Evaluating and Fine-Tuning Retrieval-Augmented Language Models to Generate Text With Accurate Citations

Vinzent Penzkofer and Timo Baumann
Faculty of Informatics and Mathematics, OTH Regensburg
vinzent.penzkofer@outlook.de and timo.baumann@oth-regensburg.de

Abstract

Retrieval Augmented Generation (RAG) is becoming an essential tool for easily accessing large amounts of textual information. However, it is often challenging to determine whether the information in a given response originates from the retrieved context, the training, or is a result of hallucination. Our contribution in this area is twofold. Firstly, we demonstrate how existing datasets for information retrieval evaluation can be used to assess the ability of Large Language Models (LLMs) to correctly identify relevant sources. Our findings indicate that there are notable discrepancies in the performance of different current LLMs in this task. Secondly, we utilise the datasets and metrics for citation evaluation to enhance the citation quality of small open-weight LLMs through fine-tuning. We achieve significant performance gains in this task, matching the results of much larger models.

1 Introduction

In Retrieval Augmented Generation (RAG) (Lewis et al., 2020) the generation process of a language model is augmented at inference time with additional textual information retrieved from a corpus of documents. This approach aims to factually ground LLMs, reduce hallucination and provide access to information after the knowledge cut-off of the language model (Lewis et al., 2020).

Our focus is on the evaluation and improvement of RAG systems. We believe that it is necessary to correctly reference the information used for answer generation in order to make the factual accuracy verifiable by users in a practical setting. While there are ways to evaluate retrieval performance (Thakur et al., 2021; Muennighoff et al., 2023) and also factual correctness (Es et al., 2024; Chen et al., 2024), we see a research gap in evaluating the ability of models to correctly reference their sources. In this paper we present RAGE (Retrieval Augmented Generation Evaluation), a framework focused on evaluating the citation performance of language models used for RAG. Furthermore, we show how the citation evaluation metrics of RAGE can be used to directly improve the citation quality through fine-tuning.

2 Related Work

Several works have focused on the evaluation of RAG systems. Es et al. (2024) evaluate several different aspects of RAG including faithfulness, answer relevance and context relevance. Chen et al. (2024) propose a benchmark focusing on noise robustness, negative rejection, information integration and counterfactual robustness. Neither consider the attribution of referenced documents.

Gao et al. (2023) provide insights into how RAG systems can be prompted to generate text with citations. They also present a way of assessing the citation quality of LLMs, which includes the use of entailment models to assess the entailment of generated model responses and cited passages. Our work differs by using lightweight information retrieval datasets for citation evaluation and by having a clearly structured dataset format, making it more adaptable to specific use-cases.

Concurrently to us, Li et al. (2024) research fine-tuning to improve source attribution in RAG and developed a somewhat similar approach to ours. They also use Supervised Fine Tuning (SFT) for aligning model responses to a desired format, but do so using public datasets rather than generating new synthetic data as we do. We argue that the use of synthetic data makes the process more adaptable to specific use cases. They use preference optimisation (Rafailov et al., 2023) to optimise for citation quality, whereas we directly use citation quality metrics as a reward function for Proximal Policy Optimization (PPO), which can be automated more directly.

3 The RAGE Framework

In this section we describe RAGE (Retrieval Augmented Generation Evaluation), our automatic evaluation framework for RAG systems.[1] RAGE is designed to assess the performance of a RAG system in correctly referencing the documents it used for answer generation.

Typically RAG involves two steps, retrieval of relevant documents and generation augmented with the relevant texts.

RAGE specializes in assessing the augmented generation component, specifically its ability to cite its sources. We define this component as any system that takes in a query with a list of documents and generates an answer to the query whilst referencing the documents used for answer generation.

The fundamental idea of RAGE is to present augmented generation systems with a query accompanied by both relevant and irrelevant documents, then assessing the systems' capability to accurately identify and cite the relevant sources.

3.1 Datasets

The design of RAGE is based on ideas from the evaluation of Information Retrieval (IR) systems. IR systems are typically evaluated using datasets consisting of three distinct components: a corpus of documents, a set of queries, and a mapping table that indicates for each query the relevance of some specific documents (Thakur et al., 2021).

For RAGE, we extend this dataset structure with two additional mapping tables. We introduce a mapping of queries to **irrelevant** and to **seemingly relevant** documents in addition to the mapping of **relevant documents**. Documents are *seemingly* relevant when they appear as if they may contain the information necessary to answer a given query but don't actually do. This results in three distinct mapping tables in addition to the documents and queries.

We base our experiments on the Natural Questions (Kwiatkowski et al., 2019) dataset which was designed for the question answering domain and use the version adjusted for information retrieval by Thakur et al. (2021).[2] We argue that datasets designed for question answering are well-suited for evaluating RAG systems due to the typical application of RAG in this domain.

We create the mapping of irrelevant documents by randomly sampling the document corpus while excluding the relevant documents for a given query.

For the mapping of seemingly relevant documents, we generate a vector representation of all documents and queries using the multilingual-e5-small embedding model (Wang et al., 2024). Subsequently, for each query, we compare its embedding to all the document embeddings using an L2 similarity measure, while again excluding the relevant documents. The ten documents that show the highest similarity to the given query were mapped.

Other IR datasets can trivially be converted to the required format by scripts that are part of RAGE[3].

3.2 Procedure

The evaluation process employed in RAGE follows two steps.

Step 1: Create a relevancy mixture of documents. For each query, a mixture of relevant, irrelevant and seemingly relevant documents is created. The proportions of relevant, irrelevant, and seemingly relevant documents in the mixture for each query can be freely adjusted in an evaluation run. A prompt is generated containing processing instructions, the document mixture, and the query itself. The prompt is then passed to the augmented generation component under evaluation. An example of a prompt and a LLM response used in our experiments is given in Appendix A.

Step 2: Analyze LLM answer and compute performance metrics. The LLM response is analysed w.r.t. various performance metrics including *Citation-Precision*, *Citation-Recall*, the number of *Distinct Citations*, and *Response Length*.

Citation-Precision is defined as the ratio of relevant citations to the total number of citations within the response. Similarly, *Citation-Recall* is determined by the ratio of relevant distinct citations to the total number of relevant documents included in the document mixture during the first step. *Response Length* is measured by the total number of words, and finally *Distinct Citations* counts the unique citations within the response. Additionally, the harmonic mean of *Citation-Precision* and *Citation-Recall* yields the *F1-Score*.

[1] The codebase is available at https://github.com/othr-nlp/rage_toolkit.

[2] We have also experimented with the HotpotQA (Yang et al., 2018) dataset. That dataset yields similar results which we here omit for brevity.

[3] Some datasets already converted to the RAGE format are available at https://huggingface.co/othr-nlp.

Model	F1 Score	Precision	Recall	Answer Length	Cited Distinct
Baseline	.16	.14	.19	-	1.47
LLaMA 2 7B	.47	.41	.55	77.7	1.88
LLaMA 2 13B	.45	.40	.51	67.6	1.63
LLaMA 2 70B	.66	.64	.67	41.3	1.40
Mistral 7B	.61	.51	.77	45.8	2.17
Mixtral 8x7B	.73	.64	.85	41.8	1.83
GPT 3.5	.78	.75	.80	18.4	1.34
GPT 4	.82	.81	.83	21.1	1.29

Table 1: RAGE evaluation results for different state-of-the-art LLMs evaluated on the Natural Questions (Kwiatkowski et al., 2019) Dataset from the BEIR Benchmark (Thakur et al., 2021). As a baseline we include an augmented generation system which randomly cites 1-3 of the provided documents.

The metrics are calculated for each query and then averaged to determine the final scores for a given evaluation dataset.

3.3 Evaluation Setup

In our experiments, we evaluate citation performance of some state-of-the-art LLMs. To achieve this, we first combine the query and the mixture of relevant, irrelevant, seemingly relevant documents into a prompt that is then to be passed to the LLM in question. For our experiments, we used 1-3 relevant, 3 irrelevant and 3 seemingly relevant documents for all runs.

The prompt furthermore contains processing instructions which state to use only the information contained within the documents and to cite in a predefined format. Prompt generation is identical for all augmented generation components (and has not undergone much prompt engineering). For an example of prompt, query and LLM response, we again refer to Appendix A.

We selected LLMs of differing model size in terms of parameters, availability (open- or closed-weight) and performance on common benchmarks for our trial run of RAGE.

The evaluation was performed on the Natural Questions dataset (Kwiatkowski et al., 2019) with the described adaptations (Section 3.1).

3.4 Results

The results of our evaluation are shown in Table 1. Baseline performance is significantly surpassed by all models, indicating an understanding of the task and citation format.

Smaller models tend to produce longer answers and more distinct citations which leads to good recall but poorer precision. There is a tendency for larger models to perform better.

GPT-3.5 and GPT-4 (OpenAI, 2023) perform best *out of the box* and produce short answers and few distinct citations, indicating concise responses.

To test the robustness of RAGE, we also conducted experiments with different proportions in the document mixtures. The results indicate that RAGE works consistently well across these variations, though higher proportions of seemingly relevant documents increase task difficulty. We included those results in Appendix B.

4 Fine-Tuning for Citation Quality

In this section, we describe our approach to fine-tune open-weight LLMs for improved citation quality. We use the metrics and datasets as defined above for RAGE and use synthetic target data produced by GPT-3.5.

4.1 High-level Approach

Our fine-tuning technique to improve citation quality is inspired by Ouyang et al. (2022). They fine-tune LLMs to follow human instructions by first applying SFT to align the model outputs to a desired format and subsequently using PPO (Schulman et al., 2017) to further align to human preferences. Similarly, our approach is also twofold:

Step 1: Use supervised fine-tuning (SFT) to align model outputs to a preferred answer format. The idea of SFT for language models is to continue the self-supervised next token prediction objective of the pretraining phase with labeled task-specific data. For our models, we use synthetic data from GPT-3.5, which showed a concise answer style

Model	Tuning	F1 Score	Precision	Recall	Answer Length	Cited Distinct
LLaMA 2 7B		.47	.41	.55	77.7	1.88
	SFT	.47	.47	.46	18.7	1.25
	PPO	.53	.43	**.68**	142.3	2.92
	SFT+PPO	**.70**	**.74**	.66	18.1	1.05
Mistral 7B		.61	.51	.77	45.8	2.17
	SFT	.56	.57	.55	20.6	1.19
	PPO	**.72**	.65	**.80**	40.5	1.68
	SFT+PPO	.70	**.74**	.66	16.4	1.02
GPT 3.5		.78	.75	.80	18.4	1.34

Table 2: Evaluation results for fine-tuned models evaluated on Natural Questions (Kwiatkowski et al., 2019). Base models and GPT-3.5 are included for comparison.

with high precision and recall, to adjust the answer format of the smaller models.

Step 2: Improve citation-quality with reinforcement learning via proximal policy optimization (PPO). Reinforcement learning is a useful approach for language model fine-tuning, as it requires only a quality measure of the generated sequences, known as the reward function, rather than labeled example responses. We use a reward function based on the RAGE evaluation metrics outlined above and the PPO algorithm to directly improve citation quality. The reward function and the datasets are described in more detail later. PPO is applied separately or on top of the SFT process.

4.2 Fine-Tuning Datasets

This section presents the composition of the datasets we used for SFT and PPO fine-tuning.

SFT: Inspired by Mukherjee et al. (2023), we used the performance gap of the small 7B models to GPT-3.5[4] to generate synthetic training data. As shown in Table 1, GPT-3.5 provides good precision and recall with a short answer length, making it ideal for aligning the smaller models. We used the Natural Questions (Kwiatkowski et al., 2019) dataset and the same process as in the evaluation to generate a set of prompts for GPT-3.5. We then collected the responses of GPT-3.5, combined them with the prompts and added model-specific special tokens to create the final SFT dataset. 250 queries of Natural Questions were withheld from the training dataset for evaluation, leaving a total of 3201 fine-tuning examples.

[4]The exact model version is gpt-3.5-turbo.

PPO: For PPO fine-tuning we also generated prompts as described in the evaluation section, each containing citation instructions, documents and query. We generated the prompts using the Natural Questions (Kwiatkowski et al., 2019) and HotpotQA (Yang et al., 2018) datasets, withholding 250 examples from each for evaluation, thereby compiling a training dataset of 10,347 examples.

4.3 PPO Reward Function

Instead of using a reward model for reward generation as done by Ouyang et al. (2022), we use a simple reward function by calculating the arithmetic mean of citation precision and citation recall:

$$\text{Reward} = \frac{\text{Recall} + \text{Precision}}{2}$$

This function directly rewards improved citation quality without the need for an expensive reward model training. To prevent the model from exploiting the reward function, we use a KL-penalty as described by Ouyang et al. (2022).

4.4 Experimental Details

We used the instruction fine-tuned versions of Llama2 7B (Touvron et al., 2023) and Mistral 7B (Jiang et al., 2023) as bases. For both base models, three versions were trained and evaluated: *SFT-only*, *PPO-only* and *PPO+SFT*. We use QLoRA (Dettmers et al., 2023) with 4-bit quantization and a rank of 64 for the adaptation matrices for both SFT and PPO. SFT was performed for three epochs and PPO for one epoch on the respective dataset.

4.5 Results and Discussion

The fine-tuned models are evaluated via RAGE using the 250 queries withheld from the fine-tuning

datasets. Results are shown in Table 2.

The PPO+SFT model versions show that fine-tuning leads to gains compared to the base models and they approach GPT-3.5's citation precision despite the significantly smaller model sizes. Mistral 7B PPO+SFT experiences a decrease in recall, likely attributable to the significantly shorter answer lengths imposed by SFT. Mistral 7B PPO-only achieves the highest scores in terms of F1-score and recall among the fine-tuned models; however, it exhibits significantly lower precision and produces longer answers compared to PPO+SFT. For both, SFT reduces the average answer length to that of GPT-3.5, while resulting in a loss of recall. Interestingly, training observations indicate that the shorter answer lengths after SFT, enhance PPO training, improving reward gains and reducing training times. This efficiency is likely due to faster answer generation and fewer token generation steps for reward distribution.

The results clearly indicate that fine-tuning is effective in improving citation performance for RAG. We find that fine-tuning improves the F1-score by .10 to .20 points or a relative reduction of F1 error of 28 - 43 %.

5 Ethical Considerations

All experiments performed in this work were conducted in accordance with the ACM Code of Ethics. We believe that there should be no conflicts and that this work does not raise any ethical issues. All datasets used are publicly available or synthetically generated. Both cases are referenced accordingly. We do not use personal data or other sensitive information.

6 Limitations

The major limitation of our work is that RAGE considers only citation quality for evaluation. More aspects have to be covered to provide a complete RAG evaluation framework. At the moment, we refer to other work to include aspects like measures for factual correctness, how good information from the documents is integrated and a general measure of how fluent the answer is. This especially becomes relevant when evaluating the fine-tuned model versions as the improvement in citation quality does not necessarily come with an improvement in the other metrics or could even worsen performance in some cases. Tests of our models do not indicate this, but it is still important to consider when applying them in practice. Also currently, there are only two evaluation datasets converted to the format used in RAGE. A greater variety of datasets would further improve the significance of the evaluation.

References

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Dongfang Li, Zetian Sun, Baotian Hu, Zhenyu Liu, Xinshuo Hu, Xuebo Liu, and Min Zhang. 2024. Improving attributed text generation of large language models via preference learning. *arXiv preprint arXiv:2403.18381*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT-4. *arXiv preprint arXiv:2306.02707*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. 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.

A Prompt Structure

Figure 1 shows an example for a prompt generated during evaluation using RAGE. Our prompt structure is strongly inspired by Gao et al. (2023). The figure also shows the original query and the response containing citations, that was generated by a Mistral 7B (Jiang et al., 2023) model. A prompt of this structure is generated for each query and contains a predefined portion of relevant, irrelevant and seemingly-relevant documents. The response of the LLM under evaluation is analysed regarding its citation quality as described in Section 3.2.

B Effects of Varying the Relevancy Mixture

Figure 2 shows the effects of using different mixtures of relevant, irrelevant and seemingly-relevant documents for a given query on citation precision and recall. The number of relevant documents was kept constant, and the proportion of irrelevant and seemingly relevant documents was varied. The evaluation was performed with five different open-weight models. Evaluation results are consistent across all evaluated models and show that RAGE is robust to using different proportions. The evaluation also shows that seemingly relevant documents significantly increase task difficulty.

Query: When did hollywood become the centre of the film industry?

Prompt:
Instruction: Write an accurate, engaging, and concise answer for the given question using only the provided documents (some of which might be irrelevant) and cite them properly using the format $[< doc\_id >]$. For example if a particular information is from document 3 reference it with [3]. Use an unbiased and journalistic tone. Always cite for any factual claim. When citing several search results, use [1][2][3]. Cite at least one document and at most three documents in each sentence. If multiple documents support the sentence, only cite a minimum sufficient subset of the documents.

Documents:
[1]: Hollywood was a small community in 1870 and was incorporated as a municipality in 1903. It was consolidated with the city of Los Angeles in 1910, and soon thereafter a prominent film industry emerged, eventually becoming the most recognizable film industry in the world.

[2]: By 1912, major motion-picture companies had set up production near or in Los Angeles. In the early 1900s, most motion picture patents were held by Thomas Edison's Motion Picture Patents Company in New Jersey, and filmmakers were often sued to stop their productions. To escape this, filmmakers began moving out west, where Edison's patents could not be enforced. Also, the weather was ideal and there was quick access to various settings. Los Angeles became the capital of the film industry.

[3]: The history of the Forest Service [. . .]

[. . .]

[6]: The cinema of the United States, often metonymously referred to as Hollywood [. . .] Since the 1920s, the film industry of the United States has had higher annual grosses than any other country's [. . .]

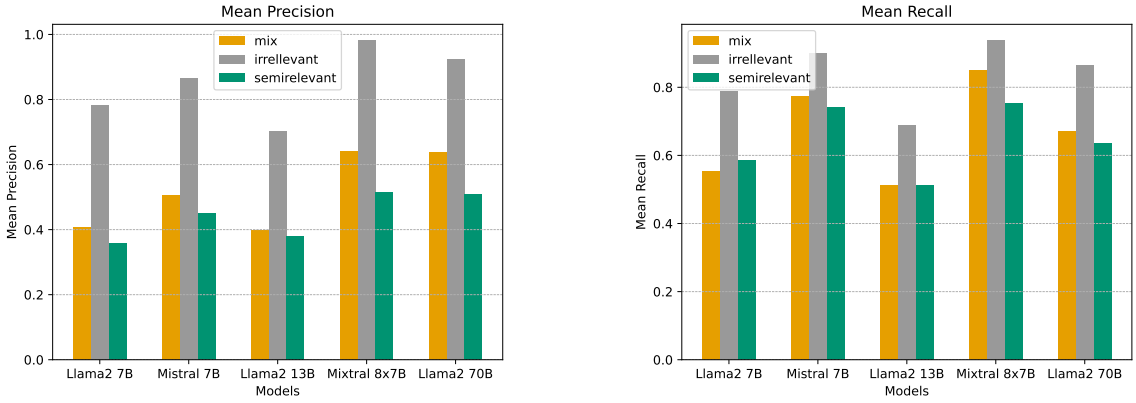
[. . .]

Question: When did hollywood become the centre of the film industry?

Remember to answer as short as possible.

Answer: Hollywood became the center of the film industry in the early 1900s as major motion-picture companies moved there to escape patent disputes and take advantage of ideal weather and easy access to various settings [1][2]. By the 1920s, Hollywood produced the largest number of films and had the highest annual grosses in the film industry [6].

Figure 1: An example prompt and LLM response produced during the evaluation process. The query was taken from the Natural Questions (Kwiatkowski et al., 2019) dataset contained in the BEIR benchmark (Thakur et al., 2021). The prompt was automatically generated and contains the documents assembled from relevant, irrelevant and apparently relevant documents as described in Section 3.2. The structure of the prompt template essentially follows the concept presented in (Gao et al., 2023). The answer was generated by a Mistral 7B model (Jiang et al., 2023).



(a) (b)

Figure 2: Comparative evaluation of mean citation precision (Figure 2a) and recall (Figure 2b) across three document relevancy mixtures in the Natural Questions (Kwiatkowski et al., 2019) dataset. The *mix* setup includes 1-4 relevant, 3 irrelevant, and 3 seemingly relevant documents. The *irrelevant* setup consists of 1-4 relevant and 6 irrelevant documents, with no seemingly relevant documents. The *seemingly-relevant* setup features 1-4 relevant and 6 seemingly relevant documents, excluding any irrelevant documents.