Benchmarking LLM Faithfulness in RAG with Evolving Leaderboards

Manveer Singh Tamber, Forrest Sheng Bao, Chenyu Xu, Ge Luo, Suleman Kazi, Minseok Bae, Miaoran Li, Ofer Mendelevitch, Renyi Qu, Jimmy Lin

Benchmarking RAG Faithfulness

Goal. Reliably measure hallucinations in retrieval-augmented generation (RAG).

Our Contributions:

- **We discuss Vectara's original hallucination leaderboard**, which, since 2023, has tracked hallucination rates in summarization for LLMs (currently over 160 LLMs evaluated).
- **We introduce FaithJudge:** an LLM-as-a-judge framework that improves upon the effectiveness of the original leaderboard by reading a few *human-annotated* peer responses to the *same source*, then judging a new response for faithfulness.

Why FaithJudge? Few-shot, source-specific examples calibrate powerful LLM judges to subtle, article-specific details, yielding much higher agreement with human labels than zero-shot or fine-tuned hallucination-detection.

Hallucination Detection Effectiveness

- **Dataset**: We use FaithBench, a challenging dataset with human-annotated hallucinated summaries from 10 diverse LLMs.
- **Fine-tuned models struggle**: Hallucination detectors like HHEM (used in the original leaderboard) have poor effectiveness.
- **Zero-shot LLMs lag**: Zero-shot prompting of powerful LLMs for classification shows moderate effectiveness.
- **FaithJudge Excels**: FaithJudge with an o3-mini-high judge achieves the strongest effectiveness.

		FaithBench	
Method	**# Params**	Acc (%)	F1 (%)
Fine-Tuned Hallucination Detection Models			
HHEM-2.1-Open	110M	52.6	32.9
AlignScore-large	355M	50.3	26.1
Bespoke-MiniCheck	7B	55.7	47.3
Zero-Shot Classification with FACTS Grounding Prompt			
GPT-4o	?	65.9	56.2
o3-mini-high	?	68.8	60.7
Zero-Shot Classification with Luo et al. Prompt			
GPT-4o	?	62.5	50.6
o3-mini-high	?	63.3	49.8
FaithJudge Prompting			
Qwen-2.5	72B	73.2	73.0
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 1. Balanced Accuracy and F1-Macro on FaithBench using different hallucination detection methods. When combining models, we apply majority voting.

FaithJudge Predictions per Model

FaithJudge's predictions closely track human labels across a wide range of models, though it can be slightly conservative (under-flagging hallucinations in general).

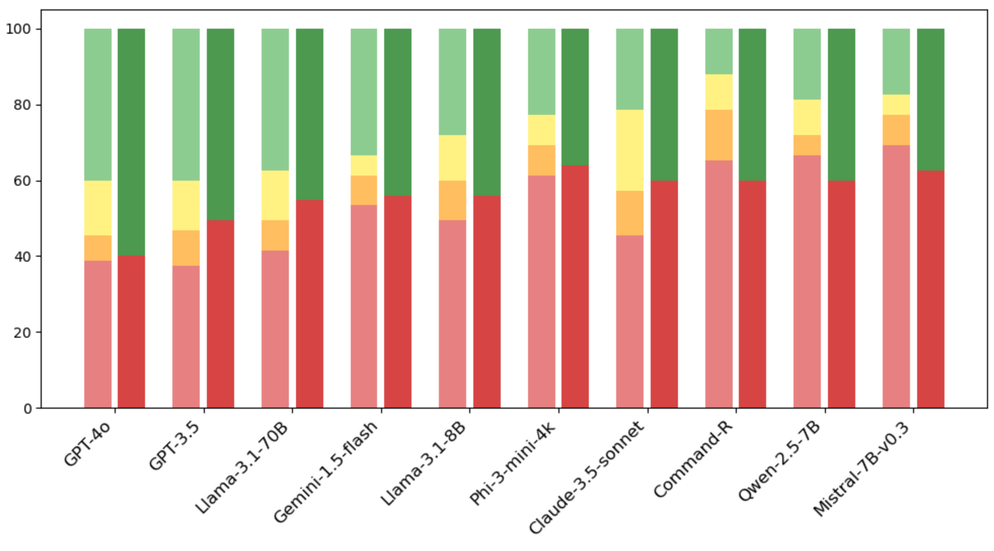

Figure 1. Proportion of summary FaithBench labels (left) and FaithJudge predictions (right) across models. For FaithBench labels, red indicates Unwanted, orange indicates Questionable, yellow indicates Benign, while green indicates Consistent. For FaithJudge predictions, red indicates Hallucinated, and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

Expanding Evaluation Tasks

While FaithBench provides hallucination annotations for 10 different LLMs, it covers summarization only. To test broader RAG capabilities, we evaluate on RAGTruth, which includes summarization, QA, and data-to-text tasks. FaithJudge consistently outperforms the zero-shot Facts Grounding approach across all tasks.

Dataset	Facts Grounding Prompt		FaithJudge Prompt	
	F1-Macro	Balanced Accuracy	F1-Macro	Balanced Accuracy
RAGTruth-Data-to-txt	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 2. 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 FaithBench, we evaluate using all summaries, treating Questionable or Benign summaries as inconsistent summaries.

FaithJudge Example

Grok-3 Summary — Verdict: Inconsistent

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 ... the plant , located in zhangzhou city , produces paraxylene (px) , a reportedly carcinogenic chemical used in the production of polyester films and fabrics . ... 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 .

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, ... 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

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.

Figure 2. Truncated Grok-3 summary judged by FaithJudge. Red highlights indicate hallucinated/unsupported details. "..." indicates text skipped for conciseness.

Hallucinations are often subtle and difficult to identify. The model incorrectly changes "reportedly carcinogenic" to "carcinogenic," a small but critical factual error. To our understanding, there is not enough evidence to say that paraxylene is carcinogenic in humans or in animals.

FaithJudge Effectiveness with # Examples

What changes with more examples? As we increase the number of annotated peer examples from the *same source*, **sensitivity** (recall of Unwanted) increases while **specificity** (recall of Consistent) stays high with a minor dip. **Takeaway:** FaithJudge becomes more willing to flag hallucinations when it has seen more source-specific mistakes.

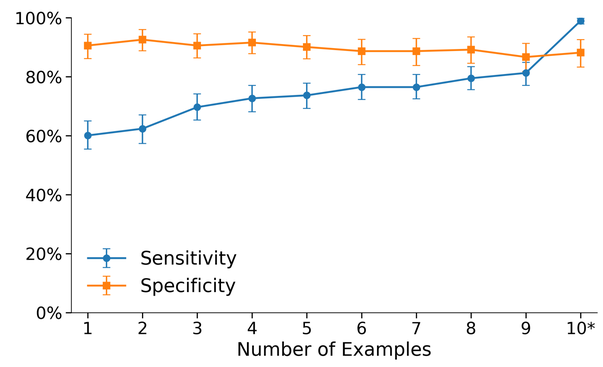

Figure 3. Sensitivity and specificity with FaithJudge as the number of examples in the prompt is increased. We place an asterisk (*) next to the 10 because, in this case, FaithJudge is shown annotations for the summary it is evaluating.

Selected Models from the FaithJudge Leaderboard

Model	Hallucination Rate
Gemini-2.5-Flash	6.26%
Gemini-2.5-Pro	6.65%
R1-0528	9.78%
GPT-4.1-2025-04-14	11.94%
GPT-4.5-preview-2025-02-27	11.94%
o3-mini-high-2025-01-31	12.52%
Grok-3	15.26%
GPT-4o-2024-11-20	15.85%
Claude-3-7-Sonnet-20250219	16.05%
Llama-3.3-70B-Instruct	16.44%
Mistral-Small-24B-Instruct-2501	17.03%
Qwen2.5-72B-Instruct	20.74%

Table 3. Hallucination rates for selected models from the FaithJudge leaderboard.

Try it Out!

Our FaithJudge leaderboard is currently live! Also, check out our paper!

https://github.com/vectara/FaithJudge