is used to encode documents. Dense retrievers are widely used in present time mainly due to the good representation capability of neural networks. Recent works suggest that the retrieval-augmented LLMs has significant improvement over the standard LLMs across various tasks especially with respect to scale (Guu *et al.*, 2020; Lewis *et al.*, 2020). REALM and RAG are the initial efforts where they train the retriever and language model by representing documents as latent variable, and minimizing the language model objective (Guu *et al.*, 2020; Lewis *et al.*, 2020). REALM leverages masked-language modeling as an objective to pretrain the model in end to end fashion. However, it is computationally very expensive to train these models that requires to retrain the entire index with new knowledge. Guu *et al.* (2020) explored the concept of query-side finetuning that only refreshes the query encoder whereas document encoder remains frozen. Izacard et al. (Izacard & Grave, 2020; Izacard *et al.*, 2021) proposed various ways to improve the retrieval augmented models, including novel learning objectives to align retriever with the language model (Izacard *et al.*, 2021). Furthermore, RETRO (Borgeaud *et al.*, 2022) shows the benefits of scaling the retrieval memory to trillions of tokens. ATLAS (Izacard *et al.*, 2022) experiments with various design and training configurations for retrieval augmented models with a focus on few shot learning ability, while ATLANTIC (Munikoti *et al.*, 2024) shows an approach to utilize heterogeneous graph information to create a structure-aware RAG technique. Finally, Wagle *et al.* (2024) provides a word of caution, finding that while RAG-based language models tend to be more confident if fine-tuned on scientific documents, they also exhibit false confidence even for incorrect predictions.

[3]https://github.com/jerryjliu/llama_index

5 Conclusion

In this study, we explored the efficacy of retrieval augmented language models on science tasks. Our experiments were based on ATLAS model which is a state of the art retrieval augmented language model with few shot capability. We performed a systematic evaluation on the performance of different ATLAS model variants in two scientific document reasoning tasks. Our experiments on the pretrained ATLAS model reveal that the model demonstrates acceptable performance in science tasks but the evidences are fabricated. We also observe that pretraining the model with scientific corpus does not alleviate evidence fabrication. We plan to develop techniques to alleviate these issues in a future work.

Acknowledgements

This work was supported by the NNSA Office of Defense Nuclear Nonproliferation Research and Development, U.S. Department of Energy, and Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DE-AC05–76RLO1830. This article has been cleared by PNNL for public release as PNNL-SA-189029.

References

Borgeaud, Sebastian, Mensch, Arthur, Hoffmann, Jordan, Cai, Trevor, Rutherford, Eliza, Millican, Katie, Van Den Driessche, George Bm, Lespiau, Jean-Baptiste, Damoc, Bogdan, Clark, Aidan, *et al.* 2022. Improving language models by retrieving from trillions of tokens. *Pages 2206–2240 of: International conference on machine learning*. PMLR.

Gao, Leo, Biderman, Stella, Black, Sid, Golding, Laurence, Hoppe, Travis, Foster, Charles, Phang, Jason, He, Horace, Thite, Anish, Nabeshima, Noa, *et al.* 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Guu, Kelvin, Lee, Kenton, Tung, Zora, Pasupat, Panupong, & Chang, Mingwei. 2020. Retrieval augmented language model pre-training. *Pages 3929–3938 of: International conference on machine learning*. PMLR.

Izacard, Gautier, & Grave, Edouard. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Izacard, Gautier, Caron, Mathilde, Hosseini, Lucas, Riedel, Sebastian, Bojanowski, Piotr, Joulin, Armand, & Grave, Edouard. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Izacard, Gautier, Lewis, Patrick, Lomeli, Maria, Hosseini, Lucas, Petroni, Fabio, Schick, Timo, Dwivedi-Yu, Jane, Joulin, Armand, Riedel, Sebastian, & Grave, Edouard. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Jiang, Zhengbao, Xu, Frank F, Gao, Luyu, Sun, Zhiqing, Liu, Qian, Dwivedi-Yu, Jane, Yang, Yiming, Callan, Jamie, & Neubig, Graham. 2023. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

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

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

Lo, Kyle, Wang, Lucy Lu, Neumann, Mark, Kinney, Rodney, & Weld, Dan S. 2019. S2ORC: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

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

Mialon, Grégoire, Dessì, Roberto, Lomeli, Maria, Nalmpantis, Christoforos, Pasunuru, Ram, Raileanu, Roberta, Rozière, Baptiste, Schick, Timo, Dwivedi-Yu, Jane, Celikyilmaz, Asli, *et al.* 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Munikoti, Sai, Acharya, Anurag, Wagle, Sridevi, & Horawalavithana, Sameera. 2024 (February). ATLANTIC: Structure-Aware Retrieval-Augmented Language Model for Interdisciplinary Science. *In: Proceedings of the Workshop on AI to Accelerate Science and Engineering (AI2ASE)*. Held in conjunction with the 38th AAAI Conference on Artificial Intelligence (AAAI 2024).

Raffel, Colin, Shazeer, Noam, Roberts, Adam, Lee, Katherine, Narang, Sharan, Matena, Michael, Zhou, Yanqi, Li, Wei, & Liu, Peter J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, **21**(1), 5485–5551.

Singh, Amanpreet, D'Arcy, Mike, Cohan, Arman, Downey, Doug, & Feldman, Sergey. 2022. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. *arXiv preprint arXiv:2211.13308*.

Wagle, Sridevi, Munikoti, Sai, Acharya, Anurag, Smith, Sara, & Horawalavithana, Sameera. 2024 (February). Empirical evaluation of uncertainty quantification in retrieval-augmented language models for science. *In: Proceedings of the Workshop on Scientific Document Understanding (SDU)*. Held in conjunction with the 38th AAAI Conference on Artificial Intelligence (AAAI 2024).