CHEF: A Comparative Hallucination Evaluation Framework for Large Language Models

Anonymous ACL submission

Abstract

We introduce **CHEF**, a novel Comparative Hallucination Evaluation Framework that leverages the HaluEval2.0 LLM-in-the-loop hallucination detection pipeline to directly measure the relative effectiveness of hallucination mitigation techniques, specifically retrieval-augmented generation (RAG) and fine-tuning. While HaluEval2.0 provides absolute hallucination scores using a single evaluator LLM, CHEF demonstrates that by evaluating an identical model architecture across three distinct configurations, we can effectively attribute the resulting differences in hallucination rates to each specific technique. Our experiments across science, biomedical, and other domains, conducted using CHEF, reveal variable effectiveness of both RAG and fine-tuning approaches, with significant domain-dependent performance differences. Offering valuable and actionable insights into mitigation strategies.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across numerous tasks, yet hallucination remains a persistent challenge for their deployment in high-stakes domains (Li et al., 2023). While various mitigation strategies exist, there is a critical gap in our ability to systematically compare their effectiveness under consistent evaluation conditions. Existing evaluation frameworks like HaluEval2.0 (Li et al., 2024b) face a fundamental limitation: evaluator hallucination confounds absolute scores (Manakul et al., 2023; Kossen et al., 2024), making it difficult to reliably compare mitigation techniques.

Our key contribution is CHEF, a comparative evaluation framework that shifts focus from single-score reporting to controlled differential analysis. By systematically applying the same evaluation pipeline to three variants of the same base model, CHEF obtains relative hallucination reductions that remain robust to evaluator error. We hypothesize that measuring percentage changes relative to a shared baseline isolates true mitigation effects from evaluator bias. This controlled experimental design isolates the effects of specific mitigation techniques while controlling for model architecture, evaluation methodology, and domain characteristics, representing a systematic comparison of RAG and fine-tuning for hallucination mitigation.

CHEF provides key advantages over traditional benchmarking approaches: (1) **Isolation of mitigation effects**: By controlling model architecture and evaluation methodology, CHEF attributes performance differences specifically to RAG or fine-tuning interventions; (2) **Robustness to evaluator inconsistency**: Relative improvements remain meaningful despite potential systematic error in absolute scores; (3) **Practical guidance**: Our results quantify the relative effectiveness of these mitigation strategies, informing cost-benefit decisions for applications.

2 Related Work

LLM-in-the-loop evaluators Recent work has developed various approaches to detect hallucinations in large language models using the models themselves as evaluators. SelfCheckGPT leverages the insight that if an LLM has knowledge of a given concept, sampled responses are likely to be similar and contain consistent facts, while hallucinated facts tend to cause stochastically sampled responses to diverge and contradict one another (Manakul et al., 2023). This sampling-based approach performs well but increases computational overhead by requiring multiple model generations.

TofuEval (Tang et al., 2024) specifically examines hallucinations in dialogue summarization, highlighting limitations in LLM-based evaluators when tasked with verifying factual consistency. Hallu-

Lens (Bang et al., 2025) extends this work by offering a dynamic taxonomy-based benchmark that distinguishes between intrinsic and extrinsic hallucinations. Meanwhile, Phare's multilingual benchmark (Dora, 2025) confirms the pervasiveness of evaluator errors across languages, emphasizing the need for our comparative framework that controls for such biases.

Mitigation via RAG vs. fine-tuning The effectiveness of RAG and fine-tuning approaches has been investigated in several studies, with complementary findings to our work. Soudani et al. (2024) and Ovadia et al. (2023) demonstrate that RAG particularly excels at addressing low-frequency knowledge queries compared to fine-tuning approaches, supporting our hypothesis that these techniques provide different benefits in hallucination mitigation.

End-to-end RAG pipelines have shown significant improvement in domain-specific factuality (Li et al., 2024a), while fine-tuning remains more resource-intensive (Lakatos et al., 2024). Our work builds on these insights by providing a direct comparative analysis of both approaches within a consistent evaluation framework, allowing for more precise quantification of their relative benefits.

Meta-evaluation and evaluator fallibility A critical challenge in hallucination research is the reliability of the evaluators themselves. McKenna et al. (2023) identify behavioral biases in Natural Language Inference (NLI) tasks that contribute to evaluator hallucinations. FACTOID (Rawte et al., 2024) introduces factual entailment for more precise detection, while HALoGEN (Ravichander et al., 2025) provides a taxonomy and multi-domain verification framework specifically designed to identify evaluator errors.

Our comparative benchmarking approach (CHEF) directly addresses these concerns by focusing on relative improvements rather than absolute scores. By controlling for evaluator biases through differential analysis, we isolate the true effects of mitigation strategies while acknowledging the inherent limitations of LLM-in-the-loop evaluation. The proposed CHEF framework approach aligns with recent work on semantic uncertainty quantification (Kossen et al., 2024), which similarly recognizes the value of comparative metrics over absolute scores for robust hallucination detection.

3 Proposed Framework

CHEF builds upon the HaluEval2.0 hallucination detection pipeline to evaluate three distinct test-time LLM configurations—the baseline test LLM, the same model augmented with Retrieval-Augmented Generation (RAG), and a version fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al., 2022)—all under a shared, LLM-in-the-loop hallucination detection setup.

See Appendix A.1 for a visual overview of the CHEF framework architecture.

The evaluation unfolds in two key stages: (1) identification of hallucinations using HaluEval2.0's extraction and verification procedure, and (2) comparative analysis across the three model variants. This structured setup enables quantification of relative hallucination rates across different mitigation strategies under consistent evaluation conditions.

3.1 Hallucination Detection Pipeline

We adopt HaluEval2.0's three-stage detection pipeline (Li et al., 2024b), applied consistently across all model configurations:

- **Answer Generation:** For each benchmark query, the test LLM generates an answer, forming a QA pair.
- **Fact Extraction:** A separate evaluation LLM identifies atomic factual claims from the QA output using a template-based prompt.
- **Fact Evaluation:** The same evaluator LLM verifies each claim, assigning one of three labels: `True`, `False` (with justification), or `Unknown`.

3.2 Mitigation Strategies

Retrieval-Augmented Generation (RAG) The RAG strategy supplements the LLM with external factual knowledge at inference time through a structured pipeline:

- **Key-Topic Extraction:** Identifying key terms from each query
- **Document Collection:** Retrieving relevant sources
- **Embedding and Retrieval:** Processing documents into chunks for contextual retrieval

The full RAG pipeline implementation is detailed in Appendix A.2.

LoRA-Based Fine-Tuning We apply Low-Rank Adaptation (LoRA) to fine-tune the base LLM with domain-grounded knowledge:

- **Synthetic QA Generation:** Creating domain-specific training examples
- **Training Procedure and Configuration:** Applying parameter-efficient adaptation techniques and balancing knowledge integration with generalization

3.3 Comparative Evaluation

By comparing each variant against the shared baseline, we quantify changes in hallucination rates attributable to each mitigation strategy, controlling for model architecture and evaluation methodology.

4 Experimental Setup

4.1 Dataset

We conduct experiments on the HaluEval2.0 benchmark, comprising 8,770 fact-intensive questions across five domains: Biomedicine (1,535 questions), Finance (1,125), Science (1,409), Education (1,701), and Open Domain (3,000) (Li et al., 2024b). Questions are drawn from BioASQ, NFCorpus, FiQA-2018, SciFact, LearningQ, and HotpotQA, filtered to include only those requiring factual reasoning.

4.2 RAG Implementation Details

For each input question, we first perform *key-topic extraction* by prompting LLM with a lightweight template. This yields a compact, semantically-focused bag of terms (e.g., "colorectal cancer," "metastases," "regional spread," "cancer statistics"), which we have found to generalize more broadly than using the raw questions themselves. We then use the Wikipedia API to retrieve the full text of the top 2–3 pages matching each extracted keyword, yielding 32 thousand pages in total across our benchmark queries. All documents are split into 512-token chunks with 50-token overlap to preserve context, embedded via a local sentence embedding model. At inference, we retrieve the top-k chunks (we set $k = 3$) for answer synthesis.

4.3 Fine-Tuning Implementation Details

Rather than fine-tuning on the original benchmark Q&A pairs, we generate a synthetic, topic-grounded dataset from our scraped documents. For each document in the Science and Bio-Medical domain, we instructed the LLM to generate up to 10 fact-checking questions along with their precise answers based solely on the provided text. This yields over 18,000 Q&A pairs that cover the same topical space as the benchmark yet differ in surface form.

We then fine-tune the base LLaMA (Team, 2024) model using Low-Rank Adaptation (LoRA) (Hu et al., 2022), targeting the Query and Value projection matrices in each attention layer. We set the LoRA rank $r = 36$ and scaling factor $\alpha = 36$ (so that $\alpha/r = 1$) to balance adaptation capacity against parameter efficiency. Training is run for 4 epochs with effective batch size of 24, which we found sufficient to integrate new factual knowledge without overfitting.

4.4 Evaluation & Comparision Metrics

We adopt the standard HaluEval2.0 metrics, recording for each predicted answer:

- **Accuracy:** proportion of claims labeled `True`.
- **False Rate:** proportion labeled `False`.
- **Unknown Rate:** proportion labeled `Unknown`.
- **Micro-Hallucination Rate (MiHR):** the average, over all responses, of the fraction of claims in a response flagged as hallucinated:
- **Macro-Hallucination Rate (MaHR):** proportion of responses with at least one hallucinated claim:
- **Comparison:** To isolate the effect of each mitigation technique, we compute percentage reductions in MiHR and MaHR, as well as accuracy differences, all relative to our shared baseline.

5 Results

5.1 Baseline Performance

The LLaMA 3.2 8B base model demonstrates varied performance across domains. In the Science domain, it achieves the highest accuracy (90.28%) with the lowest hallucination rate (MiHR 6.58%, MaHR 24.28%). In contrast, the Open-Domain exhibits the lowest accuracy (73.29%) and highest hallucination rates (MiHR 17.54%, MaHR 55.50%). Other domains fall between these extremes, with Bio-Medical and Education domains showing similar patterns.

5.2 Effects of RAG

Retrieval-Augmented Generation (RAG) demonstrates mixed effectiveness across domains. In

Table 1: Performance Metrics for Base and RAG Models Across Different Domains

Domain	LLaMA 3.2 8B Base Model				LLaMA 3.2 8B + RAG Model			
	Acc (%)	**MiHR (%)**	**MaHR (%)**	**FR (%)**	**Acc (%)**	**MiHR (%)**	**MaHR (%)**	**FR (%)**
Bio-Medical	87.32	11.50	33.62	11.48	86.78	9.89	34.33	10.57
Science	90.28	6.58	24.28	8.17	89.74	6.87	29.88	7.79
Finance	77.28	9.53	39.47	13.39	79.18	11.69	46.31	13.91
Education	87.57	11.11	35.39	10.94	85.35	8.88	34.22	10.62
Open-Domain	73.29	17.54	55.50	17.35	79.04	4.67	13.73	15.16

Table 2: Performance Metrics for Fine-Tuned Model

Domain	Acc (%)	MiHR (%)	MaHR (%)	FR (%)
Bio-Medical	78.93	16.96	50.42	16.32
Science	91.59	4.97	14.48	5.66

Table 3: RAG Model: Performance Delta vs Base Model

Domain	ΔAcc (%)	ΔMiHR (%)	ΔMaHR (%)
Bio-Medical	-0.62	14.17	-2.11
Science	-0.60	-4.41	-23.06
Finance	2.46	-22.67	-17.33
Education	-2.54	20.07	3.31
Open-Domain	7.85	73.38	75.26

Table 4: Fine-Tuned Model: Performance Delta vs Base Model

Domain	ΔAcc (%)	ΔMiHR (%)	ΔMaHR (%)
Bio-Medical	-9.61	-47.48	-49.55
Science	1.45	24.47	40.36

Open-Domain, RAG was able to drastically decrease the hallucination rates (decreased 73.38% for MiHR and 75.26% for MaHR). In the **Science** domain, RAG slightly decreases accuracy while increasing hallucination rates, particularly MaHR (a 23.06% increse). For **Bio-Medical** queries, RAG reduces MiHR by 14.17% while slightly increasing MaHR. In the **Finance** domain, RAG improves accuracy but increases both hallucination metrics, while in **Education**, it decreases accuracy but reduces hallucination rates.These mixed results suggest domain-specific factors influence RAG effectiveness.

5.3 Effects of Fine-Tuning

Our fine-tuning experiments reveal contrasting outcomes between domains. In the **Science** domain, fine-tuning produces the most promising results, with increased accuracy (90.28% to 91.59%) and substantial reductions in hallucination rates (MiHR from 6.58% to 4.96%, MaHR from 24.28% to 14.48%). In stark contrast, fine-tuning in the **Bio-Medical** domain significantly degrades performance, with decreased accuracy (87.32% to 78.93%) and dramatically increased hallucination rates. This domain-dependent variability suggests that fine-tuning effectiveness is contingent on domain-specific knowledge characteristics.

5.4 Mitigation Strategy Performance Factors

We believe RAG's inconsistent performance stems from context window limitations in our LLaMA 8B model, which struggled to process retrieved information while maintaining query focus, alongside variable Wikipedia coverage quality across domains and degraded responses when confronted with information gaps. Meanwhile, fine-tuning exhibited stark domain dependence, with Science benefiting from high-quality synthetic training data while Bio-Medical suffered significant degradation, possibly due to domain-specific synthetic data challenges or our LoRA implementation (r=36) providing insufficient capacity for specialized terminology domains.

6 Conclusion

In this paper, we introduced CHEF, a Comparative Hallucination Evaluation Framework that enables direct measurement of the relative effectiveness of hallucination mitigation techniques. By evaluating identical model architectures across three configurations CHEF successfully isolates the impact of specific mitigation strategies while controlling for evaluator biases that confound absolute hallucination scores. CHEF's comparative approach represents an important step toward more reliable hallucination benchmarking. By focusing on relative improvements rather than absolute scores, we mitigate the impact of evaluator inconsistency that has hampered previous hallucination detection frameworks.

Limitations

While CHEF provides valuable comparative insights, several limitations remain. First, our evaluation is currently limited to a single base model architecture (LLaMA), which may not generalize to other model families with different pre-training objectives or architectural designs. Second, our RAG implementation relies solely on Wikipedia, potentially limiting its effectiveness for specialized domains requiring more technical resources. Third, the HaluEval2.0 prompts we adopted may not optimally extract or evaluate claims across all domains.

Future work should address these limitations through:

1. **Model diversity:** Extending CHEF to evaluate a wider variety of model architectures (e.g., Mixtral, PaLM, GPT-4, Claude) to understand how mitigation techniques perform across different foundation models.

2. **Prompt refinement:** Enhancing the HaluEval2.0 prompts with domain-specific terminology and structured claim formats to improve fact extraction and evaluation reliability. Exploring chain-of-thought approaches may also lead to more consistent evaluations.

3. **Domain-specific knowledge sources:** Integrating domain-specific databases and literature repositories beyond Wikipedia to better address specialized knowledge domains.

4. **Comprehensive fine-tuning:** Extending our fine-tuning methodology to all domains (Finance, Education, and Open-Domain) to provide a complete comparative analysis across the entire benchmark. This would allow for more robust conclusions about the relative effectiveness of fine-tuning as a hallucination mitigation strategy across diverse knowledge areas.

5. **Evaluator uncertainty quantification:** Incorporating Semantic Entropy Probes (SEPs) as an additional comparison metric to detect and account for evaluator uncertainty. SEPs offer a computationally efficient approach to measuring semantic uncertainty by directly approximating semantic entropy from the hidden states of a single model generation, eliminating the need for multiple sampling runs. This technique would provide a more robust measure of evaluator confidence when determining hallucination rates, potentially improving the reliability of our comparative framework.

The comparative benchmarking approach pioneered in CHEF opens new possibilities for systematic evaluation of hallucination mitigation techniques. As the field continues to advance, we believe this focus on controlled differential analysis, rather than absolute scoring, will be essential for reliable progress measurement in reducing LLM hallucinations.

Acknowledgments

This research was supported by **Anonymous Research Lab**. We thank **Anonymous colleagues/reviewers** who provided feedback. We acknowledge the use of AI assistants, including Claude and GitHub Copilot, which were employed to assist with drafting portions of the manuscript, refining technical explanations, and suggesting code optimizations for our implementation. All AI-generated content was reviewed, edited, and verified by the authors to ensure accuracy and alignment with our research findings.

The HaluEval2.0 benchmark and LLaMA 3.1 model used in this work are employed in accordance with their intended research purposes. HaluEval2.0 is used as intended for benchmarking hallucination detection capabilities, and LLaMA 3.1 is used within the scope of its research license for comparative model evaluation. The synthetic datasets generated in this study are derived from publicly available information and are intended solely for research purposes.

References

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. *arXiv preprint arXiv:2504.17550*.

Matteo Dora. 2025. Good answers are not necessarily factual answers: an analysis of hallucination in leading llms (phare). Giskard.ai benchmark report.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth A. Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*.

Robert Lakatos and 1 others. 2024. Investigating the performance of rag and fine-tuning for ai-driven knowledge-based systems. *arXiv preprint arXiv:2403.09727*.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024a. Enhancing llm factual accuracy with rag to counter hallucinations. *arXiv preprint arXiv:2403.10446*.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024b. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *EMNLP*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *EMNLP*.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad J. Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*.

Abhilasha Ravichander and 1 others. 2025. Halogen: Fantastic llm hallucinations and where to find them. *arXiv preprint arXiv:2501.08292*.

Vipula Rawte and 1 others. 2024. Factoid: Factual entailment for hallucination detection. *arXiv preprint arXiv:2403.19113*.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Fine tuning vs. retrieval augmented generation for less popular knowledge. *arXiv preprint arXiv:2403.01432*.

Liyan Tang and 1 others. 2024. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. *arXiv preprint arXiv:2402.13249*.

Meta AI Team. 2024. Llama 3: A more capable and accessible foundation language model family. Technical report, Meta AI. Model release technical report.

A Appendix

To support future work in explicit content detection, we release the full dataset, annotation scripts, and category definitions at **Anonymous Repository**.

A.1 CHEF Framework Architecture

Figure 1 provides a visual overview of our CHEF framework, illustrating how we evaluate three distinct configurations of the same base model—baseline, RAG-enhanced, and fine-tuned—using a consistent hallucination detection pipeline.

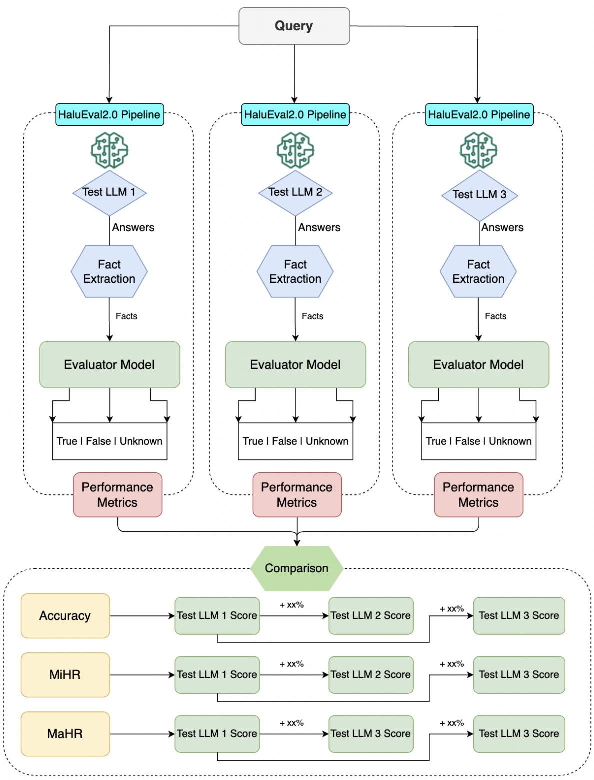

Figure 1: Detailed overview of the CHEF Comparative Benchmarking Framework architecture.

A.2 RAG Pipeline Details

Figure 2 illustrates our RAG implementation, which follows a three-stage process of key-topic extraction, document collection, and embedding-based retrieval as described in Section 3.2.

A.3 Equations

$$\text{MiHR} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{Count}\big(\text{hallucinatory facts in } r_i\big)}{\text{Count}\big(\text{all facts in } r_i\big)}$$
$$\text{MaHR} = \frac{\text{Count}\big(\text{hallucinatory responses}\big)}{n} \tag{1}$$

$$\Delta\text{MiHR} = \frac{\text{MiHR}_{\text{baseline}} - \text{MiHR}_{\text{method}}}{\text{MiHR}_{\text{baseline}}} \times 100\%,$$
$$\Delta\text{MaHR} = \frac{\text{MaHR}_{\text{baseline}} - \text{MaHR}_{\text{method}}}{\text{MaHR}_{\text{baseline}}} \times 100\%. \tag{2}$$

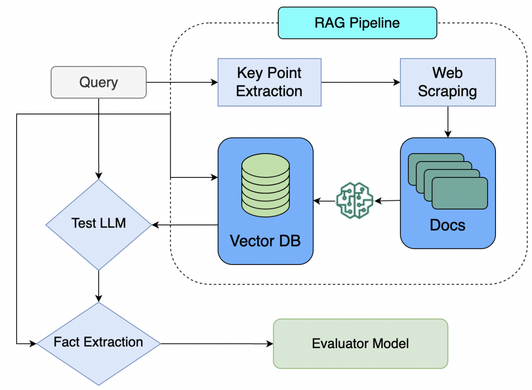

Figure 2: Detailed view of the RAG pipeline used in our experiments.

A.4 Prompts

A.4.1 Key Word Extraction Prompt

```
attached is a json file
filled with queries about
[Domain name] domain
subjects, i want you to go
through each question and
generate keywords and topics
about the question that
could be used in Wikipedia
api search to help find
documents related to that
question. The keywords and
topics should be not too
large. your output format
should be a json array in
this style : [
{
  "id": query id as integer,
  "keywords": [
    "keywords related to query",
    "topics related to query",
    ...
  ]
}, ... ].
```

A.4.2 Synthetic Q&A Generation Prompt

```
You are provided with the
following document:
"""
{document_content}
"""
Your task is to extract
straightforward, fact-based
questions and answers solely
from the document. Rules:

1. Source Strictness: Only
   use information from the
   document!
2. Extraction: Generate
   questions with answers
   from key details.
3. Clarity: Questions must
   be clear and unambiguous.
4. Question Styles: Use
   varied types (True/false,
   What/How is/are, etc.)
5. Quantity: Max 15 quality
   questions.
6. Format: JSON format as:

[{
  "question": "Question?",
  "answer": "Answer."
}, ... ]

Provide only the JSON output.
```