

# Factored Verification: Detecting and Reducing Hallucination in Summaries of Academic Papers

Charlie George

Elicit

charlie@elicit.com

Andreas Stuhlmüller

Elicit

andreas@elicit.com

## Abstract

Hallucination plagues even frontier LLMs—but how bad is it really for summarizing academic papers? We evaluate *Factored Verification*, a simple automated method for detecting hallucinations in abstractive summaries. This method sets a new SotA on hallucination detection in the summarization task of the HaluEval benchmark, achieving 76.2% accuracy. We then use this method to estimate how often language models hallucinate when summarizing across multiple academic papers and find 0.62 hallucinations in the average ChatGPT (16k) summary, 0.84 for GPT-4, and 1.55 for Claude 2. We ask models to self-correct using *factored critiques* and find that this lowers the number of hallucinations to 0.49 for ChatGPT, 0.46 for GPT-4, and 0.95 for Claude 2. The hallucinations we find are often subtle, so we advise caution when using models to synthesize academic papers.

## 1 Introduction

Hallucination—the generation of inaccurate or ungrounded information—is a largely unsolved problem for LLMs (Kryściński et al., 2019; Maynez et al., 2020; Ji et al., 2023). This is acceptable for creative use cases such as story generation and brainstorming, but would be highly problematic if common for academic summarization and Q&A where factual accuracy is key. How common is hallucination for SotA models when answering questions given the abstracts of multiple scientific papers?

To answer this question, we first construct a simple method for checking hallucination inspired by Kadavath et al. (2022) and Lightman et al. (2023): Given a summary, we automatically decompose it into key claims, assign a model-generated probability to each of the claims given the relevant sources, and combine these into an overall correctness probability. We validate this method on the hallucination detection benchmark HaluEval and

## Summary

McMorris (2007) and Ling (2009) found that creatine improved performance on cognitive tasks in adults. Benton (2010) found that creatine improved memory in vegans.

### Claims

McMorris (2007) found that creatine improved performance on cognitive tasks in adults.

Ling (2009) found that creatine improved performance on several cognitive tasks.

Benton (2010) found that creatine improved memory in vegans.

### Critiques

No, McMorris (2007) found improvements to cognitive performance in elders.

Yes, Ling (2009) notes improvements from creatine ethyl ester supplementation.

No, Benton (2019) found that creatine improved memory in vegetarians.

## Revised Summary

McMorris (2007) found that creatine improved cognitive performance in elderly individuals, Ling (2009) in the general population. Benton (2010) found improvements to memory in vegetarians.

Figure 1: Factored Verification splits a summary into claims, checks each claim, and then optionally revises the summary to address the claim critiques. Each step is a language model task.

set a new SotA, exceeding the previous chain-of-thought-based method by 10 absolute percentage points using the same language model.

We then apply Factored Verification to detecting hallucination in a real-world scientific summarization task. Given the abstracts of eight papers and a question, the task is to provide a question-relevant summary. We measure hallucination for SotA models including GPT-4 (OpenAI, 2023) and Claude 2 (Bai et al., 2022a), and estimate that the average summary has between 0.62 and 1.57 hallucinations.

Given that we can automatically detect some hallucinations, can we use this knowledge to reduce them? We treat the claim-wise critiques generated by Factored Verification as model-generated advice (Saunders et al., 2022a) and show that we can reduce detected hallucinations for every model we study, but that significant hallucination remains.

## 2 Detecting hallucination with Factored Verification

We first develop and validate Factored Verification, a simple method for using LLMs to detect hallucinations in settings where the relevant source material is provided.

### 2.1 Defining “hallucination”

We call a claim “hallucinated” if it is not backed by the source material provided in context, even if it could be supported with other sources. For example, if the source material discusses the implementation of a public transport policy and the model-generated summary infers that the policy was aimed at addressing sustainability challenges, this is a hallucination unless the source explicitly talked about this as the goal of the policy.

### 2.2 Method

Following [Lightman et al. \(2023\)](#), we break each summary into a list of claims and then assign each claim a probability of being correct, both using LLM prompting. The claim decomposition prompt is in [Appendix A.1.1](#).

To compute the likelihood that a single claim is correct we use a few-shot prompt with GPT-4 base ([OpenAI, 2023](#)) and look up the probability of the final `Yes` token ([Appendix A.1.2](#)). For ChatGPT, which doesn’t provide access to token probabilities, we ask the model to verify that each claim is supported using few-shot chain-of-thought ([Jason Wei et al.](#)), interpreting the resulting Yes/No answer as a 0/1 probability ([Appendix A.1.3](#)).

Assuming independence of the correctness of claims for simplicity, the probability that the summary is correct is the product of the probabilities of each of the individual claims:

$$P_{\text{summary}} = \prod_{i=1}^n P_{\text{claim}_i} \quad (1)$$

We classify a summary as hallucinated if  $P_{\text{summary}}$  is greater than a threshold  $\theta$ .

### 2.3 Dataset

To measure how well Factored Verification works, we use the summarization task of HaluEval, a hallucination benchmark ([Li et al., 2023](#)). Each item in this task consists of a document and two summaries, one of which contains a subtle hallucination.

Due to cost constraints, we randomly sample 5% of the dataset as a “training” set, and 20% as

a test set (1000 and 4000 items respectively). The training set is only used to set the decision threshold  $\theta$  by running GPT-4 with the aforementioned claim likelihood prompt on the training set examples and calculating the average of  $P_{\text{summary}}$  for both faithful and hallucinated summaries.

## 2.4 Results

Factored Verification is SotA for hallucination detection, exceeding prior few-shot and chain-of-thought approaches for both ChatGPT and GPT-4 ([Table 1](#)).

Model	Few-shot	CoT	Factored
GPT-4	30.9%	75.5%	<b>76.2%</b>
ChatGPT (3.5)	58.5%*	61.2%*	<b>71.2%</b>

Table 1: HaluEval summarization results, showing % accuracy,  $n = 4000$ , \* from [Li et al. \(2023\)](#).

For comparison, we also tested a single-prompt equivalent of Factored Verification, asking ChatGPT to write out each claim and verify it in one go. This achieved an accuracy of 63.3%, which is below the multi-prompt factored method (71.2%) and comparable to standard CoT results (61.2%).

We were surprised by GPT-4’s low 30.9% accuracy when using the exact few-shot prompt from [Li et al. \(2023\)](#). To check our implementation, we ran the same code with ChatGPT and found an accuracy 59.9% on our sample, roughly matching the 58.5% from [Li et al. \(2023\)](#). Given that GPT-4 is much more competitive in the CoT and Factored Verification settings, we suspect that the [Li et al. \(2023\)](#) few-shot prompt has idiosyncratic limitations.

## 3 How often do models hallucinate in scientific summaries?

We now turn to our task of interest, generating summaries based on the abstracts of scientific papers. We will apply Factored Verification to estimate how frequently models hallucinate in this context.

### 3.1 Task

Given a question and eight abstracts, the task is to generate a question-relevant summary where each claim is cited with the paper that backs it up. [Figure 2](#) shows a simplified example with only two papers.

**Query:** What is the impact of creatine on cognition?

**Summary:** The papers suggest that creatine supplementation may have a positive impact on cognition. [Ling \(2009\)](#) found that creatine ethyl ester supplementation improved performance on several cognitive tasks, although the results require replication using objective measures of compliance. [Benton \(2010\)](#) found that creatine supplementation improved memory in vegetarians, but not in those who consume meat.

Figure 2: Example of the scientific summarization task, simplified from eight to two papers.

The dataset of queries and abstracts<sup>1</sup> is generated by using 200 anonymized search queries and their associated top results from Elicit ([Stuhlmüller and Byun, 2023](#)).

### 3.2 Method

We follow the hallucination detection strategy outlined in Section 2.2, with the following modifications to reduce compute cost:

1. We directly treat sentences as claims.
2. We only check claims that have associated citations.
3. When checking each claim, we provide only the abstracts of the cited papers as sources.

We expect that these modifications lead to little degradation given that almost all sentences have citations and the simplification step from sentence to claim is not doing much work.

### 3.3 Results

We run Factored Verification with ChatGPT, GPT-4, Claude 2, and Claude Instant ([OpenAI, 2023](#); [Bai et al., 2022a](#)). Table 2 shows that for all models, our method reports at least one hallucination in the majority of summaries.

We include additional results that show interactions between ChatGPT and GPT-4 when used as generation, criticism, and judge models in Table 3 in the Appendix.

### 3.4 Interpretation

Based on the 76% accuracy of Factored Verification on HaluEval, we know that there are likely false

<sup>1</sup><https://github.com/elicit/fave-dataset>

positives and/or false negatives, so we can't take the reported hallucination rates literally.

We manually inspected about a hundred claims evaluated by GPT-4. When GPT-4 said that a claim is supported, we agreed in all cases. When GPT-4 reported an unsupported claim, we agreed 66% of the time. So, our best guess for the true hallucination rate is  $2/3$  of the reported hallucination rate.

Many of the claims we encountered were wrong in subtle ways that we would likely have missed without seeing the GPT-4 critiques, and would expect non-expert evaluators to miss, including:

- Stating that a claim is supported by two abstracts when it is only supported by one
- Slightly exaggerating the findings of a paper
- Conflating the purpose of the study with the outcome
- Implying that two independent findings are linked

This augmentation of human evaluation is consistent with prior work by [Saunders et al. \(2022b\)](#) which found that model-generated critiques help humans find flaws in summaries.

## 4 Reducing hallucination in scientific summaries with Factored Verification

It is common for LLMs to apparently fail at a task, only to then succeed with better prompting. Can we prompt models using the detected inaccuracies to automatically reduce hallucination in scientific summaries?

### 4.1 Baseline

We ask GPT-4 to self-correct by first identifying false claims in its initial summary, then revising the summary given this correction (prompts in Appendix A.4.1 and A.5). This *increased* the average number of detected hallucinations from 1.55 to 2.13. [Huang et al. \(2023a\)](#) similarly found that the GPT-4 generation of LLMs struggles to directly self-correct across a variety of reasoning datasets.

### 4.2 Method

To improve on the baseline, we propose to reduce hallucination with Factored Verification in three steps, as illustrated in Figure 1 and shown in Algorithm 1.

Model	Hallucinations per summary (reported)	Hallucinations per summary (adjusted)	% of summaries with reported hallucinations
GPT-4	1.26 → 0.69	0.84 → 0.46	63.25% → 40%
ChatGPT (3.5, 16k)	0.93 → 0.735	0.62 → 0.49	54% → 41.63%
Claude 2	2.32 → 1.43	1.55 → 0.95	83.0% → 71.50%
Claude instant	2.35 → 1.86	1.57 → 1.24	87.0% → 81.50%

Table 2: Prevalence of hallucination for models when generating summaries of academic papers, before and after revision with factored critiques. Based on manual inspection of approximately 100 data points our best guess is that the true prevalence of hallucination (“adjusted”) is 2/3 of the reports from automated evaluation.

First, we create claim-wise critiques (true/false judgments and supporting reasoning) analogous to the hallucination detection method above: We ask the model to evaluate the supportedness of each sentence based on the cited abstracts. We then concatenate the critiques of the unsupported claims to form the *factored critique*. Finally we ask the model to revise the summary given that critique.

**Algorithm 1** Factored Verification: Revising a summary by generating sentence-wise critiques

---

```

1: Initialize empty list for critiques
2: for each sentence in the summary do
3:   Critique ← LLM.critique(sentence, cited abstracts)
4:   if sentence is unsupported then
5:     Add Critique to the list of critiques
6:   end if
7: end for
8: FactoredCritique ← concat(critiques)
9: RevisedSummary ← LLM.revise(FactoredCritique)
10: return RevisedSummary

```

---

Figure 3 shows an example critique.

### 4.3 Results

Table 2 shows that Factord Verification reduces the number of summaries with reported hallucination by 5.50% to 23.25% (absolute) depending on the model, with ChatGPT being the lowest-hallucination model before critique (0.62 estimated hallucinations per summary), and GPT-4 being the lowest-hallucination model after critique and across all settings (0.46 estimated hallucinations per summary).

## 5 Related work

Hallucination is widely known to be a significant problem for LLMs (Luo et al., 2023; Peng et al., 2023; Ji et al., 2023), although to a much lesser extent for abstractive summarization where the information needed to answer is fully provided (Cao

et al., 2022; Huang et al., 2023c).

Various strategies have been proposed to mitigate hallucination. Some strategies aim to prevent their occurrence by checking how familiar models are with instructions (Luo et al., 2023). Others, including our proposed method, focus on the detection and correction of hallucinations post-generation (Cao et al., 2022; Huang et al., 2023b).

Of these, notable strategies are the use of external knowledge and retrieval augmentation, and automated feedback (Shuster et al., 2021; Peng et al., 2023; Zhang et al., 2023). While external knowledge is less relevant here, Factored Verification can be viewed as a kind of automated feedback.

In simultaneous work, Dhuliawala et al. (2023) propose an automated feedback method called Chain-of-Verification, which is effectively the same as ours: (i) Draft an initial response, (ii) plan fact-checking questions, (iii) answer the questions independently, and (iv) generate a final revised response. Their evaluation focuses on out-of-context fact checking. Consistent with our results, they find that the factored version of their approach reduces hallucination for question-answering and text generation tasks.

## 6 Relevance to AI alignment

To align powerful AI systems, we need to be able to provide accurate feedback and supervision even when systems surpass human-level performance, a property known as scalable oversight (Amodei et al., 2016).

Today, reinforcement learning from human feedback (RLHF) is often used to align LLMs (Christiano et al., 2023; Bai et al., 2022b; Ouyang et al., 2022). However, GPT-4 already surpasses the performance of the average human on many academic tasks (OpenAI, 2023), making it difficult for non-experts to provide effective reward signals. In our

**Critique:**

- For the sentence "*Giuntella (2015) found that immigration reduced waiting times for outpatient referrals, suggesting that policies affecting population dynamics can indirectly impact waiting times and thus healthcare costs.*" in the summary: The claim accurately reflects the findings of [Giuntella \(2015\)](#) that immigration reduced waiting times for outpatient referrals. However, the claim extends the findings of the study to suggest that policies affecting population dynamics can indirectly impact waiting times and thus healthcare costs. While this may be a reasonable inference, it is not directly supported by the abstract.

- For the sentence "*Propper (2002) found that GP fundholders were able to secure shorter waiting times for their patients, suggesting that giving family doctors the ability to choose the hospital where their patients are treated and the means to pay for some services can reduce waiting times.*" in the summary: The claim is not directly supported by the abstract. The abstract states that the study investigates whether GP fundholders were able to secure shorter waiting times for their patients, but it does not provide the results of this investigation. Therefore, the claim that [Propper \(2002\)](#) found that GP fundholders were able to secure shorter waiting times for their patients is not supported by the abstract.

Figure 3: Example of a factored critique generated with GPT-4

attempts to delegate evaluation of academic claims to non-expert contractors, we observed only 38% inter-rater agreement for unsupported claims, a sign of similar difficulties. In the short term this can be solved by using contractors with specialized domain knowledge. However, this won't work if models surpass the capabilities of the best humans.

We have shown that factored critiques let models correct some of their own mistakes without need for human supervision. If similar approaches can be extended beyond hallucination reduction to richer tasks, they could help us scale supervision in lockstep with future model capabilities.

## 7 Discussion

Our main finding is that the absolute rate of hallucination of SotA models like ChatGPT, Claude 2, and GPT-4 is surprisingly high for academic summarization. This is true even with revision using factored critiques, which results in 0.46 to 1.24 estimated hallucinations per summary.

A natural question to ask in this context is whether we can finetune on model-revised summaries, incrementally bootstrapping to more and more accurate summaries, initially detecting and eliminating the most egregious failure modes, then more subtle ones with each training iteration.

Overall, despite incredible advances, language models still struggle with accurate summarization in academic contexts. Many mistakes are only clear upon careful inspection of the sources and look identical to genuine answers otherwise. For now, we advise caution in situations where accuracy matters, as we would for human summaries as well.

## 8 Acknowledgments

We thank Noah Goodman, William Saunders, Jungwon Byun and James Brady for helpful early discussions. We would also like to thank Owain Evans and Justin Reppert for their feedback on early drafts. In addition, we are grateful to Ben Rachbach for help with human labeling and the rest of the Elicit team for supporting this work.

## References

- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. [Concrete Problems in AI Safety](#). ArXiv:1606.06565 [cs].
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. [Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback](#). arXiv:2204.05862 [cs].
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022b. [Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback](#). ArXiv:2204.05862 [cs].
- Meng Cao, Yue Dong, and Jackie Cheung. 2022. [Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.
- Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. [Deep reinforcement learning from human preferences](#). ArXiv:1706.03741 [cs, stat].
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. [Chain-of-Verification Reduces Hallucination in Large Language Models](#).
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023a. [Large Language Models Cannot Self-Correct Reasoning Yet](#). ArXiv:2310.01798 [cs].
- Kung-Hsiang Huang, Hou Pong Chan, and Heng Ji. 2023b. [Zero-shot Faithful Factual Error Correction](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5660–5676, Toronto, Canada. Association for Computational Linguistics.
- Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2023c. [The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey](#).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. [Chain of Thought Prompting Elicits Reasoning in Large Language Models](#).
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2023. [Survey of Hallucination in Natural Language Generation](#). *ACM Computing Surveys*, 55(12):1–38.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. [Language Models \(Mostly\) Know What They Know](#).
- Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. [Evaluating the Factual Consistency of Abstractive Text Summarization](#).
- Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. [HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models](#).
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Karl Cobbe, and Ilya Sutskever. 2023. [Let’s Verify Step by Step](#).
- Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. [Zero-Resource Hallucination Prevention for Large Language Models](#).
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On Faithfulness and Factuality in Abstractive Summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- OpenAI. 2023. [GPT-4 Technical Report](#).
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). ArXiv:2203.02155 [cs].

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. [Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback](#).

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022a. Self-critiquing models for assisting human evaluators. <https://arxiv.org/abs/2206.05802v2>.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022b. [Self-critiquing models for assisting human evaluators](#). ArXiv:2206.05802 [cs].

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. [Retrieval Augmentation Reduces Hallucination in Conversation](#).

Andreas Stuhlmüller and Jungwon Byun. 2023. [Elicit: The AI Research Assistant](#). <https://elicit.org/>.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023. [Mitigating Language Model Hallucination with Interactive Question-Knowledge Alignment](#).

## A Appendix

### A.1 Factored Verification prompts (HaluEval)

#### A.1.1 Decomposing a summary into claims

Below is a summary of a document.  
Please extract ALL the claims from the document. You should give your answer as a list separated by "-" and start by saying "The claims are:"

[summary].

#### A.1.2 Verifying the correctness of a claim with GPT-4 base

Below are a set of documents and claims.  
We will check if the document that the claim is supported by the document or otherwise inaccurate. Below are some examples. It can sometimes be the case that a claim is very subtly wrong.

[Few-shot examples]

Example 5:

Document: [Document]

Claim: [Claim]

Supported: Yes

#### A.1.3 Verifying the correctness of a claim with ChatGPT

Below is a claim and a document. Check that the claim is supported by the document. If it is, say "Yes". If it is not, say "No".

Document: [document]

Claim: [claim]

Give your answer in the following format:

Reasoning: [give your reasoning (including quotes) here]

Supported: [Yes/No]

Remember you MUST include quotes in your reasoning.

## A.2 Prompt template for generating summaries of academic papers

I now need you to help me summarize many more papers in the same way as above. Our research question is "[question]".

I've collected many papers that might address this research question.

Paper [number]: [reference]

Title: [title]

Abstract: [abstract]

Write a summary of what the papers collectively say about the research question. Use the same format as the summary above.

You must cite the papers in your summary. You can use the following format:  
Author (year)

You will only include the findings that directly answer our research question, ignoring other findings that are only loosely relevant. Remember to include citations in the final summary. Your final summary should use varied and engaging language.

## A.3 Prompt templates for Factored Verification (academic papers)

### A.3.1 Generating claim-wise critiques

I need some more help verifying some claims from scientific papers.

The claim is from [paper references]:  
[reference]:

Title: [title]

Abstract: [abstract]

==

Claim: [claim]

First give a critique of the claim.

Then, say whether it is supported by the abstract["s" if we have multiple abstracts]. Finally, if claim is not supported give a revised claim that is supported by the abstract["s" if we have multiple abstracts].

If the claim is partially supported say "No" for the "Supported" field and give a revised claim that is fully supported by the abstract.

Format:

Critique: [critique]

Supported: "Yes" or "No"

Revised Claim: [revised claim] or "N/A" if claim is supported.

### A.3.2 Revision based on claim-wise critiques

As a follow-up to the papers and model-provided summary:

Ok, after reading your summary, I have some feedback:

Feedback:

I have some concerns about the factual accuracy of the summary:

- For the sentence "[original false claim]" in the summary: [critique]

===

Can you correct your summary incorporating each piece of my feedback? The concerns are MOST important to address. Start by writing "Corrected summary:" and then your corrected summary. Keep everything not mentioned in my feedback the same.

## A.4 Prompt templates for self-correction baseline (academic papers)

### A.4.1 Generating self-correction feedback

Below is a list of academic papers.

[Papers]

This is a summary of the papers:

[summary]

Please read the papers and the summary and give feedback. The feedback should ONLY look the at factual accuracy of the summary and make sure that any claims made are FULLY supported by the relevant papers. Write "Feedback:" and then your feedback. You should give a VERY harsh long and detailed piece of feedback.

### **A.5 Revision based on self-generated feedback**

Ok, after reading your summary, I have some feedback:

Feedback:

[Model feedback from prompt above]

Can you correct your summary incorporating each piece of my feedback? The concerns are MOST important to address. Start by writing "Corrected summary:" and then your corrected summary. Keep everything not mentioned in my feedback the same.

### **A.6 Additional results**

See Table 3.

Summary model	Critique model	Judge	Hallucinations per summary (reported)	% of summaries with reported hallucinations
ChatGPT	-	GPT-4	0.89	51.00%
ChatGPT	ChatGPT	GPT-4	0.98	52.00%
ChatGPT	GPT-4	GPT-4	0.45	28.00%
GPT-4	-	GPT-4	1.55	69.50%
GPT-4	ChatGPT	GPT-4	1.19	67.00%
GPT-4	GPT-4	GPT-4	0.51	29.50%
GPT-4	-	ChatGPT	0.84	48.00%
GPT-4	ChatGPT	ChatGPT	0.37	23.50%
ChatGPT	-	ChatGPT	0.97	57.00%
ChatGPT	GPT-4	ChatGPT	0.85	49.50%
ChatGPT	ChatGPT	ChatGPT	0.66	37.00%

Table 3: Interaction effects between ChatGPT and GPT-4 as summary, critique, and judge models. ChatGPT refers to the GPT-3.5 series with 16k context. Dashes indicate that no revision was used. In the main paper, for ChatGPT and GPT-4, we average over { ChatGPT, GPT-4 } as critique generation and evaluation models to reduce interaction effects. For Claude models, we use ChatGPT as a judge.