LEAF: Learning and Evaluation Augmented by Fact-Checking to Improve Factualness in Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated exceptional capabilities in natural language processing tasks but often fall short in maintaining factual accuracy, particularly in knowledge-intensive domains like healthcare. This study introduces LEAF: Learning and Evaluation Augmented by Fact-Checking, a novel framework aimed at improving the factual reliability of LLMs in medical question answering (QA). LEAF comprises three key contributions: (1) the **Retrieval-Augmented Factuality Evaluator (RAFE)**, a robust fact-checking system using open-source LLMs and domain-specific retrieval corpora to evaluate response accuracy; (2) **Fact-Check-then-RAG**, an enhanced Retrieval-Augmented Generation method that incorporates fact-checking to guide retrieval without requiring parameter updates; and (3) **Learning from Fact Check via Self-Training**, a strategy to improve LLM performance through supervised fine-tuning or preference-based learning, using fact-checking results as pseudo-labels. Experimental results show that RAFE outperforms Factcheck-GPT in detecting inaccuracies, Fact-Check-then-RAG effectively corrects errors, and Learning from Fact Check improves performance without labeled data. These findings suggest LEAF as a scalable and robust solution for low-resource settings [1].

1 Introduction

Large language models (LLMs) have revolutionized natural language processing (NLP), bringing remarkable advancements to tasks such as question answering (QA). As a cornerstone task in NLP, QA involves generating accurate and contextually appropriate answers to questions posed in natural language. Their ability to comprehend complex prompts and generate human-like responses has significantly enhanced the utility of QA systems in practical applications like knowledge retrieval, decision support, and education (Cai et al., 2023; Liu et al., 2023; Jin et al., 2024).

Medical QA highlights the significant demands and challenges faced by LLMs in QA tasks, particularly in ensuring factual accuracy and integrating relevant domain-specific knowledge. Accurate answers in medical QA often rely on retrieval-augmented generation (RAG) techniques, where models augment their responses by retrieving authoritative and up-to-date information from external knowledge sources (Singhal et al., 2023; Wu et al., 2024). For example, determining the appropriate treatment for a patient may require accessing the latest clinical guidelines, retrieving evidence-based medical literature, or verifying specific diagnostic criteria. This dependence on knowledge retrieval underscores the critical importance of integrating reliable and domain-specific retrieval systems to address factual inaccuracies and ensure contextually relevant responses. Without robust retrieval mechanisms and rigorous factuality checks, LLMs risk generating plausible but incorrect information, which is particularly problematic in high-stakes domains like healthcare.

To address this challenge, fact-checking has become a promising solution. Fact-checking mechanisms work by verifying the factual accuracy of generated content against reliable data sources, serving as a filter to detect and correct misinformation (Quelle and Bovet, 2024; Wang et al., 2024b). Prior research has explored various methods for integrating fact-checking into LLM workflows, including verification techniques such as Factcheck-GPT, Factscore and SAFE (Wang et al., 2023; Wei et al., 2024; Min et al., 2023). However, these approaches have notable limitations. For instance, frameworks like Factcheck-GPT and SAFE rely on proprietary model such as ChatGPT-3.5, which cannot be deployed on private datasets, restricting

[1] We will release prompts, codes, and dataset upon acceptance.

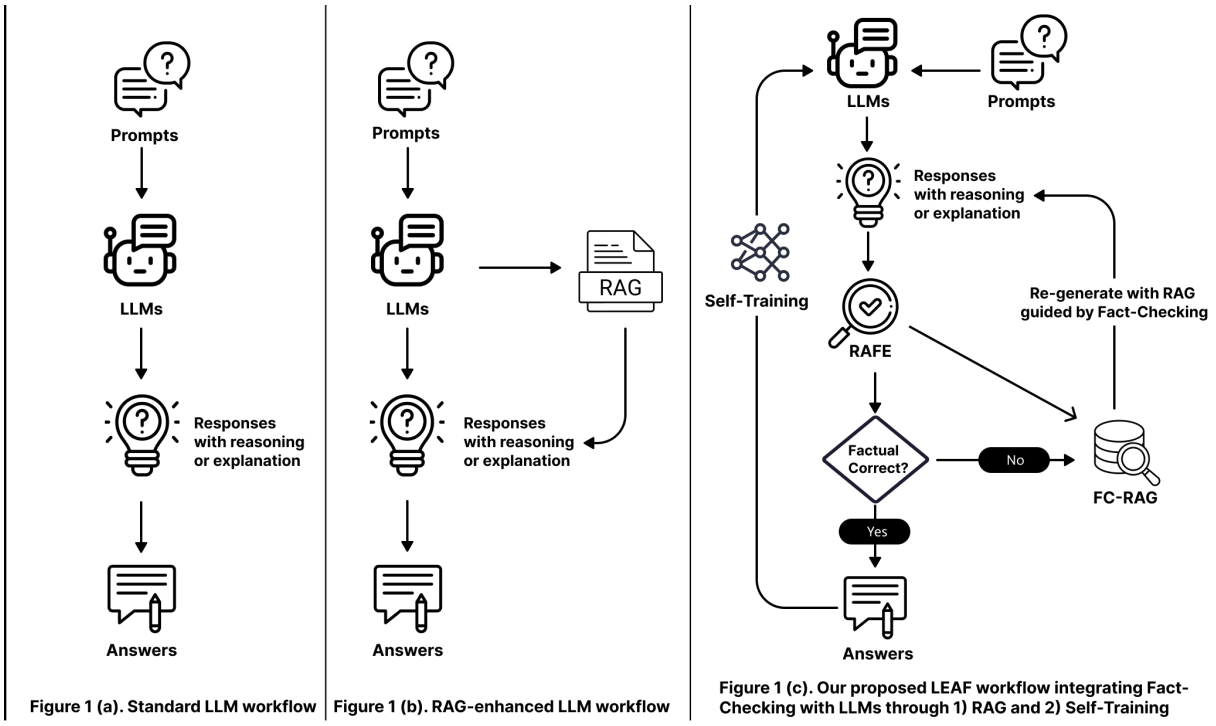

Figure 1: Comparison of workflows: standard LLM workflow (left), RAG-enhanced LLM workflow (middle), and our proposed Fact-Checking integrated workflow (right).

their applicability in sensitive fields like healthcare. Additionally, reliance on Google Search for retrieving information exposes these frameworks to vulnerabilities, including inconsistent results and the potential inclusion of malicious content from the open web, further compromising reliability.

In this study, we introduce **LEAF: Learning and Evaluation Augmented by Fact-Checking**, a novel framework designed to enhance the factual accuracy and reliability of LLMs. LEAF introduces three complementary contributions to address the challenges of improving factual accuracy across diverse use cases:

- **Retrieval-Augmented Factuality Evaluator (RAFE)**: We propose RAFE, a robust fact-checking system that combines open-source LLM(Qwen2-72B-Instruct) with a corpus-based retrieval system tailored to general-domain knowledge and medical-domain resources. This systematic evaluation enhances the reliability of responses while ensuring domain specificity and accessibility, surpassing limitations of prior approaches like Factcheck-GPT.

- **Fact-Check-then-RAG**: We propose an innovative approach to Retrieval-Augmented Generation where retrieval is informed by fact-checking results. This method selectively retrieves information to address factual inaccuracies in the model's initial outputs, significantly improving contextual relevance and factual correctness without updating the underlying LLM parameters. This approach is particularly beneficial for proprietary models like ChatGPT that cannot be fine-tuned.

- **Learning from Fact-Check via Self-Training**: We explore two self-training mechanisms leveraging fact-checked responses to improve LLM parameters: **Supervised Fine-Tuning (SFT)**: Fact-checked responses are used to fine-tune the model, reinforcing factual correctness. This involves generating responses, evaluating them with fact-checking, and fine-tuning on verified outputs. **Preference-Based Learning with SimPO (Meng et al., 2024)**: Fact-checking is used as a ranking mechanism, with high-scoring responses labeled as "chosen" and low-scoring ones as "rejected" to guide preference-based training. This method further refines the model's ability to prioritize factual responses.

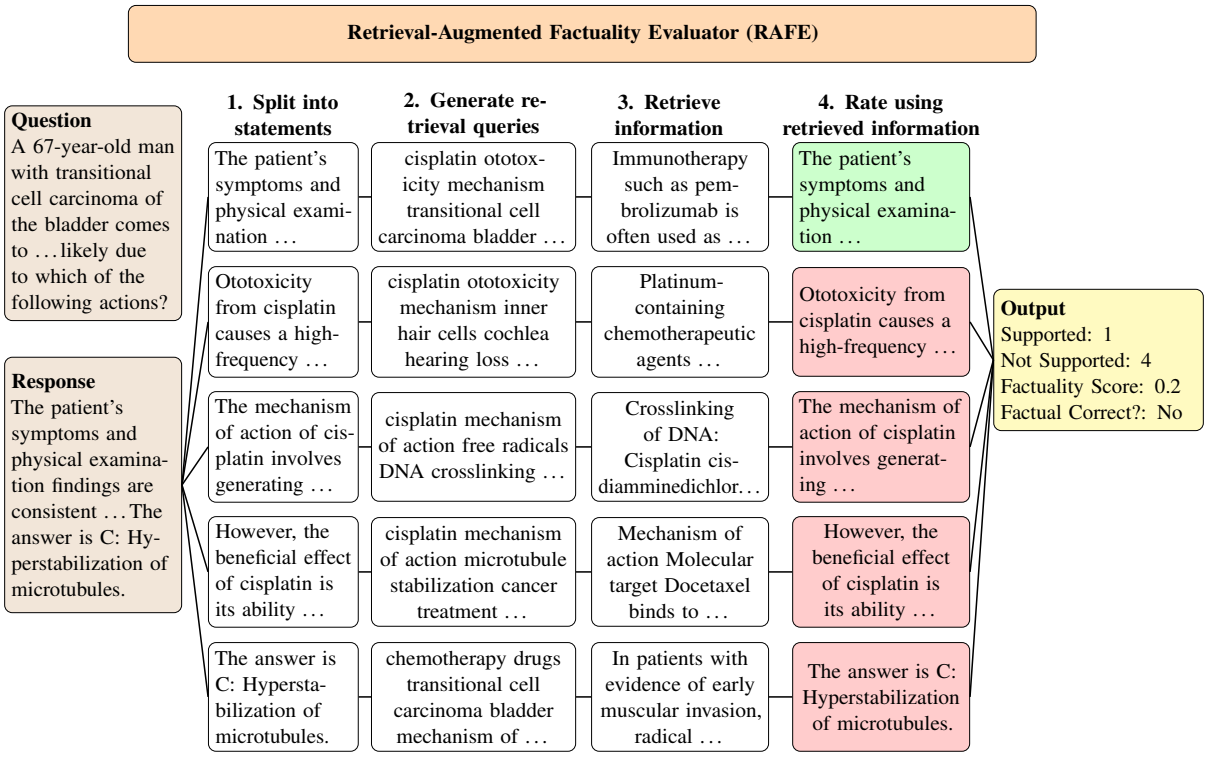

Figure 2: The Retrieval-Augmented Factuality Evaluator (RAFE) assesses the factual accuracy of response in four steps. (1) Split into sentences: The response is divided into individual statements. (2) Generate retrieval queries: For each statement, an LLM generates multiple retrieval queries aimed at retrieving relevant information. (3) Retrieve information: The retrieval system gathers supporting information based on these queries. (4) Rate using retrieved information: Each statement is evaluated against the retrieved information and labeled as Supported or Not Supported. The final output includes a factuality score, calculated as the proportion of supported statements, which aids in selecting the most factually reliable response.

2 Methodology

In this section, we describe our proposed methodology to enhance the factual accuracy and reliability of large language models in generating responses. Our approach, LEAF, integrates fact-checking, retrieval-augmented generation, and self-training mechanisms to systematically improve factuality in LLM outputs. The workflow of our proposed method is illustrated in Figure 1.

The proposed workflow aims to enhance the factual accuracy of responses generated by large language models through an integrated fact-checking process. In the conventional LLM workflow (Figure 1(a)), the model generates responses to prompts with reasoning or explanations and directly provides the final answers. However, this approach does not inherently guarantee the factual correctness of the output. In standard Retrieval-Augmented Generation (Figure 1(b)), the prompts are supplemented with relevant documents retrieved using the question as a retrieval query. This process can add valuable contextual information but sometimes introduces noise due to the inclusion of irrelevant documents, which can adversely affect the accuracy of the generated response.

To address these limitations, we introduce Fact-Check-then-RAG, a fact-checking-integrated workflow (Figure 1(c)). In this enhanced approach, after the LLM generates a response, it is evaluated by a fact-checking system. If the response is factually correct, it is retained as the final output. However, if the response is identified as incorrect, the workflow triggers a FC-RAG process, incorporating relevant documents retrieved during fact-checking into the prompt. This refined prompt guides the model to regenerate a more accurate response. This iterative process continues until a factually correct answer is achieved. In addition, factually verified responses are used for self-training. The model is fine-tuned on these fact-checked outputs, further improving its performance and reliability in generating factual responses. The following subsections provide a detailed breakdown of each component of our methodology.

2.1 Retrieval-Augmented Factuality Evaluator (RAFE)

Inspired by fact-checking systems that combine LLMs with external search engines, such as Factcheck-GPT, Factscore, and Search-Augmented Factuality Evaluator (SAFE) (Wang et al., 2023; Min et al., 2023; Wei et al., 2024), which use an LLM (ChatGPT-3.5) integrated with Google Search to evaluate the factuality of responses, we introduce the **Retrieval-Augmented Factuality Evaluator (RAFE)**. RAFE adapts these approaches by replacing the closed source ChatGPT-3.5 model with the open source Qwen2-72B-Instruct model and substituting Google Search with a corpus-based retrieval system that includes both general-domain knowledge (Wikipedia) and medical domain resources (PubMed, StatPearls, and Medical Textbooks). This adaptation enhances domain specificity and accessibility for tasks that require specialized knowledge.

To assess the factual accuracy of generated responses, RAFE evaluates each response in four systematic steps, as illustrated in Figure 2:

1. **Split into Statements**: Each response is divided into individual statements.

2. **Generate Retrieval Queries**: For each statement, RAFE employs an LLM to generate multiple retrieval queries designed to retrieve contextually relevant evidence.

3. **Retrieve Information**: The retrieval system gathers documents that corresponds to each generated query. This evidence provides a factual basis for evaluating the consistency of each statement with external sources.

4. **Rate Using Retrieved Information**: Each statement is compared against the retrieved evidence and labeled as either **Supported** or **Not Supported**, based on alignment with the information. The overall factuality score for the response is calculated as the proportion of supported statements, indicating the response's factual reliability. A response is considered factually correct if all statements are supported by retrieved knowledge.

2.2 Fact-Check-then-RAG

Our second innovative mechanism, Fact-Check-then-RAG, seamlessly integrates the fact-checking stage with Retrieval-Augmented Generation (RAG). This approach leverages the documents retrieved during the fact-checking process to enhance the generation of responses. The key idea is to utilize the knowledge retrieved from the fact-checking stage, specifically for individual statements that did not pass the fact-check test. This strategy ensures that when a statement is not supported by the retrieved knowledge sources, the relevant documents are included in the RAG prompt to help the LLM refine its reasoning or answer, potentially improving performance. As illustrated in Figure 3, the methodology involves several steps:

First, during the fact-checking stage, each statement in a response is evaluated for factual correctness using RAFE. If a statement is not supported by the knowledge retrieved, it indicates a gap between the LLM and the knowledge base. For these unsupported statements, relevant documents are retrieved from a comprehensive corpus (MedCORP (Xiong et al., 2024)), which includes authoritative sources like Wikipedia, PubMed, textbooks, and StatPearls. The ColBERT (Khattab and Zaharia, 2020) retrieval model is used to extract these documents.

Next, the retrieved documents are included in the RAG prompt. This additional context provides the LLM with the necessary information to adjust its reasoning or answer, addressing the knowledge gap identified during the fact-checking stage. The LLM then generates new responses using the RAG framework, which is now enhanced with the relevant knowledge retrieved earlier.

By integrating fact-checking with RAG, our approach effectively addresses the knowledge gaps identified during the fact-checking process. This method enhances the LLM's ability to produce accurate and reliable responses, demonstrating improved performance over traditional RAG methods, particularly in increasing the factualness of generated content.

2.3 Learning from Fact-Check via Self-Training

We explore self-training mechanisms using fact-checked responses to enhance the performance of LLMs. This approach consists of two main parts: supervised fine-tuning on factually correct responses and preference-based learning with Simple Preference Optimization.

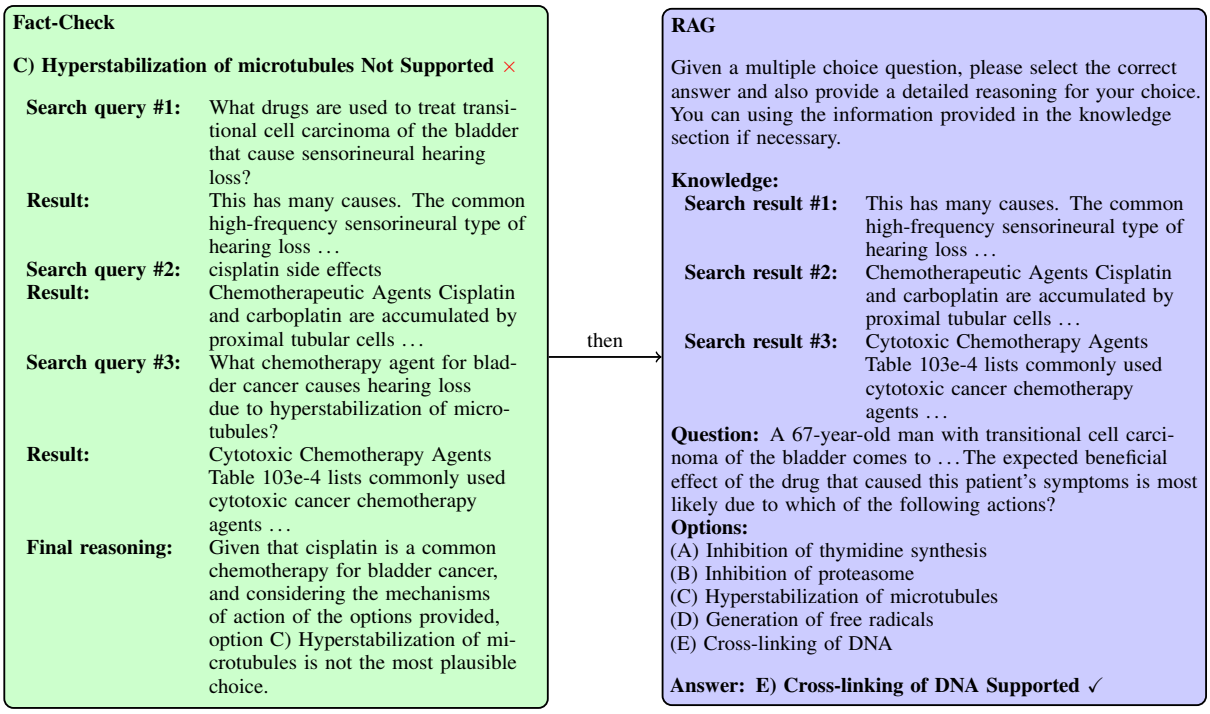

Figure 3: Fact-Check-then-RAG is able to change the answer of LLMs by leveraging the knowledge retrieved from fact-check stage to regenerate the responses.

2.3.1 Supervised Fine-Tuning on Factually Correct Responses

This phase involves fine-tuning the model on responses that have passed fact-checking, ensuring training on verified, accurate information and enhancing overall model performance. The LLM generates multiple responses to a given prompt, which are evaluated by the fact-checking system. Only responses with a factuality score of 1 are selected for fine-tuning. The model is then fine-tuned on these factually correct responses, reinforcing its ability to produce accurate and reliable outputs.

2.3.2 Preference-based Learning with SimPO

The second part of our self-training approach utilizes Simple Preference Optimization (Meng et al., 2024), SimPO aligns the reward formulation directly with the generation metric, eliminating the need for a reference model. This process involves Fact-Checking as a Ranking Model: The fact-checking system assigns scores to generated responses based on their factual accuracy. The highest-scoring responses are selected as "chosen" and the lowest-scoring ones as "rejected". By using the fact-checking system as a ranking model, SimPO effectively guides the model to prefer factually accurate responses.

3 Experiments

In this section, we present experiments to evaluate each component of our proposed workflow. Due to constraints in time and computational resources, we were able to run only a single iteration, meaning each component was executed once without repeating until a factually correct answer was reached. We anticipate that increasing the number of iterations would yield improved results, albeit with greater time and computational costs.

We conducted two main experimental setups across different model configurations. For the large LLaMA 3 70B Instruct model, we applied the Fact-Check-then-RAG technique to enhance the model's performance without updating its parameters, as fine-tuning such a large model is computationally intensive. In contrast, with the smaller LLaMA 3 8B Instruct model, we explored self-training techniques where the model parameters were updated based on fact-checking outcomes rather than labeled data. The self-training was conducted using either supervised fine-tuning or preference-based learning, with training data curated through a rigorous fact-checking process.

3.1 Alignment Between Factuality and Correctness

We assess the alignment between factuality and correctness in responses generated by LLaMA 3 70B across five datasets—MedQA, MMLU Medical, PubMedQA, BioASQ, and MedMCQA—using precision, recall, and F1 scores to compare the performance of Factcheck-GPT and RAFE. These metrics measure the effectiveness of the fact-checking systems in ensuring that responses are both factually and actually correct.

Precision represents the proportion of responses labeled as factually correct by the fact-checking system that are actually correct. A higher precision indicates fewer false positives, meaning the system is effective in validating responses that truly align with the ground truth. Recall, on the other hand, measures the proportion of factually correct responses identified by the fact-checking system out of all actually correct responses. This metric reflects the system's sensitivity in capturing all correct answers, including those that might be challenging to validate. The F1 score provides a balance between precision and recall, offering a single metric to assess the system's overall performance in aligning factuality with correctness.

As shown in Table 1, RAFE consistently surpasses Factcheck-GPT across all datasets. RAFE achieves significantly higher precision, indicating its superior ability to accurately validate factually correct responses while minimizing false positives. For instance, RAFE achieves a precision of 96.27% on BioASQ compared to Factcheck-GPT's 85.29%, demonstrating its robustness in distinguishing correct responses. Similarly, RAFE's recall outperforms Factcheck-GPT on every dataset, reflecting its stronger capability to capture a larger proportion of correct responses. On MMLU Medical, RAFE achieves a recall of 58.79%, compared to 44.77% for Factcheck-GPT, highlighting its ability to identify more correct answers. The F1 scores consistently show RAFE's dominance, combining high precision and recall to deliver more aligned and reliable results across datasets.

Overall, RAFE demonstrates a superior alignment between factuality and correctness, outperforming Factcheck-GPT by significant margins across all datasets. These results underscore RAFE's effectiveness in fact-checking, ensuring that validated responses are not only factually accurate but also aligned with the actual ground truth. This makes RAFE a robust and scalable solution for enhancing factual reliability in knowledge-intensive domains like medical QA.

Dataset	Factcheck-GPT			RAFE		
	Precision	Recall	F1	Precision	Recall	F1
MedQA	77.35	29.91	43.14	**86.52**	**75.43**	**80.59**
MMLU-M	84.00	44.77	58.41	**93.00**	**58.79**	**72.04**
PubMedQA	50.93	63.37	56.47	**72.76**	**69.64**	**71.16**
BioASQ	85.29	29.12	43.41	**96.27**	**51.81**	**67.36**
MedMCQA	75.31	28.92	41.79	**81.62**	**43.90**	**57.09**

Table 1: Precision, recall, and F1 scores for Factcheck-GPT and RAFE across five medical QA datasets, MMLU-M mean MMLU-Medical. Bold values indicate higher scores.

3.2 Fact-Check-then-RAG

To evaluate the effectiveness of our Fact-Check-then-RAG (FC-RAG) approach, we present the experiments conducted comparing it to the original performance of the LLaMA 3 70B Instruct model and the standard RAG setting in MedRAG (Xiong et al., 2024). In MedRAG, the question is used as a query to retrieve relevant documents, which are then included in the prompt. In our FC-RAG approach, we use information obtained in the fact-checking stage to include in the prompt.

Table 2 compares the performance of the Llama 3 70B Instruct model across five medical QA datasets. While RAG is designed to improve the model's contextual grounding by providing additional information, the results reveal that it actually harms performance on the MedQA and MMLU-Medical datasets—consistent with findings in original paper (Xiong et al., 2024). This suggests that while RAG can be beneficial in certain contexts, it may introduce noise or irrelevant information in others, leading to decreased accuracy. In contrast, the FC-RAG approach consistently improves accuracy across all datasets. By incorporating fact-checking results into the RAG process, FC-RAG ensures that the model's outputs are more reliable and factually correct. This method leverages verified information during generation, leading to significant performance gains: a 4.99% improvement on MedQA, 1.66% on MMLU-Medical, 13.0% on PubMedQA, 7.28% on BioASQ, and 1.56% on MedMCQA compared to the original model performance. These results demonstrate the robustness and efficiency of FC-RAG in enhancing the outputs of large language models, particularly in domains where factual accuracy is critical.

Dataset	MedQA	MMLU-M	PubMedQA	BioASQ	MedMCQA	Average
CoT	73.53	85.12	60.60	80.58	71.21	74.21
RAG	68.58	82.46	70.80	87.70	68.78	75.66
FC-RAG	**77.52**	**86.78**	**73.60**	**87.86**	**72.77**	**79.71**

Table 2: Comparison of LLaMA 3 70B Instruct CoT, performance when using RAG, and FC-RAG on five medical QA datasets. Note that all of model's parameters remained unchanged. MMLU-M mean MMLU Medical

3.3 Computational Analysis

Table 3 presents the inference costs on MedQA for two method: RAG and FC-RAG. The table reports the average number of model calls and the total number of tokens generated per question during the generation process. Traditional RAG required 1 call and average of generated 467.0 tokens, while for the FC-RAG, the number of calls is 3, with 1050.8 tokens generated on average.

	RAG	FC-RAG
Avg. calls	1.0	3.0
Avg. generated tokens	467.0	1050.8

Table 3: Inference costs on MedQA with RAG and FC-RAG. We show the average number of inferences and generated tokens required to answer a question.

3.4 Supervised Fine-Tuning on Factually Correct Responses

In order to assess the effectiveness of a model fine-tuned on fact-checked generated responses, the LLaMA 3 8B Instruct model was tested on prompts drawn from five datasets, generating responses that were subsequently fact-checked. We perform supervised fine-tuning on the responses that pass the fact-check test(the response with factuality score is 1.0). We compare the performance of the SFT model with the original model and also conduct the same experiments on the Factcheck-GPT (Wang et al., 2023).

Table 4 shows that SFT with fact-checked responses significantly improves accuracy across all datasets. Using RAFE, the model achieved notable gains, including a 4.71% increase on MedQA and a 6.60% increase on PubMedQA, compared to the original model. Additionally, RAFE outperformed Factcheck-GPT, demonstrating its robustness and efficiency in ensuring factually accurate outputs. These results highlight the potential of combining fact-checking with fine-tuning to enhance LLM performance.

3.5 Preference-based Learning on Ranked Responses

We design experiments to evaluate the effectiveness of preference-based learning with SimPO on responses ranked by our fact-checking system and by ArmoRM (Wang et al., 2024a). For each question, we generate five responses using the Llama 3 8B Instruct model with a temperature setting of 0.8. We then use our fact-checking system and ArmoRM to score these responses, selecting the lowest-scored responses as "rejected" and the highest-scored responses as "chosen". We then run preference-based learning on these chosen and rejected responses.

As shown in Table 4, the preference-based learning with SimPO on RAFE-ranked responses results in better performance compared to ArmoRM-ranked responses. Specifically, the SimPO approach using RAFE shows significant improvements: an increase of 4.08% on MedQA, 2.67% on MMLU-Medical, 6.80% on PubMedQA, 7.45% on BioASQ, and 2.89% on MedMCQA compared to the original model performance. This is attributed to the larger gap between the highest and lowest-scored responses in our fact-checking system, as demonstrated in Table 5. A larger gap indicates a more significant distinction between high-quality and low-quality responses, leading to more effective learning and ultimately better performance after preference-based learning.

4 Related Work

Evaluating factuality in Model Responses Evaluating the factuality of model responses is crucial for ensuring the reliability of large language models. Recent studies have demonstrated that LLMs can serve as effective tools for fact verification (Guan et al., 2024; Tian et al., 2023). Improvements in human evaluation techniques have further enhanced factuality assessment (Cheng et al., 2024). Factcheck-GPT (Wang et al., 2023) presents an end-to-end solution for annotating factuality in LLM outputs, offering fine-grained labels for verifiability and factual inconsistencies. Inspired by methods that break down responses for evaluation (Chern et al., 2023), SAFE (Wei et al., 2024) applies a similar approach in the long-form factuality setting, leveraging search-augmented models. While methods like Factcheck-GPT and SAFE offer innovative approaches to factuality evaluation, they face no-

Method	MedQA	MMLU-Medical	PubMedQA	BioASQ	MedMCQA	Average
Original	55.46	70.98	55.20	74.27	57.78	62.74
Supervised Fine-Tuning (SFT)						
SFT(Factcheck-GPT)	57.03	71.99	59.60	75.40	58.71	64.55
SFT(RAFE)	**60.17**	**75.85**	**61.80**	**78.80**	**60.75**	**67.47**
Preference-based Learning (SimPO)						
SimPO(ArmoRM)	56.40	72.82	59.00	76.70	59.05	64.79
SimPO(RAFE)	**59.54**	**73.65**	**62.00**	**81.72**	**60.67**	**67.52**

Table 4: Comparison of performance on five medical QA datasets using Supervised Fine-Tuning (SFT) and Preference-Based Learning approaches with Llama 3 8B Instruct.

table limitations, such as reliance on proprietary models and Google Search, making them unsuitable for private or sensitive domains like healthcare. Our approach overcomes these issues using Qwen2-72B-Instruct for factuality evaluation and MedCorp with ColBERT for secure, domain-specific retrieval.

Retrieval-Augmented Generation Retrieval-Augmented Generation, proposed by (Yih, 2020), integrates relevant retrieved information into the generation process of LLMs, enhancing their performance on knowledge-intensive tasks. This approach helps improve factualness by grounding the LLMs on provided contexts and supplying up-to-date knowledge that might not be encoded in the models. Many studies have built upon the original RAG framework to further improve its effectiveness, including works by (Borgeaud et al., 2022; Ram et al., 2023; Gao et al., 2023; Jiang et al., 2023). In the biomedical field, RAG has been explored for literature information-seeking and clinical decision-making (Frisoni et al., 2022; Naik et al., 2022; Jin et al., 2023).

Learning from Fact-Check via Self-Training Building on self-training methods like Med-Gemini (Saab et al., 2024), which integrate web search to enhance clinical reasoning, we propose a fact-checking-based approach tailored to the medical domain. Unlike existing methods that often rely on external web searches or curated labels, our approach generates multiple responses, evaluates their factuality using domain-specific retrieval systems, and fine-tunes the model on validated outputs. This ensures greater reliability and domain adaptation. Our method also addresses limitations in prior work, such as SCoRe (Kumar et al., 2024), which focuses on general self-correction, and rationale-based self-improvement (Huang et al., 2022), by explicitly incorporating medical context and robust factuality checks to reduce hallucinations and improve clinical relevance.

5 Conclusion

In this study, we explored the potential of fact-checking mechanisms to enhance the factual accuracy of large language models in medical question-answering tasks.

Firstly, we demonstrated that the **Retrieval-Augmented Factuality Evaluator** can effectively replace closed-source LLMs integrated with Google Search by utilizing open-source LLMs and a specialized corpus retrieval system. This architecture offers a more controllable, cost-effective, and domain-adaptable solution, reducing reliance on external APIs while enabling precise tuning for specific datasets and domains.

Additionally, we proposed **Fact-Check-then-RAG**, an innovative approach that integrates fact-checking into Retrieval-Augmented Generation workflows. This method improves the correctness of generated responses without requiring updates to model parameters.

Finally, we introduced **two methods for learning from fact-checking results**, providing a novel framework to enhance LLM performance without the need for labeled data. These methods leverage fact-checking outputs as pseudo-labels, enabling supervised fine-tuning on factually correct responses and preference-based learning to refine model outputs. This flexibility demonstrates the robustness of fact-checking mechanisms in model training, particularly in low-resource scenarios.

Overall, our findings highlight the versatility and scalability of fact-checking systems like LEAF in improving LLM accuracy, offering practical solutions for knowledge-intensive domains such as medical QA, even in resource-constrained settings.

6 Limitations

Despite the promising results, our study has several limitations that need to be addressed in future work. One significant limitation is the speed and computational efficiency of the fact-checking system. The current implementation requires multiple iterations of inference with LLMs and several retrieval operations for each sentence in the responses. This process can be time-consuming and computationally intensive, potentially limiting the scalability and real-time applicability of our approach.

Additionally, our study primarily focused on the medical domain, leveraging datasets and corpora specific to healthcare. While this domain specificity ensured relevance and precision, it also limits the generalizability of our findings to other fields. Extending our approach to diverse domains and evaluating its effectiveness across various types of knowledge-intensive tasks will be crucial for broader applicability.

Our future works will also explore RAFE's performance upper bounds by leveraging more comprehensive medical corpora and investigating the impact of multiple rounds of self-training. Additionally, we plan to integrate stronger fact-checking models, such as Meta's LLaMA 405B, to enhance the precision of our fact-verification process and extend RAFE's applicability to other knowledge-intensive domains beyond healthcare.

References

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

Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, et al. 2023. Paniniqa: Enhancing patient education through interactive question answering. *Transactions of the Association for Computational Linguistics*, 11:1518–1536.

Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2024. Relic: Investigating large language model responses using self-consistency. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai-a tool augmented framework for multi-task and multi-domain scenarios. corr, abs/2307.13528, 2023. doi: 10.48550. *arXiv preprint arXiv.2307.13528*.

Giacomo Frisoni, Miki Mizutani, Gianluca Moro, and Lorenzo Valgimigli. 2022. Bioreader: a retrieval-enhanced text-to-text transformer for biomedical literature. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 5770–5793.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv e-prints*, pages arXiv–2312.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. 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 1090–1111, Mexico City, Mexico. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Qiao Jin, Robert Leaman, and Zhiyong Lu. 2023. Retrieve, summarize, and verify: how will chatgpt affect information seeking from the medical literature? *Journal of the American Society of Nephrology*, 34(8):1302–1304.

Qiao Jin, Zifeng Wang, Charalampos S Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. 2024. Matching patients to clinical trials with large language models. *Nature Communications*, 15(1):9074.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. 2024. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*.

Siru Liu, Aileen P Wright, Barron L Patterson, Jonathan P Wanderer, Robert W Turer, Scott D Nelson, Allison B McCoy, Dean F Sittig, and Adam Wright. 2023. Using ai-generated suggestions from chatgpt to optimize clinical decision support. *Journal of the American Medical Informatics Association*, 30(7):1237–1245.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv e-prints*, pages arXiv–2405.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. 2022. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Dorian Quelle and Alexandre Bovet. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7:1341697.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv e-prints*, pages arXiv–2404.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024a. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, et al. 2023. Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output. *arXiv e-prints*, pages arXiv–2311.

Yuxia Wang, Minghan Wang, Hasan Iqbal, Georgi Georgiev, Jiahui Geng, and Preslav Nakov. 2024b. Openfactcheck: A unified framework for factuality evaluation of llms. *arXiv preprint arXiv:2405.05583*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, et al. 2024. Long-form factuality in large language models. *arXiv e-prints*, pages arXiv–2403.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv e-prints*, pages arXiv–2402.

Scott Yih. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Conference on Neural Information Processing Systems, Vancouver, Canada*.

A Appendix

A.1 Overview

This appendix provides supplementary information and detailed examples to support the methodology and results presented in the main paper. It is structured as follows:

- **Datasets:** A comprehensive description of the five medical datasets used in our experiments, including MedQA, MMLU-Medical, PubMedQA, BioASQ, and MedMCQA.
- **Factuality Confusion Matrixes**
- **Fact-Checking as a Ranking Model**
- **Self-Training Experimental Setup:** Detailed information about the infrastructure, hyperparameters, and training procedures used in our experiments.
- **Prompts:** Examples of prompts used for query generation, fact-checking, and retrieval-augmented generation, demonstrating how our system interacts with the language models.
- **Fact-Checking Process:** A step-by-step walkthrough of our fact-checking methodology, including:
  1. Query generation with context
  2. Retrieval from the MedCorp corpus
  3. Fact-checking with context
- **Fact-Check-Then-RAG process:** A walkthrough of how to use the fact-checking results to guide the RAG process.
- **Impact of Fact-Checking and Sample Questions:** An analysis of how fact-checking influences the selection of correct options, illustrated with examples and visualizations. This section includes a set of sample questions from the MedQA dataset to demonstrate the system's performance and allow for experiment reproduction.

Each section builds upon the previous ones, providing a comprehensive view of our methodology and its application. The examples and figures throughout the appendix are designed to illustrate key concepts and provide empirical support for our approach.

A.2 Datasets

In this subsection, we describe the datasets used in our experiments. We utilize the MIRAGE benchmark (Xiong et al., 2024), which comprises five medical QA datasets, including three medical examination QA datasets and two biomedical research QA datasets. Specifically, the datasets are as follows:

MMLU-Med (Hendrycks et al., 2020): This dataset includes multiple-choice questions from medical examinations, testing the model's knowledge and reasoning in various medical domains.

MedQA (Jin et al., 2021): This dataset contains multiple-choice questions from the US medical licensing examination, designed to evaluate the model's understanding of medical concepts and clinical practices.

MedMCQA (Pal et al., 2022): This dataset features multiple-choice questions from Indian medical examinations, providing a diverse set of questions that test the model's knowledge in clinical medicine and medical science.

PubMedQA* (Jin et al., 2019): Following the setting in the MIRAGE paper, we use a modified version of PubMedQA where all ground-truth supporting contexts are excluded, resulting in PubMedQA*. This dataset focuses on yes/no questions derived from biomedical research articles, testing the model's ability to answer questions based solely on the questions without additional context.

BioASQ-Y/N (Tsatsaronis et al., 2015): This dataset contains yes/no questions from the BioASQ challenge, which aims to test the model's ability to understand and answer questions based on biomedical literature.

We adhere to the same settings as the MIRAGE paper, including only multiple-choice questions related to biomedicine and excluding all ground-truth supporting contexts for the questions. For example, in PubMedQA, we remove the contexts and only use the questions, resulting in PubMedQA*. It is important to note that while we focus on medical QA tasks in this work, our workflow of integrating LLMs with fact-checking is generalizable to any domain and can be applied to various tasks beyond QA. We chose the QA task for its popularity in evaluating LLMs and demonstrating the effectiveness of our proposed workflow.

Dataset	MedQA	MMLU-Medical	PubMedQA	BioASQ	MedMCQA	Average
Lowest ArmoRM score	51.92	68.69	58.40	74.60	57.54	62.23
Highest ArmoRM score	56.80	73.19	60.20	78.32	59.91	65.68
Δ(ArmoRM)	4.88	4.50	1.80	3.72	2.37	3.45
Lowest RAFE score	48.78	68.69	53.20	73.79	55.99	60.09
Highest RAFE score	60.33	73.55	64.60	79.94	61.42	67.97
Δ(RAFE)	**11.55**	**4.86**	**11.40**	**6.15**	**5.43**	**7.88**

Table 5: Comparison of lowest and highest scored responses using ArmoRM and RAFE across five medical QA datasets on LLaMA 3 8B Instruct. Δ represents the difference between the highest and lowest performance for each system.

A.3 Factuality Confusion Matrixes

We evaluate the alignment between factual correctness and actual correctness of LLaMA 3 70B responses across five datasets—MedQA, MMLU Medical, PubMedQA, BioASQ, and MedMCQA—using Factcheck-GPT and RAFE. The alignment ratio, defined as the proportion of True Positives (TP) and True Negatives (TN) to total samples, quantifies the effectiveness of each fact-checking system.

MedQA: As shown in Table 6, Factcheck-GPT achieves an alignment ratio of 0.42, while RAFE improves this to 0.73, a 31% increase. RAFE significantly reduces misaligned predictions (false positives and false negatives).

MMLU Medical: In Table 7, RAFE improves the alignment ratio from 0.46 (Factcheck-GPT) to 0.61, a 15% gain, by increasing true positives (545 vs. 414) and true negatives (121 vs. 82).

PubMedQA: Table 8 shows RAFE improving the alignment ratio from 0.41 to 0.66 (+25%). RAFE achieves this by increasing alignment in both factual and actual correctness.

BioASQ: On BioASQ (Table 9), RAFE achieves an alignment ratio of 0.60 compared to 0.39 for Factcheck-GPT (+21%), with significant improvements in true positives (258 vs. 145).

MedMCQA: As seen in Table 10, RAFE achieves an alignment ratio of 0.53, compared to 0.43 for Factcheck-GPT (+10%). Despite the dataset's size, RAFE consistently improves aligned predictions.

Summary: RAFE consistently outperforms Factcheck-GPT across all datasets, with alignment ratio gains ranging from 10% to 31%. These results highlight RAFE's effectiveness in enhancing factual and actual correctness alignment.

Method	Type	Actual Correct	Actual Incorrect	Alignment
Factcheck-GPT	Factual Correct	280	82	0.42
	Factual Incorrect	656	255	
RAFE	Factual Correct	706	110	**0.73**
	Factual Incorrect	230	227	

Table 6: Confusion matrix for the MedQA dataset, comparing Factcheck-GPT and RAFE.

Method	Type	Actual Correct	Actual Incorrect	Alignment
Factcheck-GPT	Factual Correct	414	80	0.46
	Factual Incorrect	513	82	
RAFE	Factual Correct	545	41	**0.61**
	Factual Incorrect	382	121	

Table 7: Confusion matrix for the MMLU Medical dataset, comparing Factcheck-GPT and RAFE.

Method	Type	Actual Correct	Actual Incorrect	Alignment
Factcheck-GPT	Factual Correct	192	185	0.41
	Factual Incorrect	111	12	
RAFE	Factual Correct	211	79	**0.66**
	Factual Incorrect	92	118	

Table 8: Confusion matrix for the PubmedQA dataset, comparing Factcheck-GPT and RAFE.

Method	Type	Actual Correct	Actual Incorrect	Alignment
Factcheck-GPT	Factual Correct	145	25	0.39
	Factual Incorrect	353	95	
RAFE	Factual Correct	258	10	**0.60**
	Factual Incorrect	240	110	

Table 9: Confusion matrix for the BioASQ dataset, comparing Factcheck-GPT and RAFE.

Method	Type	Actual Correct	Actual Incorrect	Alignment
Factcheck-GPT	Factual Correct	863	283	0.43
	Factual Incorrect	2121	916	
RAFE	Factual Correct	1310	295	**0.53**
	Factual Incorrect	1674	904	

Table 10: Confusion matrix for the MedMCQA dataset, comparing Factcheck-GPT and RAFE.

A.4 Fact-Checking as a Ranking Model

We conducted an experiment to assess the effectiveness of our fact-checking system as a ranking model for responses generated by large language models. Five responses were generated using the LLaMA 3 8B Instruct model with a temperature setting of 0.8. Each response was then scored by our

fact-checking system, and the performance of the highest and lowest-scored responses was analyzed. For comparison, we also ran similar experiments using ArmoRM. (Wang et al., 2024a), a reward model designed to align LLMs with human preferences. ArmoRM is trained using human preference data, employing a Mixture-of-Experts (MoE) strategy to select suitable reward objectives based on context.

LLaMA 3 8B (Lowest ArmoRM score): Performance of the lowest scored response using the ArmoRM reward model.

LLaMA 3 8B (Highest ArmoRM score): Performance of the highest scored response using the ArmoRM reward model.

Δ(ArmoRM): This indicates the difference in performance between the highest and lowest-scored responses using ArmoRM.

LLaMA 3 8B (Lowest RAFE score): Performance of the lowest scored response using RAFE.

LLaMA 3 8B (Highest RAFE score): Performance of the highest scored response using RAFE.

Δ(RAFE): This indicates the difference in performance between the highest and lowest-scored responses using our fact-checking system.

As evident from table 5, our fact-checking system(RAFE) effectively ranks the responses to highlight the best-performing ones. The larger Δ values for our system compared to ArmoRM demonstrate the robustness and efficiency of our fact-checking approach in differentiating between high-quality and low-quality responses.

A.5 Self-Training Experimental Setup

Optimization with SimPO The second part of our self-training approach utilizes Simple Preference Optimization (Meng et al., 2024) to rank and optimize responses based on their factual accuracy. SimPO aligns the reward formulation directly with the generation metric, eliminating the need for a reference model. This process involves Fact-Check as Ranking Model:

- Fact-Check as Ranking Model: The fact-checking system assigns scores to generated responses based on their factual accuracy. The highest-scoring responses are selected as "chosen" and the lowest-scoring as "rejected".

- SimPO Objective: The SimPO objective is designed to maximize the difference in rewards between the chosen and rejected responses. The reward is calculated as:

$$r_{SimPO}(x, y) = \frac{\beta}{|y|} \sum_{i=1}^{|y|} \log \pi_\theta(y_i | x, y_{<i}) \tag{1}$$

where β is a scaling constant.

- Target Reward Margin: Additionally, we introduce a target reward margin term, $\gamma > 0$, to the Bradley-Terry objective to ensure that the reward for the winning response, $r(x, y_w)$, exceeds the reward for the losing response, $r(x, y_l)$, by at least γ:

$$p(y_w \succ y_l | x) = \sigma(r(x, y_w) - r(x, y_l) - \gamma). \tag{2}$$

Finally, we obtain the SimPO objective by incorporating the length-normalized reward:

$$L_{SimPO}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim D} \left[\log \sigma \left(\frac{\beta}{|y_w|} \log \pi_\theta(y_w|x) - \frac{\beta}{|y_l|} \log \pi_\theta(y_l|x) - \gamma \right) \right]. \tag{3}$$

A.5.1 Hyperparameters for Training

The training of the LLaMA 3 8B Instruct model was carefully configured using a set of hyperparameters designed to optimize the model's performance on the selected tasks. The key hyperparameters and their settings are summarized in Table 11.

The learning rate was set to 1.0×10^{-6}, a value selected after initial experimentation to balance the rate of convergence with the stability of training. A batch size of 4 per device was chosen to ensure that the model could effectively utilize the available GPU memory, while the gradient accumulation steps were set to 8 to allow for a larger effective batch size without exceeding memory limits.

The maximum sequence length was set to 2048 tokens, with a prompt length of 1800 tokens, ensuring that the model could process lengthy inputs and generate comprehensive responses. The AdamW optimizer was selected for its effectiveness in handling weight decay during training, and the cosine learning rate scheduler was used to gradually reduce the learning rate, facilitating smoother convergence.

The warmup ratio of 0.1 was implemented to gently ramp up the learning rate at the beginning of

training, reducing the risk of instability in the early stages. The number of training epochs was set to 5, balancing training time with the need for thorough model training.

Specific to SimPO, the beta and gamma hyperparameters were set to 2.5 and 1.4, respectively. These values were selected based on prior research and experimentation, optimizing the model's preference ordering during training. Finally, a seed of 42 was used to ensure reproducibility of the results.

Hyperparameter	Value
Learning Rate	1.0e-6
Batch Size per Device	4
Gradient Accumulation Steps	8
Max Sequence Length	2048
Max Prompt Length	1800
Optimizer	AdamW
LR Scheduler Type	Cosine
Warmup Ratio	0.1
Number of Training Epochs	5
Beta (SimPO)	2.5
Gamma (SimPO)	1.4
Seed	42

Table 11: Summary of Hyperparameters for Training with SimPO.

A.5.2 Infrastructure

All experiments presented in this paper were conducted using a computing environment equipped with four NVIDIA H100 80GB GPUs. These GPUs are built on the Hopper architecture and feature HBM3 memory, providing exceptional performance for large-scale AI and machine learning tasks.

This high-performance hardware configuration enabled efficient handling of the computationally intensive tasks required for training and evaluating large language models across multiple medical datasets.

A.5.3 Self-Training Experiments

In this set of experiments, we focused on evaluating the impact of self-training using the Llama 3 8B Instruct model across five medical datasets. The process began by generating five responses for each prompt, with each prompt corresponding to a question in the selected medical datasets: MedQA, MMLU-Medical, PubMedQA, BioASQ, and MedMCQA.

After generating the responses, we applied two different approaches for each dataset:

- Supervised Fine-Tuning on Fact-Checked Responses: In this approach, we fine-tuned the model using only the responses that passed a rigorous fact-checking process. This ensured that the model learned from the most accurate data available.
- Simple Preference Optimization with Fact-Check Ranking: Here, we utilized fact-check scores to rank the generated responses. The highest-ranked responses were used for further optimization of the model via SimPO, refining the model's output quality based on factual correctness.

Each of these self-training methods—SFT and SimPO—was performed separately on each dataset to assess their individual impact on the model's performance. After the training process, we evaluated the accuracy and reliability of the fine-tuned models across the same medical QA datasets, allowing us to determine the effectiveness of each self-training approach.

It is important to note that all fine-tuning in this experiment was conducted as full fine-tuning without the use of any LoRA (Low-Rank Adaptation) techniques.

A.6 Prompts

In this section, we provide an overview of the various prompts used in our experiments (Table 12). These prompts were designed to guide the LLM through different stages of processing, including query generation, fact-checking, and retrieval-augmented generation. Each prompt is tailored to specific tasks, ensuring the model receives clear instructions to perform the required actions effectively.

- **{_KNOWLEDGE_PLACEHOLDER}**: This represents the background information or facts that are provided to the model. It typically includes retrieved documents, or previously established facts that can help the model in its reasoning process.
- **{_CONTEXT_PLACEHOLDER}**: This contains the specific scenario or question that the model needs to address. In medical QA tasks, this often includes patient information, symptoms, and other relevant details of the

case. For example, in MedQA, this part is dynamically filled with a question and the corresponding answer options.

- **{_STATEMENT_PLACEHOLDER}**: This represents a specific claim or assertion that the model needs to evaluate or fact-check based on the given knowledge and context. In our medical QA experiments, this placeholder is filled with individual sentences from the LLM's initial response to a question. Each sentence is fact-checked separately to assess the factual accuracy of the entire response at a granular level.

- **{_QUESTION_PLACEHOLDER}**: In the Fact-Check-Then-RAG prompt, this represents the full question text that the model needs to answer.

- **{_OPTIONS_PLACEHOLDER}**: In the Fact-Check-Then-RAG prompt, this contains the list of multiple-choice options that the model can choose from when answering the question.

These placeholders are dynamically filled with appropriate content during the execution of our system, allowing for flexible and context-specific interactions with the language model.

A.7 Fact-Checking Process

To evaluate the effectiveness of our fact-checking system, we conducted experiments using the Llama 3 70B Instruct model on several samples of the MedQA dataset. For each question, ten responses were generated with a temperature setting of 1.2. These responses were subsequently evaluated using our fact-checking system. The figure 8 displays the frequency of each answer option along with the average fact-check score assigned to those options. Notably, the fact-check scores tend to be higher for the correct answers, which are highlighted in gold. This visualization illustrates the correlation between the frequency of selected options and their factual accuracy, as determined by the fact-checking system. The results demonstrate that the fact-checking system can reliably identify and score correct responses, supporting its utility in enhancing the factual accuracy of model outputs.

We present an example from the MedQA dataset to illustrate the fact-checking process. The example involves a 13-year-old boy presenting with severe knee, hip, and groin pain. The prompt for the model was:

> **An example of MedQA Question** *A 13-year-old boy presents to the emergency department with severe knee, hip, and groin pain. The patient has a past medical history notable only for obesity and asthma. His temperature is 98°F (36.7°C), blood pressure is 124/65 mmHg, pulse is 128/min, respirations are 14/min, and oxygen saturation is 99% on room air. Physical exam is notable for an inability of the patient to bear weight on his left leg and limited range of motion of the left hip. Which of the following is the best management for this patient?*

The available choices were:

- (A) Casting and crutches
- (B) Immobilization of the hip in a Pavlik harness
- (C) Supportive therapy and observation
- (D) Surgical drainage of the hip
- **(E) Surgical pinning of the femoral head** (Correct)

For this prompt, we generated 5 responses using the Llama 3 70B Instruct model with a temperature of 1.2. The responses were then fact-checked, with each sentence in the response being evaluated for factual accuracy against retrieved knowledge. The fact-check score for each response was calculated as the ratio of sentences supported by the retrieved knowledge to the total number of sentences in the response.

Table 13 illustrates the LLM original generated responses, and their selected options, corresponding fact-check scores. In the markup text, sentences that were **not supported** by the retrieved knowledge are highlighted in red, while sentences that were supported remain in black.

We will take the first response in Table 13 as an example to show how to do fact-check with context.

> **An Example of LLM original response** ***(D) Surgical drainage of the hip** **Reasoning:** This patient's symptoms and physical exam findings are highly suggestive of a septic hip, also known*

Type	Prompt
Query generation with context	Instructions: 1. You have been given a STATEMENT, a CONTEXT and some KNOWLEDGE points. 2. Your goal is to try to find evidence that either supports or does not support the factual accuracy of the given STATEMENT in the given CONTEXT. 3. To do this, you are allowed to issue ONE Google Search query that you think will allow you to find additional useful evidence. 4. Your query should aim to obtain new information that does not appear in the KNOWLEDGE. This new information should be useful for determining the factual accuracy of the given STATEMENT. 5. Format your final query by putting it in a markdown code block. KNOWLEDGE: {_KNOWLEDGE_PLACEHOLDER} CONTEXT: {_CONTEXT_PLACEHOLDER} STATEMENT: {_STATEMENT_PLACEHOLDER}
Fact-check with context	Instructions: 1. You have been given a STATEMENT, a CONTEXT and some KNOWLEDGE points. 2. Determine whether the given STATEMENT is supported by the given CONTEXT, you can use the given KNOWLEDGE to support your decision if necessary. The STATEMENT is supported if it is a proper action or reasoning given the CONTEXT. 3. Before showing your answer, think step−by−step and show your specific reasoning. 4. If the STATEMENT is supported by the CONTEXT, be sure to show the supporting evidence. 5. After stating your reasoning, restate the STATEMENT and then determine your final answer based on your reasoning and the STATEMENT. 6. Your final answer should be either "{SUPPORTED_LABEL}" or "{NOT_SUPPORTED_LABEL}". Wrap your final answer in square brackets. KNOWLEDGE: {_KNOWLEDGE_PLACEHOLDER} CONTEXT: {_CONTEXT_PLACEHOLDER} STATEMENT: {_STATEMENT_PLACEHOLDER}
Fact-Check-Then-RAG	Given a multiple choice question, please select the correct answer and also provide a detailed reasoning for your choice. You can using the information provided in the knowledge section if necessary. KNOWLEDGE: {_KNOWLEDGE_PLACEHOLDER} QUESTION: {_QUESTION_PLACEHOLDER} OPTIONS: {_OPTIONS_PLACEHOLDER} ANSWER:

Table 12: All prompts used in our work.

*as infectious arthritis of the hip. The patient's severe pain, limited range of motion, and inability to bear weight on the left leg are all consistent with an inflammatory response in the hip joint. The fact that the patient is afebrile (normal temperature) does not rule out septic arthritis, as some patients may not develop a fever. The past medical history of obesity and asthma are not directly related to the patient's current symptoms. The patient's vital signs indicate tachycardia (elevated heart rate), which may be due to pain or inflammation. Given the high likelihood of septic arthritis, the best management for this patient is surgical drainage of the hip (option D). Delaying drainage can lead to permanent damage to the joint and surrounding bone. Surgical drainage will allow for removal of infected fluid and tissue, as well as collection of a sample for microbiological analysis to guide antibiotic therapy. The other options are not appropriate in this scenario: * (A) Casting and crutches are not indicated for a potentially infected joint. * (B) Pavlik harness is a device used to treat developmental dysplasia of the hip in infants, not for septic arthritis in adolescents. * (C) Supportive therapy and observation would not adequately address the potential infection and may lead to further harm. * (E) Surgical pinning of the femoral head is not a treatment for septic arthritis; it may be used for fractures or other orthopedic conditions, but not for an infected joint. Therefore, the best management for this patient is surgical drainage of the hip (option D).*

A.7.1 Query Generation with Context

First, we use the "Query generation with context" prompt (as shown in Table 12) to generate a search query. For this example, the CONTEXT is the full question text, and the STATEMENT is a response generated by the LLM. A sample STATEMENT from the above response could be:

> *Given the high likelihood of septic arthritis, the best management for this patient is surgical drainage of the hip (option D).*

The LLM then generates a search query based on this information:

```
13−year−old boy knee hip groin pain unable to bear weight
    ↪ best management
```

The detail is shown in Figure 4.

A.7.2 Retrieval from MedCorp Corpus

Next, we use this query to retrieve 3 relevant documents from the MedCorp corpus using ColBERT retrieval. The retrieved documents provide additional context for fact-checking. This process is as shown in Figure 5.

A.7.3 Fact-Check with Context

Finally, we use the "Fact-check with context" prompt (Table 12) to evaluate the factual accuracy of the STATEMENT. The KNOWLEDGE placeholder is filled with the retrieved documents, and the CONTEXT and STATEMENT remain the same as in the query generation step. The filled prompt is shown in Figure 6.

The LLM then evaluates the statement and provides a reasoning:

> *The STATEMENT is not supported by the CONTEXT. The CONTEXT does not provide any information that would support a diagnosis of septic arthritis or the need for surgical drainage of the hip. The patient's symptoms are more consistent with conditions such as slipped capital femoral epiphysis (SCFE), which is common in obese adolescents and typically requires surgical pinning of the femoral head.*
>
> *Final answer: [Not Supported]*

This process is repeated for each sentence in the LLM's response, allowing for a granular assessment of the response's factual accuracy. The overall fact-check score for the response is then calculated as the ratio of supported sentences to the total number of sentences.

The results demonstrate that our fact-checking system can reliably identify and score correct responses, supporting its utility in enhancing the factual accuracy of model outputs.

A.8 Fact-Check-Then-RAG

After the initial fact-checking process, if the LLM's response is found to contain inaccuracies, we employ the Fact-Check-Then-RAG approach to improve the response. This method leverages the

Selected Option	Fact-Check Score	Response
D	0.75	**(D) Surgical drainage of the hip** **Reasoning:** This patient's symptoms and physical exam findings are highly suggestive of a septic hip, also known as infectious arthritis of the hip. The patient's severe pain, limited range of motion, and inability to bear weight on the left leg are all consistent with an inflammatory response in the hip joint. The fact that the patient is afebrile (normal temperature) does not rule out septic arthritis, as some patients may not develop a fever. The past medical history of obesity and asthma are not directly related to the patient's current symptoms. The patient's vital signs indicate tachycardia (elevated heart rate), which may be due to pain or inflammation. Given the high likelihood of septic arthritis, the best management for this patient is surgical drainage of the hip (option D). Delaying drainage can lead to permanent damage to the joint and surrounding bone. Surgical drainage will allow for removal of infected fluid and tissue, as well as collection of a sample for microbiological analysis to guide antibiotic therapy. The other options are not appropriate in this scenario: * (A) Casting and crutches are not indicated for a potentially infected joint. * (B) Pavlik harness is a device used to treat developmental dysplasia of the hip in infants, not for septic arthritis in adolescents. * (C) Supportive therapy and observation would not adequately address the potential infection and may lead to further harm. * (E) Surgical pinning of the femoral head is not a treatment for septic arthritis; it may be used for fractures or other orthopedic conditions, but not for an infected joint. Therefore, the best management for this patient is surgical drainage of the hip (option D).
B	0.5	**Answer**: (B) Immobilization of the hip in a Pavlik harness **Reasoning**: Based on the symptoms and physical exam findings, the most likely diagnosis is a slipped capital femoral epiphysis (SCFE). SCFE is a common condition affecting adolescents, particularly obese individuals, where the epiphysis (growth plate) at the proximal end of the femur slips, causing hip pain and limited mobility. The patient's age, symptoms, and physical exam findings are consistent with SCFE. The inability to bear weight on the left leg and limited range of motion of the left hip suggest a hip-related issue. SCFE is more common in obese individuals, and the patient's history of obesity supports this diagnosis. Immobilization of the hip in a Pavlik harness (option B) is the best management for this patient. The Pavlik harness is a device that maintains the hip in a position that reduces the risk of further slippage and allows for gentle reduction of the epiphysis. This approach is often used for SCFE, especially in cases where the slippage is mild to moderate. The other options are not appropriate for this patient: * Casting and crutches (option A) would not be effective in reducing the risk of further slippage or promoting healing. * Supportive therapy and observation (option C) would not provide adequate stability or treatment for the condition. * Surgical drainage of the hip (option D) is not indicated, as SCFE is not typically associated with an infectious process. * Surgical pinning of the femoral head (option E) may be necessary in some cases of SCFE, but it is not the initial management of choice. Surgery is often reserved for more severe cases or when orthopedic management fails. In conclusion, immobilization of the hip in a Pavlik harness is the best management for this patient, given the likely diagnosis of SCFE and the need to reduce the risk of further slippage and promote healing.
E	1.0	(E) Surgical pinning of the femoral head **Reasoning**: The patient's symptoms, particularly the sudden onset of severe knee, hip, and groin pain, along with the inability to bear weight and limited range of motion of the left hip, suggest a diagnosis of slipped capital femoral epiphysis (SCFE). SCFE is a common condition in adolescents, particularly in those who are obese, which is consistent with the patient's past medical history. SCFE occurs when the femoral head slips out of place at the growth plate, leading to a mismatch between the head and neck of the femur. This can cause significant pain, stiffness, and impaired mobility in the affected hip and surrounding areas. Now, let's evaluate the options: (A) Casting and crutches are typically used for fractures or sprains, which are not likely diagnoses based on the patient's symptoms. (B) Immobilization in a Pavlik harness is usually used for developmental dysplasia of the hip (DDH) in infants, not for SCFE in adolescents. (C) Supportive therapy and observation may be appropriate for minor injuries or self-limiting conditions, but SCFE requires more prompt and definitive treatment to prevent long-term complications, such as osteoarthritis and hip deformities. (D) Surgical drainage of the hip is not typically indicated for SCFE, as there is no clear indication of infection or abscess formation. (E) Surgical pinning of the femoral head is the most appropriate management for this patient. This procedure involves stabilizing the femoral head with one or more pins to prevent further slippage and promote healing. Early surgical intervention can help prevent long-term complications and improve outcomes for patients with SCFE. Therefore, the correct answer is (E) Surgical pinning of the femoral head.
D	0.64	**Answer**: D) Surgical drainage of the hip **Reasoning**: Based on the patient's symptoms and physical exam findings, the most likely diagnosis is a septic hip, which is a medical emergency. The patient's severe pain, inability to bear weight, and limited range of motion of the left hip suggest an inflammatory process in the hip joint. Although the patient's temperature is normal, the patient's pulse is elevated, which could indicate an underlying infection. Given the patient's symptoms and physical exam findings, surgical drainage of the hip is the best management option. Septic arthritis of the hip joint requires urgent surgical intervention to drain the infected joint fluid, wash out the joint, and administer antibiotics to prevent long-term damage to the joint. Option A, casting and crutches, is not appropriate for a suspected septic hip, as it would not address the underlying infection. Option B, immobilization of the hip in a Pavlik harness, is used to treat developmental dysplasia of the hip in infants, not septic hip in adolescents. Option C, supportive therapy and observation, would not be sufficient to manage a septic hip, as it would not address the underlying infection and could lead to long-term joint damage. Option E, surgical pinning of the femoral head, is not indicated in this scenario, as it is used to treat femoral neck fractures or slipped capital femoral epiphyses, not septic hip. In conclusion, surgical drainage of the hip is the best management option for this patient to ensure prompt treatment of the suspected septic hip and prevent long-term joint damage.

Table 13: Generated responses, selected options, fact-check scores, and response generated for the example question.

An Example Prompt for Query Generation with Context

Instructions:
1. You have been given a STATEMENT, a CONTEXT and some KNOWLEDGE points.
2. Your goal is to try to find evidence that either supports or does not support the factual accuracy of the given STATEMENT in the given CONTEXT.
3. To do this, you are allowed to issue ONE Google Search query that you think will allow you to find additional useful evidence.
4. Your query should aim to obtain new information that does not appear in the KNOWLEDGE. This new information should be useful for determining the factual accuracy of the given STATEMENT.
5. Format your final query by putting it in a markdown code block.
KNOWLEDGE: {_KNOWLEDGE_PLACEHOLDER}
CONTEXT: A 13-year-old boy presents to the emergency department with severe knee, hip, and groin pain. The patient has a past medical history notable only for obesity and asthma. His temperature is 98°F (36.7°C), blood pressure is 124/65 mmHg, pulse is 128/min, respirations 14/min, and oxygen saturation is 99% on room air. Physical exam is notable for an inability of the patient to bear weight on his left leg and limited range of motion of the left hip. Which of the following is the best management for this patient. The available choices were: (A) Casting and crutches (B) Immobilization of the hip in a Pavlik harness (C) Supportive therapy and observation (D) Surgical drainage of the hip (E) Surgical pinning of the femoral head.
STATEMENT: Given the high likelihood of septic arthritis, the best management for this patient is surgical drainage of the hip (option D).

LLM Response:

```
13-year-old boy knee hip groin pain unable to bear weight best
management
```

Figure 4: An Example Prompt for Query Generation with Context

An Example Query to MedCorp Corpus

13-year-old boy knee hip groin pain unable to bear weight best management

---

Retrieved Docs from MedCorp Corpus with ColBERT Retrieval:

in extreme cases and in older children.Slipped Capital Femoral EpiphysisChildren ages 10 to 16 years can develop displacement of the epiphysis on the femoral neck with no history of injury. The slippage occurs through the weak zone (hypertrophic zone) of the growth plate. When slippage occurs in young patients, check for endocrine disorders such as hypothyroid-ism, renal osteodystrophy, and growth hormone deficiency. Slipped capital femoral epiphysis (SCFE) is associated with African-American heritage and obesity, and it is more com-mon in boys than in girls. One-quarter of cases are bilateral. In patients with endocrine etiology, the condition is usually bilateral. Patients generally present with groin and anterior thigh pain, and the patient may have antalgic gait and a limp.
The mean age at onset is 6 years, with a range of 3 to 8 years. It is twice as common in male children. The patient or family will describe an acute onset of pain in the groin/hip, anterior thigh, or knee. Irritation of the obturator nerve can cause referred pain in the thigh and knee when the pathology is at the hip. Patients with transient synovitis are often afebrile, walk with a painful limp, and have normal to minimally elevated white blood cell count, C-reactive protein, and erythrocyte sedimentation rate compared with bacterial diseases of the hip (Table 199-1). Table 197-3 lists the differential diagnosis of a limping child. Anteroposterior and frog-leg radiographs of the hip are usually normal. Ultrasonography may reveal a joint effusion.
and pelvic osteoto-mies, are done in older age groups and in more severe cases. Osteonecrosis of the femoral head is a possible complication of treatment and can result in pain and decreased range of motion.Legg-Calvé-Perthes DiseaseOsteonecrosis of the proximal femoral epiphysis can cause flattening of the femoral head called Legg-Calvé Perthes disease. The age at presentation is between 4 and 8 years of age and occurs more in males, usually affecting one side. Younger age at presentation (less than 6 years old) will have a better prognosis. The patient presents with groin or knee pain, decreased hip motion, and a limp.

Figure 5: An example query to MedCorp Corpus and 3 retrieved documents

knowledge retrieved during the fact-checking stage to generate a more accurate answer.

Using our example question about the 13-year-old boy, let's walk through the Fact-Check-Then-RAG process:

First, we use the "Fact-Check-then-RAG" prompt (as shown in Table 12). The KNOWLEDGE placeholder is filled with the relevant information retrieved during the fact-checking process. For our example, this might include:

> *Slipped capital femoral epiphysis (SCFE) is associated with African-American heritage and obesity, and it is more common in boys than in girls. Patients generally present with groin and anterior thigh pain, and the patient may have antalgic gait and a limp. Patient may present with knee pain that can lead to missing the diagnosis. In pediatric patients with knee pain, the ipsilateral hip should be assessed as well.*

The QUESTION placeholder contains the original question text, and the OPTIONS placeholder lists the available choices. The prompt for the LLM would then look like Figure 7.

The LLM then generates a new response based on this prompt. It excludes the option D based on the knowledge retrieved from previous fact-checking, and reaches the correct answer:

> *(D) Surgical drainage of the hip is not typically indicated for SCFE, as there is no clear indication of infection or abscess formation.*
> ...
> *Therefore, the correct answer is (E) Surgical pinning of the femoral head.*

This Fact-Check-Then-RAG process allows the LLM to generate a more accurate and well-reasoned response by incorporating the relevant medical knowledge retrieved during the fact-checking stage. The resulting answer is not only correct but also provides a detailed explanation grounded in factual information.

A.9 Impact of Fact-Checking and Sample Questions

To demonstrate the effectiveness of our fact-checking system, we conducted experiments using the Llama 3 70B Instruct model on multiple samples from the MedQA dataset. Figure 8 illustrates the results of these experiments, showing the frequency of selected answer options and their corresponding fact-check scores.

For each of the six sample questions, we generated ten responses using a temperature setting of 1.2. Our fact-checking system then evaluated these responses, assigning scores to each option. The results reveal several key insights:

Correlation with Correct Answers: Across all samples, the correct answers (highlighted in gold) consistently received higher fact-check scores. This strong correlation demonstrates the ability of our fact-checking system to identify factually accurate responses.

Handling of Ambiguity: In some cases, such as sample 4, multiple options received relatively high fact-check scores. This suggests that our system can capture nuanced differences in factual accuracy, even when multiple options may have some degree of correctness.

Consistency Across Samples: The pattern of higher fact-check scores for correct answers is consistent across all six samples, indicating the robustness of our approach across different types of medical questions.

Potential for Improving Model Performance: The clear distinction in fact-check scores between correct and incorrect answers suggests that our system could be effectively used to enhance the model's decision-making process, potentially improving its overall performance on medical QA tasks.

To provide context for these results, we present the six sample questions from the MedQA dataset used in this analysis, shown in Figure 9 and Figure 10.

These sample questions cover a range of medical scenarios and concepts, demonstrating the versatility of our fact-checking system across different types of medical knowledge and reasoning tasks.

An Example Prompt for Fact-Check with Context

Instructions:
1. You have been given a STATEMENT, a CONTEXT and some KNOWLEDGE points.
2. Determine whether the given STATEMENT is supported by the given CONTEXT, you can use the given KNOWLEDGE to support your decision if necessary. The STATEMENT is supported if it is a proper action or reasoning given the CONTEXT.
3. Before showing your answer, think step-by-step and show your specific reasoning.
4. If the STATEMENT is supported by the CONTEXT, be sure to show the supporting evidence.
5. After stating your reasoning, restate the STATEMENT and then determine your final answer based on your reasoning and the STATEMENT.
6. Your final answer should be either "{SUPPORTED_LABEL}" or "{NOT_SUPPORTED_LABEL}". Wrap your final answer in square brackets.
KNOWLEDGE:
(1). in extreme cases and in older children . . .
(2). The mean age at onset is 6 years, with a range of 3 to 8 years . . .
(3). and pelvic osteoto-mies, are done in older age groups and in more severe cases . . .
(4) . . .

CONTEXT: A 13-year-old boy presents to the emergency department with severe knee, hip, and groin pain . . .
STATEMENT: Given the high likelihood of septic arthritis, the best management for this patient is surgical drainage of the hip (option D).

LLM Response:

```
The given STATEMENT is "Given the high likelihood of septic
arthritis, the best management for this patient is Surgical
drainage of the hip."
To determine if this STATEMENT is supported by the CONTEXT, I
will analyze the information provided in the CONTEXT and compare
it with the STATEMENT.
The CONTEXT describes a 13-year-old boy ...
Final answer:  [Not Supported]
```

Figure 6: An example prompt for Fact-Check with context. The final answer to the statement is [Not Supported].

An Example Prompt for Fact-Check-Then-RAG

Given a multiple choice question, please select the correct answer and also provide a detailed reasoning for your choice. You can using the information provided in the knowledge section if necessary.

KNOWLEDGE:
(1). in extreme cases and in older children ...
(2). The mean age at onset is 6 years, with a range of 3 to 8 years...
(3). and pelvic osteoto-mies, are done in older age groups and in more severe cases...
(4). ...

QUESTION: A 13-year-old boy presents to the emergency department with severe knee, hip, and groin pain. The patient has a past medical history notable only for obesity and asthma. His temperature is 98°F (36.7°C), blood pressure is 124/65 mmHg, pulse is 128/min, respirations 14/min, and oxygen saturation is 99% on room air. Physical exam is notable for an inability of the patient to bear weight on his left leg and limited range of motion of the left hip. Which of the following is the best management for this patient.

OPTIONS: (A) Casting and crutches (B) Immobilization of the hip in a Pavlik harness (C) Supportive therapy and observation (D) Surgical drainage of the hip (E) Surgical pinning of the femoral head.

ANSWER:

LLM Response:

```
(E) Surgical pinning of the femoral head
**Reasoning**:
The patient's symptoms, particularly the sudden onset of severe
knee, hip, and groin pain, along with the inability to bear
weight and limited range of motion of the left hip, suggest a
diagnosis of slipped capital femoral epiphysis (SCFE)...
Therefore, the correct answer is (E) Surgical pinning of the
femoral head.
```

Figure 7: An example prompt for Fact-Check-Then-RAG

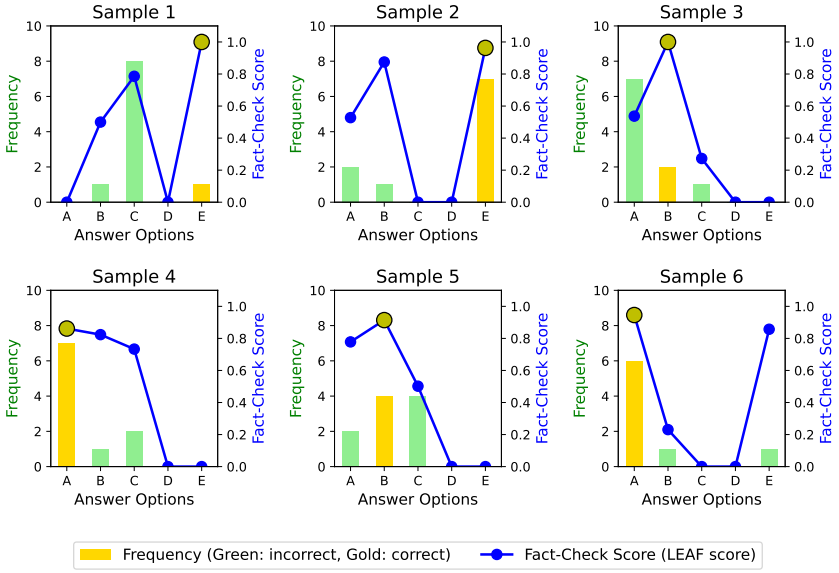

Figure 8: Frequency of selected answer options and their corresponding fact-check scores across multiple samples from the MedQA dataset using the Llama 3 70B Instruct model. For each question, ten responses were generated with a temperature setting of 1.2. The fact-check system assigned scores to each option, with higher scores indicating higher factual accuracy. The correct answers, highlighted in gold, consistently received higher fact-check scores.

Sample 1: A 13-year-old boy presents to the emergency department with severe knee, hip, and groin pain. The patient has a past medical history notable only for obesity and asthma. His temperature is 98°F (36.7°C), blood pressure is 124/65 mmHg, pulse is 128/min, respirations are 14/min, and oxygen saturation is 99% on room air. Physical exam is notable for an inability of the patient to bear weight on his left leg and limited range of motion of the left hip. Which of the following is the best management for this patient?
Choices:

(A) Casting and crutches

(B) Immobilization of the hip in a Pavlik harness

(C) Supportive therapy and observation

(D) Surgical drainage of the hip

(E) Surgical pinning of the femoral head

Sample 2: A 36-year-old nursing home worker presents to the clinic with the complaints of breathlessness, cough, and night sweats for the past 2 months. She further expresses her concerns about the possibility of contracting tuberculosis as one of the patients under her care is being treated for tuberculosis. A PPD skin test is done and reads 11 mm on day 3. Chest X-ray demonstrates a cavitary lesion in the right upper lobe. The standard anti-tuberculosis medication regimen is started. At a follow-up appointment 3 months later the patient presents with fatigue. She has also been experiencing occasional dizziness, weakness, and numbness in her feet. Physical exam is positive for conjunctival pallor. Lab work is significant for a hemoglobin level of 10 g/dL and mean corpuscular volume of 68 fl. What is the most likely cause of her current symptoms?
Choices:

(A) Decreased methionine synthesis

(B) Inhibition of ferrochelatase

(C) Increased homocysteine degradation

(D) Increased GABA production

(E) Decreased ALA synthesis

Sample 3: A 72-year-old woman is admitted to the hospital for treatment of unstable angina. Cardiac catheterization shows occlusion that has caused a 50% reduction in the diameter of the left circumflex artery. Resistance to blood flow in this vessel has increased by what factor relative to a vessel with no occlusion?
Choices:

(A) 64

(B) 16

(C) 8

(D) 4

(E) 32

Figure 9: Sample questions 1-3 from the MedQA dataset

Sample 4: A 49-year-old woman is brought to the emergency department with progressive dyspnea and cough which she developed approx. 8 hours ago. 2 weeks ago she had a prophylactic ovariectomy because of a family history of ovarian cancer. She is known to have type 2 diabetes mellitus and stage 1 hypertension, but she does not take her antihypertensives because she is not concerned about her blood pressure. Also, she has a history of opioid abuse. She takes metformin 1000 mg and aspirin 81 mg. She has been smoking 1 pack of cigarettes per day for 22 years. Her vital signs are as follows: blood pressure 155/80 mm Hg, heart rate 101/min, respiratory rate 31/min, and temperature 37.9C (100.2F). Blood saturation on room air is 89%. On examination, the patient is dyspneic and acrocyanotic. Lung auscultation reveals bilateral rales over the lower lobes. A cardiac examination is significant for S2 accentuation best heard in the second intercostal space at the left sternal border and S3 presence. There is no leg edema. Neurological examination is within normal limits. Arterial blood gases analysis shows the following results: pH 7.49 PaO2 58 mm Hg PaCO2 30 mm Hg HCO3- 22 mEq/L Based on the given data, which of the following could cause respiratory failure in this patient?
Choices:

(A) Increased alveolar dead space due to absent perfusion of certain alveoli

(B) Ischemia of the medullary respiratory center neurons

(C) Alveolar fibrosis

(D) Depression of the respiratory center via opioid receptors activation

(E) Decreased V/Q due to bronchial obstruction

Sample 5: While in the ICU, a 62-year-old male undergoes placement of a Swan-Ganz catheter to evaluate his right heart pressures. All pressures are found to be within normal limits, and the cardiology fellow records a pulmonary wedge pressure of 10 mmHg. Which of the following are normal values for the pressures that will be obtained from this patient's right ventricle?
Choices:

(A) 25/10 mmHg

(B) 25/5 mmHg

(C) 10/0 mmHg

(D) 100/5 mmHg

(E) 100/70 mmHg

Sample 6: A previously healthy 6-year-old boy is brought to the physician because of generalized malaise and a palpable swelling in the left axilla. The parents report that 2 weeks ago, his daycare group visited an animal shelter, after which he developed a rash on the left hand. His temperature is 38.5°C (101.3°F). Physical examination shows three linear crusts on an erythematous background on the dorsum of the left hand. There is tender left-sided axillary and cervical lymphadenopathy. Histopathologic examination of an axillary lymph node shows necrotizing granulomas. The most likely causal organism of this patient's clinical findings is also involved in the pathogenesis of which of the following conditions?
Choices:

(A) Bacillary angiomatosis

(B) Burkitt lymphoma

(C) Condylomata lata

(D) Brucellosis

(E) Bubonic plague

Figure 10: Sample questions 4-6 from the MedQA dataset