Benchmarking LLM Faithfulness in RAG with Evolving Leaderboards

Manveer Singh Tamber[1*], **Forrest Sheng Bao**[3*], **Chenyu Xu**[2,3], **Ge Luo**[2], **Suleman Kazi**[2], **Minseok Bae**[4*], **Miaoran Li**[3*], **Ofer Mendelevitch**[2], **Renyi Qu**[2], **Jimmy Lin**[1]

[1] University of Waterloo [2] Vectara [3] Iowa State University [4] Stanford University

Correspondence: {mtamber, jimmylin}@uwaterloo.ca, ofer@vectara.com, forrest.bao@gmail.com

Abstract

Retrieval-augmented generation (RAG) aims to reduce hallucinations by grounding responses in external context, yet large language models (LLMs) still frequently introduce unsupported information or contradictions even when provided with relevant context. This paper presents two complementary efforts at Vectara to measure and benchmark LLM faithfulness in RAG. First, we describe our original hallucination leaderboard, which has tracked hallucination rates for LLMs since 2023 using our HHEM hallucination detection model. Motivated by limitations observed in current hallucination detection methods, we introduce FaithJudge, an LLM-as-a-judge framework that leverages a pool of diverse human-annotated hallucination examples to substantially improve the automated hallucination evaluation of LLMs. We introduce an enhanced hallucination leaderboard centered on FaithJudge that benchmarks LLMs on RAG faithfulness in summarization, question-answering, and data-to-text generation tasks. FaithJudge enables a more reliable benchmarking of LLM hallucinations in RAG and supports the development of more trustworthy generative AI systems: https://github.com/vectara/FaithJudge.

1 Introduction

Large language models (LLMs) excel in various tasks, but frequently produce hallucinations, generating false or misleading information unsupported by provided contexts or world knowledge (Ji et al., 2023; Huang et al., 2025; Lin et al., 2022; Tang et al., 2023). While retrieval-augmented generation (RAG) approaches (Guu et al., 2020; Lewis et al., 2020b; Shuster et al., 2021) seek to mitigate hallucinations by grounding responses in trusted contexts, they do not fully eliminate hallucinations, as LLMs often introduce details unsupported by retrieved contexts, misrepresent information, or generate outright contradictions (Niu et al., 2024).

An ongoing challenge within RAG is evaluating and ensuring context-faithfulness (Niu et al., 2024; Jia et al., 2023; Ming et al., 2024). In this paper, we mainly focus on evaluating faithfulness in summarization tasks, building upon extensive prior research on summary consistency evaluation. Summarization tasks provide a practical benchmark for faithfulness, thanks to rich available hallucination datasets and established automated evaluation methods. However, despite recent progress, both fine-tuned detection models and LLM-as-a-judge techniques (Zheng et al., 2023; Luo et al., 2023; Jacovi et al., 2025) continue to struggle with accurately identifying hallucinations in LLM outputs.

We present two complementary efforts at Vectara for measuring and benchmarking LLM faithfulness in RAG. First, we describe our original hallucination leaderboard, which since 2023 has tracked hallucination rates for LLMs, currently ranking over 160 LLMs on hallucinations in summarization tasks. Second, motivated by the limitations of current hallucination detection approaches, we introduce FaithJudge, an LLM-as-a-judge framework that leverages a pool of diverse, human-annotated hallucination examples to improve automated faithfulness evaluation.

FaithJudge leverages labelled hallucination annotations from diverse LLM generations to automate the evaluation of LLMs on their propensity to hallucinate when summarizing the same articles or using the same articles to respond to queries. This approach results in notably higher agreement with human judgments compared with existing automated methods.

While our main investigation focuses on summarization tasks, we expand FaithJudge to other RAG tasks (including QA and data-to-text generation) using RAGTruth (Niu et al., 2024), further detailed in Section 7.

[*] Work done while at Vectara.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 799–811
November 4-9, 2025 ©2025 Association for Computational Linguistics

Unlike previous work in the automated evaluation of LLMs for faithfulness in RAG, FaithJudge recasts LLM hallucination evaluation: the judge LLM studies a handful of human-annotated peer responses to the same source and then rules on a fresh candidate. Because the judge learns directly from the examples in its prompt, it needs no extra training, adapts naturally to new domains and tasks given annotations, and reaches state-of-the-art agreement with human annotators compared to existing methods while remaining fully automated.

2 Background

Numerous datasets have been developed for evaluating hallucinations in summarization tasks. Earlier datasets, such as SummaC (Laban et al., 2022) and AggreFact (Tang et al., 2023), aggregated multiple resources and standardized labels and classification taxonomies. However, these primarily focused on summaries from pre-ChatGPT models like fine-tuned T5 (Raffel et al., 2020), BART (Lewis et al., 2020a), and PEGASUS (Zhang et al., 2020), potentially limiting their relevance to contemporary LLMs that may produce more nuanced and difficult-to-identify hallucinations.

Recent benchmarks address this limitation by incorporating summaries generated by modern LLMs. TofuEval (Tang et al., 2024b) provided hallucination labels on topic-focused dialogue summarization tasks with LLMs including GPT-3.5-Turbo, Vicuna (Chiang et al., 2023) and WizardLM (Xu et al., 2023). Similarly, HaluEval (Li et al., 2023) included ChatGPT-generated hallucinations across summarization, question-answering (QA), and dialogue tasks, while RAGTruth (Niu et al., 2024) also annotated responses from models including GPT-3.5, GPT-4 (OpenAI, 2023), Llama-2 (Touvron et al., 2023), and Mistral (Jiang et al., 2023). FaithBench (Bao et al., 2025) presented human annotations of challenging hallucinations in summaries from 10 modern LLMs from 8 different model families (detailed further in Section 4).

Due to limited large-scale, human-annotated data for training hallucination detectors, early detection methods relied heavily on natural language inference (NLI) or question-answering (QA) systems (Fabbri et al., 2022). For instance, SummaC aggregated sentence-level NLI entailment scores between document-summary sentence pairs. AlignScore (Zha et al., 2023) extended this by training detection models on multiple semantic alignment tasks evaluated at the chunk level. MiniCheck (Tang et al., 2024a) addressed data scarcity by synthesizing hallucinated examples using GPT-4 for model training.

The strong zero-shot instruction-following capabilities of modern LLMs have also enabled LLM-as-a-judge methods (Zheng et al., 2023; Luo et al., 2023; Jacovi et al., 2025; Gao et al., 2023). Instead of evaluating entire generated summaries, approaches like FACTSCORE (Min et al., 2023) and RAGAS (Es et al., 2024) decompose summaries into claims for granular hallucination detection.

Like our original hallucination leaderboard, efforts such as FACTS Grounding (Jacovi et al., 2025) and Galileo's Hallucination Index (Galileo, 2023) also provide leaderboards to benchmark hallucinations in LLMs, relying on LLM judges employed in a zero-shot manner to evaluate responses.

LLM judges have also been employed to evaluate various aspects of RAG, including the relevance of retrieved passages (Thomas et al., 2024), citation faithfulness (Liu et al., 2023), and the factuality and completeness of responses (Pradeep et al., 2025).

Nonetheless, hallucination detection in RAG remains challenging, with weak effectiveness observed across current methods. Benchmarks such as AggreFact, RAGTruth, TofuEval, and FaithBench consistently show limitations in existing methods, including LLM-based ones. Notably, FaithBench highlighted that current methods, including using LLMs for classification, achieved near 50% accuracy, suggesting negligible ability to identify hallucinated responses. Further, both RAGTruth and TofuEval suggest that smaller, fine-tuned detection models can perform competitively with or even outperform zero-shot LLM-based evaluation approaches.

3 Vectara's Original Hallucination Leaderboard

In 2023, Vectara's hallucination leaderboard was released using Vectara's hallucination detection model, HHEM-1.0-open. This model was later updated to HHEM-2.0 with improved effectiveness, the ability to handle longer contexts, and multilingual capabilities. The current leaderboard relies on the open version, HHEM-2.1-open, publicly released on HuggingFace. To date, HHEM has been downloaded over 4 million times, reflecting strong community interest and adoption. While specific training details remain confidential, we note that

HHEM-2.1-open was trained using the RAGTruth training set among other datasets.

To build our original hallucination leaderboard, articles were selected from diverse sources such as BBC, CNN, Wikipedia, and the Daily Mail, following prior work on summarization evaluation and factuality verification (Narayan et al., 2018; Maynez et al., 2020; Schuster et al., 2021; Thorne et al., 2018; Fabbri et al., 2021; Huang et al., 2020; Pagnoni et al., 2021; Hermann et al., 2015). Articles containing objectionable content, which LLMs may refuse to summarize, were excluded. The resulting dataset is composed of 955 articles with a median length of approximately 217 words.

LLMs are evaluated by prompting them to generate concise summaries strictly grounded on the provided passages. HHEM then assesses the proportion of summaries generated by the LLM containing hallucinations. Refusals are tracked by measuring the proportion of short responses (5 words or fewer). Users are also invited to submit specific models for evaluation. Continuously updated, the leaderboard now benchmarks hallucination rates of over 160 different LLMs, typically evaluating new models as soon as they become publicly available to track ongoing advances in LLMs.

4 FaithBench

FaithBench (Bao et al., 2025) examined hallucinations in diverse LLM-generated summaries and assessed the effectiveness of hallucination detection methods through human annotations. It included summaries from ten state-of-the-art LLMs, including GPT-4o, GPT-3.5, Claude-3.5-Sonnet, Gemini-1.5-Flash (Gemini Team, 2024), and open-source models like Llama-3.1 (Grattafiori et al., 2024), revealing that hallucinations remain frequent and detection methods generally fail to identify them reliably or accurately.

Human annotators labelled hallucinations as Unwanted when the summary contained contradictory or unsupported information, Benign when the information was supported by world knowledge, but absent from the article, or Questionable when the classification was unclear.

Articles in FaithBench were selected based on frequent disagreements on summaries among hallucination detection models. True-NLI, TrueTeacher, HHEM-2.1-open, and GPT-4o/GPT-3.5 judges using the chain-of-thought (CoT) prompt from Luo et al. (2023) were used to identify articles where summary hallucination classifications were most disagreed upon. The dataset includes 75 articles, each with ten annotated summaries from different LLMs, allowing for many diverse LLM summaries to be studied per article. We show that this diversity in summaries also proves to be useful for FaithJudge when judging new summaries.

5 FaithJudge

Human annotation is the gold standard for hallucination detection, but it is time-consuming and expensive. FaithJudge offers a scalable alternative by leveraging hallucination annotations to guide an LLM judge in evaluating new summaries. We also expand FaithJudge to other RAG tasks, including question-answering (QA) and writing overviews from structured data in the JSON format using the RAGTruth dataset (Niu et al., 2024). This is detailed further in Section 7.

To assess an LLM's response, FaithJudge involves prompting an LLM judge with other responses to the same prompt, along with their corresponding hallucination annotations. These annotations include hallucination spans, source references from the context, and labels of either Benign, Unwanted, or Questionable, identified by multiple human annotators.

To evaluate the effectiveness of FaithJudge, we use the fact that each FaithBench article has summaries from ten different LLMs. The judge is given the other nine annotated summaries as context, and its assessments on each summary from FaithBench are compared to human annotations. As shown in Section 6, FaithJudge substantially improves automated hallucination evaluation, outperforming existing detection methods by leveraging human-labelled examples. This allows for more accurate automated hallucination evaluation, where existing hallucination detection methods continue to lag.

6 Evaluating Hallucination Detectors

6.1 Evaluation Datasets

We evaluate hallucination detection methods on four summarization datasets: FaithBench, AggreFact (Tang et al., 2023), RAGTruth (Niu et al., 2024), and TofuEval-MeetingBank (Tang et al., 2024b). While each of these datasets has previously analyzed hallucination detection individually, we provide a comparison across all four, motivating the need for our FaithJudge approach.

Method	# Params	AggreFact-SOTA Acc (%)	AggreFact-SOTA F1 (%)	RAGTruth-Summ Acc (%)	RAGTruth-Summ F1 (%)	TofuEval-MB Acc (%)	TofuEval-MB F1 (%)	FaithBench Acc (%)	FaithBench F1 (%)	**Average** Acc (%)	**Average** F1 (%)
Claim-wise											
Fine-Tuned Hallucination Detection Models											
HHEM-1.0-Open	184M	76.0	71.0	66.2	52.2	54.4	49.9	59.3	58.8	64.0	58.0
HHEM-2.1-Open	110M	73.2	69.7	67.7	56.1	60.9	61.2	66.7*	63.7*	67.1	62.7
AlignScore-base	125M	69.5	61.9	60.2	42.4	51.7	44.0	60.6	59.9	60.5	52.1
AlignScore-large	355M	73.9	69.3	67.9	54.1	56.1	52.9	62.8	59.6	65.2	59.0
MiniCheck-Roberta-L	355M	75.7	72.5	70.5	58.6	67.6	68.5	61.6	60.0	68.8	64.9
Bespoke-MiniCheck	7B	74.3	70.1	73.3	62.9	76.9	78.4	60.1	58.0	71.2	67.3
TrueTeacher	11B	71.8	70.5	56.5	56.1	58.5	58.4	59.8*	51.8*	61.7	59.2
Zero-Shot Hallucination Detection with LLMs											
RAGAS Prompt											
Qwen-2.5	7B	71.1	69.0	68.2	64.4	64.3	57.7	57.9	51.3	65.4	60.6
Qwen-2.5	72B	74.4	69.9	75.3	64.1	69.2	70.6	64.3	57.3	70.8	65.5
Llama-3.1	8B	69.7	65.9	68.5	59.9	72.6	74.4	60.3	57.9	67.8	64.5
Llama-3.3	70B	77.3	74.9	80.0	75.1	73.2	70.6	58.9	49.6	72.3	67.5
GPT-4o	?	75.9	70.3	75.7	63.7	75.2	76.7	65.3	59.0	73.0	67.4
o3-mini-high	?	77.3	72.6	74.6	62.9	69.2	70.6	67.4	60.7	72.1	66.7
Summary-wise											
Fine-Tuned Hallucination Detection Models											
HHEM-1.0-Open	184M	78.9	79.7	53.4	51.4	56.5	39.8	50.5	40.1	59.8	52.7
HHEM-2.1-Open	110M	76.6	76.2	64.4	67.1	69.4	62.1	52.6*	32.9*	65.8	59.6
AlignScore-base	125M	73.8	73.9	57.6	58.2	65.6	52.8	51.3	33.8	62.1	54.7
AlignScore-large	355M	72.7	74.2	52.8	49.6	57.4	39.2	50.3	26.1	58.3	47.3
MiniCheck-Roberta-L	355M	74.2	72.1	66.3	60.9	54.4	45.4	55.0	53.2	62.5	57.9
Bespoke-MiniCheck	7B	79.9	80.4	79.4	77.1	78.8	78.6	55.7	47.3	73.5	70.8
TrueTeacher	11B	77.6	78.4	61.6	62.8	57.4	39.2	53.3*	36.7*	62.5	54.3
Zero-Shot Hallucination Detection with LLMs											
FACTS Grounding Prompt											
Qwen-2.5	7B	66.9	68.7	61.5	63.4	62.8	54.4	52.6	33.5	60.9	55.0
Qwen-2.5	72B	71.6	73.7	74.0	77.5	68.8	58.8	55.2	35.5	67.4	61.4
Llama-3.1	8B	55.5	55.5	62.9	62.7	55.3	54.5	60.9	49.7	58.6	55.6
Llama-3.3	70B	79.3	78.1	81.6	74.9	70.1	71.3	66.6	58.4	74.4	70.7
GPT-4o	?	81.6	78.8	82.6	76.6	76.3	76.0	65.9	56.2	76.6	**71.9**
o3-mini-high	?	82.1	77.8	79.8	70.6	69.2	70.6	68.8	60.7	75.0	69.9
Luo et al. Prompt											
Qwen-2.5	7B	72.8	73.5	67.6	70.2	69.0	66.3	53.4	39.0	65.7	62.2
Qwen-2.5	72B	78.4	78.0	81.3	81.1	83.4	80.0	58.3	44.3	75.3	70.8
Llama-3.1	8B	60.8	51.2	63.7	52.1	57.1	55.8	51.3	51.0	58.2	52.5
Llama-3.3	70B	79.2	79.1	81.3	82.9	73.6	66.5	58.8	43.6	73.2	68.0
GPT-4o	?	80.4	77.5	85.1	80.9	81.6	78.7	62.5*	50.6*	**77.4**	**71.9**
o3-mini-high	?	82.6	80.9	83.2	80.6	75.6	73.7	63.3	49.8	76.2	71.2

Table 1: Balanced Accuracy and F1-Macro of hallucination detection methods across four datasets. The final two columns report the simple average across the four datasets. We note that certain models marked with an asterisk (*) were used to select articles for the adversarially challenging FaithBench dataset. Consequently, these models, including HHEM, may perform worse on FaithBench in summary-wise classification than they otherwise would.

For FaithBench, we assign each summary the most severe hallucination label given by a majority of the annotators. We evaluate using summaries labelled either Unwanted or Consistent, excluding Benign and Questionable cases due to their more ambiguous nature. This slightly differs from the original FaithBench evaluation, which pooled the worst label across all annotators for each summary and combined Benign cases with Consistent ones, while combining Unwanted cases with Questionable ones for the binary classification problem.

For AggreFact, we evaluate on the SOTA subset, which involves annotated summaries generated by fine-tuned T5 (Raffel et al., 2020), BART (Lewis et al., 2020a), and PEGASUS (Zhang et al., 2020) models. For RAGTruth, we evaluate on the annotated summaries. Lastly, for TofuEval, we evaluate on summaries generated using articles from the MeetingBank dataset (Hu et al., 2023).

6.2 Existing Hallucination Detectors

Table 1 compares the effectiveness of fine-tuned hallucination detectors and zero-shot LLM-based methods across various datasets. We evaluate Vectara's HHEM models alongside AlignScore, MiniCheck, including Bespoke-MiniCheck (Bespoke, 2024), and TrueTeacher. We also include current LLMs, used in a zero-shot setting, such as GPT-4o and o3-mini-high, as well as open-source models Qwen2.5 (7B and 72B), Llama-3.1 (8B), and Llama-3.3 (70B). The o3-mini model, in particular, excels in reasoning tasks.

Classification methods are separated into claim-wise and summary-wise classification. Claim-wise evaluation involves decomposing sentences from summaries into individual claims using Llama-3.3 (70B) and a similar prompt from Tang et al. (2024a), while summary-wise methods assess the entire summary at once.

Method	# Params	FaithBench Acc (%)	FaithBench F1 (%)
FACTS Grounding Prompt			
GPT-4o	?	65.9	56.2
o3-mini-high	?	68.8	60.7
Luo et al. Prompt			
GPT-4o	?	62.5	50.6
o3-mini-high	?	63.3	49.8
FaithJudge Prompting			
Qwen-2.5	7B	71.9	66.6
Qwen-2.5	72B	73.2	73.0
Llama-3.1	8B	60.8	61.0
Llama-3.3	70B	77.5	77.8
GPT-4o	?	79.5	81.1
o3-mini-high	?	**84.0**	**82.1**
Majority Vote (Qwen 72B, Llama 70B, GPT-4o)		80.7	81.3

Table 2: Balanced Accuracy and F1-Macro for FaithJudge on FaithBench using different LLM judges. When combining models, we apply majority voting and break ties by defaulting to a classification of Inconsistent.

For LLM-based detection, we test three prompts: (1) the RAGAS prompt (Es et al., 2024), which verifies lists of claims instead of entire responses, (2) the FACTS Grounding JSON prompt, shown to be the most effective of the prompts tested in Jacovi et al. (2025) for GPT-4o, and (3) the CoT-based prompt from Luo et al. (2023). We modify prompts slightly as needed for clearer final outputs and to specifically evaluate summaries.

Table 1 shows that, similar to findings in previous work, hallucination detection remains challenging. Zero-shot classification using GPT-4o and o3-mini-high tends to perform best, both using summary-wise classification with either the FACTS Grounding JSON prompt or the Luo et al. (2023) prompt. However, their average effectiveness remains modest, with balanced accuracy below 78% and F1-macro below 72%. Considering FaithBench, the highest balanced accuracy is achieved by o3-mini-high at 68.8% while the highest F1-macro of 63.7% is achieved by the HHEM model when considering claim-wise classification.

The table illustrates improved effectiveness with increased model size: larger open-source models generally outperform smaller ones, and GPT-4o and o3-mini-high achieve the highest overall effectiveness. However, although HHEM-2.1-open is the smallest model tested, it performs strongly, outperforming several larger models. Among the fine-tuned models, only the 7B-parameter MiniCheck achieves higher average scores for summary-wise classification, while both MiniCheck variants outperform it in claim-wise classification.

Binary Classification

Gold Truth	Consistent	Inconsistent
Unwanted	74	322
Questionable	29	38
Benign	50	34
Consistent	176	27

Ternary Classification

Gold Truth	Consistent	Benign	Unwanted
Unwanted	84	18	294
Questionable	28	13	26
Benign	51	10	23
Consistent	179	4	20

Table 3: Confusion matrices for FaithJudge prompted for classification on FaithBench summaries.

Overall, fine-tuned models can achieve stronger scores than smaller prompted LLMs, but the largest LLMs typically yield the best results, even while being zero-shot methods. This suggests room for further improvement with LLM-based classification by learning from hallucination labels. Regardless, the examined methods demonstrate modest effectiveness in general, with particularly weak effectiveness on FaithBench, which captures a diverse set of LLM summaries but is designed to be challenging for hallucination detection models.

6.3 Evaluating FaithJudge

Table 2 presents the effectiveness of FaithJudge on FaithBench using various LLMs. The highest effectiveness is achieved using the o3-mini-high judge, reaching a balanced accuracy of 84% and an F1-macro of 82.1%, allowing for much higher agreement with human annotation on FaithBench

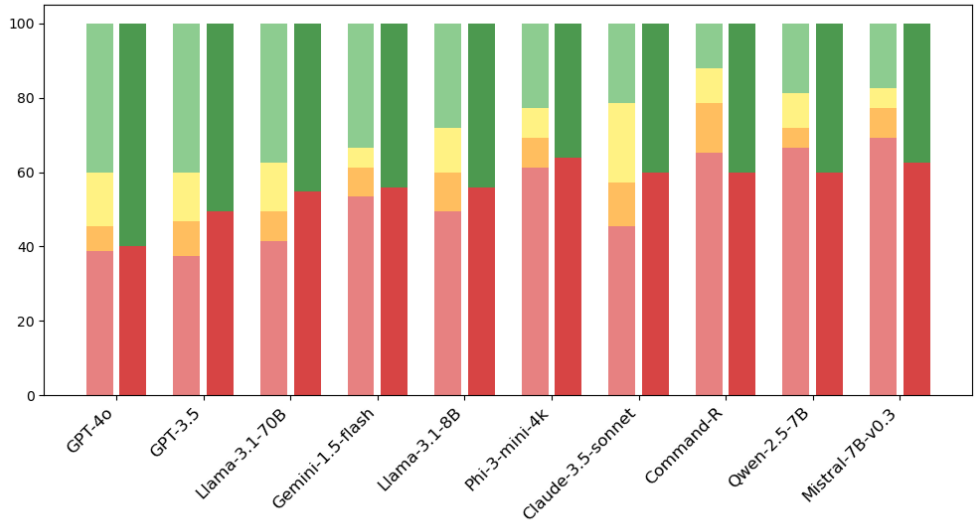

Figure 1: Proportion of summary FaithBench labels (left) and FaithJudge predictions (right) across models. For FaithBench labels, red indicates Unwanted, orange indicates Questionable, yellow indicates Benign, while green indicates Consistent. For FaithJudge predictions, red indicates Hallucinated, and green indicates Consistent summaries. Each bar shows the proportion of summaries falling into each category.

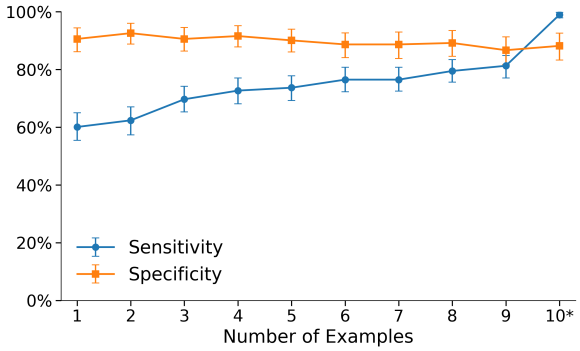

Figure 2: Sensitivity and specificity with FaithJudge as the number of examples in the prompt is increased. We place an asterisk (*) next to the 10 because, in this case, FaithJudge is shown annotations for the summary it is evaluating.

than the existing hallucination detection methods discussed. Although the effectiveness of FaithJudge is not perfect, this may be partly explained by disagreements in human annotation. While human annotation is the gold standard, the FaithBench authors (Bao et al., 2025) noted imperfect inter-annotator agreement in general and low inter-annotator agreement on more ambiguous and challenging to identify Benign and Questionable hallucination labels.

Effectiveness generally improves with increasing model size. We also tested an ensemble approach, but found that combining predictions from multiple models, including Qwen2.5 (72B), Llama-3.3 (70B), and GPT-4o with a majority vote did not outperform o3-mini-high alone. Therefore, we adopt the o3-mini-high judge as the standard for FaithJudge, with the possibility of using a stronger LLM judge in the future.

Figure 1 displays the distribution of FaithJudge predictions across LLMs. While effective, FaithJudge with o3-mini-high tends to underpredict hallucinations. This is evident for Command-R, Mistral, and Qwen, where fewer summaries were flagged as hallucinated compared to the number labelled Unwanted by annotators in FaithBench.

Table 3 presents confusion matrices for binary and ternary (including Benign labels) hallucination classification using FaithJudge. We observe that Benign summaries are difficult for FaithJudge to classify correctly. In the ternary setting, FaithJudge often misclassifies Benign summaries, generally labelling them as Consistent. Similarly, Questionable summaries are classified unreliably, though this aligns with expectations. For simplicity, we only employ FaithJudge for binary classification.

Figure 2 shows the sensitivity and specificity of FaithJudge as the number of annotated examples provided increases. Specificity remains consistently high, though slightly decreasing as more examples are given, while sensitivity notably improves as the number of examples increases. This

Dataset	FACTS Grounding Prompt		FaithJudge Prompt	
	F1-Macro	Balanced Accuracy	F1-Macro	Balanced Accuracy
RAGTruth-Data-to-text	77.1	75.1	86.3	85.1
RAGTruth-QA	76.9	81.6	83.4	85.4
RAGTruth-Summary	73.6	80.3	80.2	84.9
FaithBench-Summary	54.3	65.2	70.8	77.6

Table 4: Comparison between the FACTS Grounding zero-shot prompting approach (JSON Prompt) and the FaithJudge prompting approach on the subsets of data used in our leaderboard. In all cases we use an o3-mini-high LLM judge. For FaithJudge, we prompt the judge to evaluate LLM responses by providing the responses from the other LLMs in the dataset with their corresponding annotations. For FaithBench, we evaluate using all summaries, treating summaries labelled as Questionable or Benign as inconsistent summaries.

indicates that providing more annotated examples leads FaithJudge to predict hallucinated cases more often and better identify hallucinations.

7 Adding More Evaluation Tasks

While FaithBench provides hallucination annotations across 10 different LLMs, it is limited to evaluating summaries only. To broaden the scope of FaithJudge beyond summarization, we incorporate annotated responses from the RAGTruth dataset (Niu et al., 2024). RAGTruth includes three types of tasks: summarization, question-answering, and a data-to-text generation task that requires generating an overview of a business from JSON data sourced from the Yelp Open Dataset (Yelp, 2021). RAGTruth contains human-annotated hallucination labels for responses generated by six different LLMs: GPT-3.5, GPT-4 (OpenAI, 2023), Llama-2 (7B, 13B, and 70B) (Touvron et al., 2023), and Mistral-7B (Jiang et al., 2023).

For each RAGTruth task, we take up to 150 sources (articles for summarization, queries and passages for question-answering, and JSON data for data-to-text) with their corresponding annotated responses primarily from the test set and supplemented by the dev set where necessary. We remove sources where none of the LLM responses have a hallucination annotation.

Table 4 compares the effectiveness of FaithJudge against the zero-shot FACTS Grounding JSON prompt, which was previously shown to be an effective prompt in Jacovi et al. (2025), on the FaithBench and RAGTruth subsets used in our leaderboard. In each setting, FaithJudge achieves stronger agreement with human hallucination annotations, highlighting its strength across tasks beyond summarization.

8 Conclusion

In this paper, we presented our efforts at Vectara in evaluating and benchmarking hallucinations in RAG, discussing and building on our established hallucination leaderboard, and proposing FaithJudge. We identified effectiveness limitations in existing hallucination detection methods, including our own HHEM model. To address these challenges, we proposed FaithJudge, an approach that leverages human hallucination annotations to enhance automated hallucination detection, achieving greater effectiveness, but requiring annotations from summaries of the same articles.

Our leaderboard is live on the FaithJudge GitHub, and we share some of these results in Appendix C, providing a framework for more accurate faithfulness evaluation across diverse models and RAG tasks. We plan to continue to update our leaderboard to evaluate new models and to use improved LLM judges.

Limitations

There are some limitations to our evaluation methodology. First, our evaluation focuses exclusively on faithfulness and does not address the overall quality or helpfulness of summaries and answers. Though summary and answer quality are important in RAG applications, we consider this evaluation largely orthogonal to faithfulness.

One issue to consider is that an extractive summarizer or an LLM that simply copies parts of the article or the entire article in its response would avoid hallucinations. Nonetheless, we maintain that evaluating LLMs through hallucinations in generated summaries is promising because these hallucinations remain persistent in current LLMs.

Finally, while the o3-mini-high judge demonstrates strong effectiveness, there remains room

for enhancing accuracy and agreement with human annotators. We hope that as LLMs continue to improve, replacing o3-mini-high in FaithJudge may allow for more accurate and reliable evaluation.

Acknowledgements

We respectfully acknowledge the late Simon Mark Hughes, who led the development of the original HHEM model and Vectara's original hallucination leaderboard. His contributions laid important groundwork for Vectara's ongoing research and continue to leave a lasting influence on our work.

References

Forrest Sheng Bao, Miaoran Li, Renyi Qu, Ge Luo, Erana Wan, Yujia Tang, Weisi Fan, Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, Mike Qi, Ruixuan Tu, Chenyu Xu, Matthew Gonzales, Ofer Mendelevitch, and Amin Ahmad. 2025. FaithBench: A diverse hallucination benchmark for summarization by Modern LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 448–461, Albuquerque, New Mexico. Association for Computational Linguistics.

Bespoke. 2024. Bespoke-MiniCheck-7B.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

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

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Galileo. 2023. LLM Hallucination Index. https://galileo.ai/blog/hallucination-index.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like Summarization Evaluation with ChatGPT. *arXiv:2304.02554*.

Google Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 Herd of Models. *arXiv:2407.21783*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, Toronto, Canada. Association for Computational Linguistics.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What Have We Achieved on Text Summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2).

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, et al. 2025. The FACTS Grounding Leaderboard: Benchmarking LLMs' Ability to Ground Responses to Long-Form Input. *arXiv:2501.03200*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12).

Qi Jia, Siyu Ren, Yizhu Liu, and Kenny Zhu. 2023. Zero-shot Faithfulness Evaluation for Text Summarization with Foundation Language Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11017–11031, Singapore. Association for Computational Linguistics.

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

Reena Kandyala, Sumanth Phani C Raghavendra, and Saraswathi T Rajasekharan. 2010. Xylene: An overview of its health hazards and preventive measures. *Journal of oral and maxillofacial pathology*, 14(1):1–5.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. ChatGPT as a Factual Inconsistency Evaluator for Text Summarization. *arXiv:2303.15621*.

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.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows". *arXiv:2410.03727*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

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

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2025. The Great Nugget Recall: Automating Fact Extraction and RAG

Evaluation with Large Language Models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 180–190, New York, NY, USA. Association for Computing Machinery.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding Factual Errors in Summarization: Errors, Summarizers, Datasets, Error Detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024b. TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models can Accurately Predict Searcher Preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1930–1940, New York, NY, USA. Association for Computing Machinery.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. WizardLM: Empowering large pre-trained language models to follow complex instructions. *arXiv:2304.12244*.

Yelp. 2021. Yelp open dataset. https://www.yelp.com/dataset. Accessed: 2023-11-03.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Appendix

A Summaries with FaithJudge Verdicts

Figure 3 shows examples of two different LLM summaries for a news source from FaithBench. We generally observe that hallucinations can often be subtle and require great effort to identify. For example, while the source identifies Paraxylene as a "reportedly carcinogenic chemical", Grok-3 drops the "reportedly" modifier and describes Paraxylene as a "carcinogenic chemical". This removed detail is pertinent because it changes the information from the source. Notably, to the best of our understanding, there is not enough evidence to say that Paraxylene is carcinogenic in humans or in animals (Kandyala et al., 2010).

B Judge Bias

The FACTS Grounding leaderboard (Jacovi et al., 2025) uses three different LLM judges to mitigate bias arising from any single judge favoring its own outputs. Inspired by this, we analyze judge bias using Tables 5 and 6, which evaluate the impact of using different judges across all subsets included in our leaderboard.

Table 5 reports the effectiveness of three different LLMs when used as judges. The table shows that o3-mini-high remains a relatively effective LLM for FaithJudge, often scoring the highest. The table also shows that while using multiple judges can improve effectiveness further, in some cases, individual LLMs can score higher than the majority vote approach between the three LLMs. For example, o3-mini-high scores higher than the ensembling approach when evaluating on the RAGTruth QA subset.

Table 6 explores how each judge model ranks the other LLMs. Interestingly, o3-mini-high and llama-4-maverick both rank gemini-2.0-flash as having the fewest hallucinated responses, while gemini-2.0-flash ranks itself second to o3-mini-high, with only a small difference in hallucinated response counts (29 vs. 31).

While using multiple judges might enhance robustness and reduce individual model bias, we currently rely on a single judge to reduce computational costs. As newer and stronger LLMs become available, we plan to update FaithJudge by substituting the current o3-mini-high judge model with a more effective one.

C FaithJudge Rankings

Table 7 presents FaithJudge rankings for a small set of LLMs. In addition to detecting hallucinations, we also prompt FaithJudge to flag responses that are invalid, for example, when a model fails to meaningfully summarize an article. For simplicity, since these cases are rare, we count these as hallucinated responses. Models are ranked based on their overall hallucination rate, calculated as the total number of hallucinated or invalid responses across all four evaluation subsets. We plan to continue evaluating LLMs using FaithJudge. Please see the current leaderboard, with many more LLMs evaluated, at `https://github.com/vectara/FaithJudge`.

SOURCE

hong kong (cnn) six people were hurt after an explosion at a controversial chemical plant in china 's southeastern fujian province sparked a huge fire , provincial authorities told state media. the plant , located in zhangzhou city , produces paraxylene (px) , a reportedly carcinogenic chemical used in the production of polyester films and fabrics . the blast occurred at an oil storage facility monday night after an oil leak , though local media has not reported any toxic chemical spill . five out of six people were injured by broken glass and have been sent to the hospital for treatment , xinhua news agency reported . more than 600 firefighters were sent to battle the blaze and the fire is now under control , the zhangzhou fire department wrote on their official microblogging account tuesday morning . residents living close to the plant had heard the explosion and took to weibo to post photos of the fire . one user wrote that he heard a loud blast and felt slight tremors . the plant was hit by another explosion in july 2013 , although there were no reports of casualties or toxic leaks at the time . though demonstrations are illegal in china , the construction of px plants has sparked protests , which have occasionally turned violent , in several cities in recent years . the zhangzhou plant was slated for xiamen – a densely populated city in the southeast of the country . however , it provoked an angry backlash in 2007 due to pollution concerns and prompted the local government to relocate the factory to its current , more remote location .

Gemini-2.5-Pro **Verdict:** Inconsistent

Summary

An explosion at a chemical plant in Zhangzhou, China, injured six people and caused a large fire. The plant produces paraxylene (PX), a reportedly carcinogenic chemical. The blast occurred on a Monday night at an oil storage facility after an oil leak, though no toxic chemical spill has been reported. Five of the six people were injured by broken glass. Over 600 firefighters were sent to battle the blaze, which is now under control. The plant had a previous explosion in July 2013 and was relocated to its current location from Xiamen in 2007 due to public protests over pollution concerns.

FaithJudge Response

The summary accurately reflects almost all details from the article. It correctly states that the explosion injured six people, that five of them were injured by broken glass, that the blast occurred at an oil storage facility on Monday night after an oil leak, that no toxic chemical spill was reported, that over 600 firefighters controlled the fire, and that there was a previous explosion in July 2013.
However, the summary's last sentence introduces a key hallucination. Instead of stating that the plant "was slated for Xiamen" and that it "provoked an angry backlash in 2007 due to pollution concerns" (which then led the government to relocate the factory), the summary claims it "was relocated to its current location from Xiamen in 2007 due to public protests over pollution concerns." This rephrasing mischaracterizes the source in two ways:
1. It implies the plant was actually located in Xiamen before being moved, rather than originally being planned there.
2. It changes "an angry backlash" to "public protests," which is not how the source describes the reaction.
Because these details are not supported by the source text, the summary contains hallucinations.
Final classification: Inconsistent

Grok-3 **Verdict:** Inconsistent

Summary

An explosion at a chemical plant producing paraxylene (PX), a carcinogenic chemical used in polyester production, in Zhangzhou city, Fujian province, China, injured six people, five of whom were hurt by broken glass and hospitalized. The blast, caused by an oil leak at an oil storage facility, occurred on Monday night and led to a massive fire, which was controlled by over 600 firefighters by Tuesday morning. No toxic chemical spill was reported. Local residents heard the explosion, felt tremors, and shared images of the fire on Weibo. This plant, previously the site of a 2013 explosion with no casualties or leaks, was relocated from Xiamen to its current remote location due to pollution concerns and public backlash in 2007. PX plants have been controversial in China, sparking protests, sometimes violent, despite demonstrations being illegal.

FaithJudge Response

This summary introduces several details that do not exactly match the source. For example, it describes PX as "a carcinogenic chemical" instead of "a reportedly carcinogenic chemical," changing the certainty level. It also states that the plant "was relocated from Xiamen" rather than noting that it was originally slated for Xiamen before the backlash prompted relocation. In addition, the summary phrases the 2013 explosion as having "no casualties or leaks" rather than saying there were "no reports" of casualties or toxic leaks. These deviations, even if minor, represent hallucinated details.
Final classification: Inconsistent

Figure 3: Two LLM summaries of a news source from FaithBench with FaithJudge hallucination judgements. The red text indicates portions of the summaries that contain hallucinations.

Dataset	Judge Model	F1-Macro	Balanced Accuracy
RAGTruth (Data-to-text)	o3-mini-high	86.3	85.1
	gemini-2.0-flash	83.6	84.0
	llama-4-maverick	82.1	80.6
	Majority Vote	**86.4**	**85.8**
RAGTruth (QA)	o3-mini-high	**83.4**	**85.4**
	gemini-2.0-flash	81.8	84.2
	llama-4-maverick	77.5	81.2
	Majority Vote	81.0	83.8
RAGTruth (Summary)	o3-mini-high	80.2	84.9
	gemini-2.0-flash	83.6	82.7
	llama-4-maverick	78.0	83.7
	Majority Vote	**84.6**	**88.0**
FaithBench (Summary)	o3-mini-high	70.8	77.6
	gemini-2.0-flash	66.1	75.5
	llama-4-maverick	**74.7**	76.9
	Majority Vote	72.4	**79.1**

Table 5: Evaluation results for three models and an ensemble approach on the subsets of data used in our leaderboard. Here, for FaithBench, we evaluate using all summaries, treating summaries labelled as Questionable or Benign as inconsistent summaries.

	Judged by o3-mini-high		Judged by gemini-2.0-flash		Judged by llama-4-maverick	
Evaluated Model	Hallucinated Responses	Rank	Hallucinated Responses	Rank	Hallucinated Responses	Rank
gemini-2.0-flash	52	1	31	2	71	1
o3-mini-high	64	2	29	1	94	2
llama-4-maverick	105	3	72	3	110	3

Table 6: Total number of hallucinated responses per evaluated model, as judged by each model. Rankings indicate relative effectiveness in terms of hallucination frequency, from least to most.

Rank	Model	Overall Hallucination Rate	FaithBench (Summary)	RAGTruth (Summary)	RAGTruth (QA)	RAGTruth (Data-to-Text)
1	gemini-2.5-pro	6.65%	18/72	7/150	2/139	7/150
2	gemini-2.0-flash	10.18%	21/72	10/150	1/139	20/150
3	gpt-4.5-preview	11.94%	27/72	15/150	7/139	12/150
4	o3-mini-high	12.52%	25/72	12/150	9/139	18/150
5	gpt-3.5-turbo	14.87%	32/72	13/150	8/139	23/150
6	gpt-4o	15.85%	29/72	15/150	7/139	30/150
7	claude-3.7-sonnet	16.05%	28/72	22/150	13/139	19/150
8	llama-3.3-70b	16.44%	32/72	13/150	6/139	33/150
9	phi-4	17.03%	32/72	12/150	6/139	37/150
10	mistral-small-24b	17.03%	31/72	15/150	14/139	27/150
11	llama-4-maverick	20.55%	37/72	20/150	13/139	35/150
12	llama-3.1-8b	28.38%	32/72	19/150	17/139	77/150

Table 7: FaithJudge rankings for 12 LLMs, based on the number of hallucinated responses across four evaluation subsets: article summarization from FaithBench and RAGTruth, as well as question answering and data-to-text writing from RAGTruth.