CHEF: a comparative hallucination evaluation framework for large language models

Anonymous Author(s)
Affiliation
Address
email

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 three critical advantages over traditional benchmarking approaches:

1. **Isolation of mitigation effects**: By holding constant the model architecture and evaluation methodology, CHEF allows us to attribute performance differences specifically to RAG or fine-tuning interventions.
2. **Robustness to evaluator inconsistency**: Relative improvements measured by CHEF remain meaningful even when absolute scores contain systematic error.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

3. **Practical guidance**: Our results, derived using CHEF, quantify the relative effectiveness of two popular mitigation strategies, informing cost-benefit decisions for real-world 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. HalluLens 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.

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.

See Appendix A.2 for a visual overview of the CHEF framework architecture.

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.3.

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 GPT-4 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 GPT-4 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.

LoRA parameters are set with rank $r = 36$ and $\alpha = 36$ following prior parameter-efficient adaptation studies Hu et al. [2022], balancing adaptation capacity against training cost. Training is run for 4 epochs with effective batch size of 24, which we found sufficient to integrate new factual knowledge without overfitting.

Due to compute limits, we restrict fine-tuning to Science and Biomedical domains, which we judge most sensitive to hallucination.

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 and Discussion

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.

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

5.2 Effects of RAG

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

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

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.

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

5.4 diagnostics and analysis

While RAG improved open-domain performance, it degraded biomedical results. Closer inspection shows biomedical queries often contain specialized terminology (e.g., gene variants, compound names) poorly covered by Wikipedia, leading to retrieval noise.

By contrast, science queries were generally well-represented in Wikipedia, yet RAG sometimes increased hallucination rates due to context-window overload from irrelevant retrieved passages.

Fine-tuning improved science outcomes, likely because synthetic QAs were unambiguous and factual. In biomedicine, however, synthetic data introduced subtle factual noise (e.g., oversimplified descriptions of complex mechanisms), which the LoRA adapter amplified, leading to degraded accuracy.

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.

References

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

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

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

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

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

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

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

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

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

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

Abhilasha Ravichander et al. Halogen: Fantastic llm hallucinations and where to find them. *arXiv preprint arXiv:2501.08292*, 2025.

Vipula Rawte et al. Factoid: Factual entailment for hallucination detection. *arXiv preprint arXiv:2403.19113*, 2024.

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

Liyan Tang et al. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. *arXiv preprint arXiv:2402.13249*, 2024.

Meta AI Team. Llama 3: A more capable and accessible foundation language model family. Technical report, Meta AI, 2024. URL `https://ai.meta.com/llama/`. 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 Hardware usage

all experiments have been done in **Anonymous Cluster** with single Nvidia A100 gpu. Fine tuning took 4 hours of GPU time for each field. Evaluation per topic took 6 hours on average.

A.2 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.

A.3 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.4 Equations

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

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

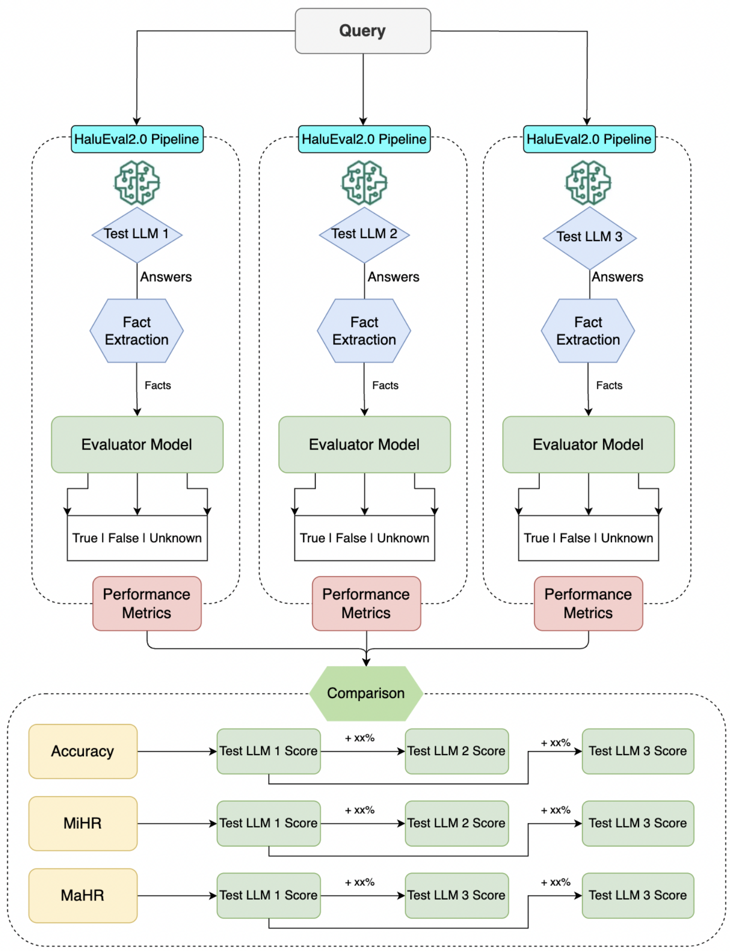

Figure 1: Overview of the CHEF comparative benchmarking framework.

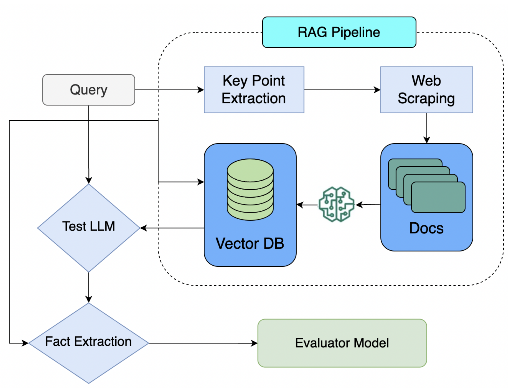

Figure 2: Detailed view of our proposed rag pipeline.

A.5 Prompts

A.5.1 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.5.2 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.

A.6 Licenses and terms of use for existing assets

This work makes use of several existing assets, each with their respective licenses and terms of use:

A.6.1 Models and frameworks

- **LLaMA 3.2 8B**: Licensed under the Meta Llama 3.2 Community License Agreement. Version 3.2 was used for all experiments.
- **GPT-4**: Used via OpenAI API under OpenAI's Terms of Use.
- **LoRA (Low-Rank Adaptation)**: The technique is described in Hu et al. (2022) and the implementation follows the Apache 2.0 licensed PEFT library.

A.6.2 Datasets and benchmarks

- **HaluEval2.0**: Licensed under MIT License.

A.6.3 External knowledge sources

- **Wikipedia**: Content retrieved via Wikipedia API is licensed under Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA 4.0) and GNU Free Documentation License (GFDL). We comply with Wikipedia's API terms of use including appropriate rate limiting and user agent identification. All Wikipedia content used in this work maintains the CC BY-SA 4.0 license requirements.

All assets were used in accordance with their respective licenses and terms of use. For assets requiring attribution, proper citations are provided throughout the paper. No modifications were made to the original datasets beyond the filtering and processing described in the methodology section.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state that CHEF is a comparative framework for evaluating hallucination mitigation techniques, specifically comparing RAG and fine-tuning approaches. The experimental results support these claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: "Limitations" section discusses constraints including single model architecture testing, Wikipedia-only RAG implementation, and incomplete domain coverage for fine-tuning.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper is empirical and does not include theoretical results or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Experimental setup including dataset descriptions (Section 4.1), RAG implementation details (Section 4.2), fine-tuning parameters (Section 4.3), evaluation metrics (Section 4.4) is provided and source code is available in appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Source code with experimental setup and training dataset has been submitted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies LoRA parameters , batch size, epochs, model architecture, and evaluation pipeline details in Sections 4.2-4.4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper reports raw percentage metrics without error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For compute used in evaluation and Fine-Tuning has been explained in appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research focuses on improving hallucination evaluation in LLMs, which aligns with responsible AI development and poses no apparent ethical concerns.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper focuses on technical contributions without discussing broader societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper presents an evaluation framework rather than releasing potentially harmful models or datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets used in the paper are properly credited with citations, and their licenses and terms of use are explicitly documented in Appendix Section

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: source code with experimental setup and training dataset has been submitted.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:The paper does not involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper describes using GPT-4 for evaluation and fact extraction, and LLaMA 3.2 8B as the test model, which are core components of the research methodology.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.