FaithBench: A Diverse Hallucination Benchmark for Summarization by Modern LLMs

Forrest Sheng Bao*[1], Miaoran Li*[1,2], Renyi Qu[1], Ge Luo[1], Erana Wan[3], Yujia Tang[4], Weisi Fan[2], Manveer Singh Tamber [5], Suleman Kazi[1], Vivek Sourabh[1], Mike Qi[6], Ruixuan Tu[6,7], Chenyu Xu[2], Matthew Gonzales[1], Ofer Mendelevitch[1], Amin Ahmad[1]

[1]Vectara, Inc. Palo Alto, CA [6]Funix.io, Iowa City, IA [2]Iowa State University, Ames, IA
[3]Univ. of Southern California, Los Angeles, CA [4]Entropy Technologies, Melbourne, Australia
[5]University of Waterloo, Waterloo, ON [7]University of Wisconsin–Madison, Madison, WI

Correspondence: {forrest.bao, limiaoran.lm, amin.ahmad}gmail.com

Abstract

Summarization is one of the most common tasks performed by large language models (LLMs), especially in applications like Retrieval-Augmented Generation (RAG). However, existing evaluations of hallucinations in LLM-generated summaries, and evaluations of hallucination detection models both suffer from a lack of diversity and recency in the LLM and LLM families considered. This paper introduces FaithBench, a summarization hallucination benchmark comprising challenging hallucinations made by 10 modern LLMs from 8 different families, with ground truth annotations by human experts. "Challenging" here means summaries on which popular, state-of-the-art hallucination detection models, including GPT-4o-as-a-judge, disagreed on. Our results show GPT-4o and GPT-3.5-Turbo produce the least hallucinations. However, most state-of-the-art hallucination detection models have near 50% accuracies on FaithBench, indicating lots of room for future improvement.

1 Introduction

With the increasing use of Large Language Models (LLMs) to process textual data, ensuring their trustworthiness has become a critical concern. In applications such as Retrieval Augmented Generation (RAG) (Lewis et al., 2020), LLMs are used to generate answers or summaries from textual input. When the generated text includes unsupported information, it is considered a hallucination, which can be misleading or harmful.

Understanding the state of hallucinations in LLMs is crucial but hard. Existing hallucination leaderboards, such as Vectara's Hallucination Leaderboard [*] and Galileo's Hallucination Index [*], detect hallucinations using models such as Google's TrueTeacher (Gekhman et al., 2023), Vectara's HHEM-2.1-Open (Bao et al., 2024), or even GPT series models in a zero-shot, LLM-as-a-judge fashion (Luo et al., 2023; Liu et al., 2023). These detection models are known to have an accuracy below 80% on benchmarks such as AggreFact (Tang et al., 2023) and RAGTruth (Niu et al., 2024). Moreover, existing benchmarks often rely on a narrow selection of LLMs, many of which are outdated and lack diversity across model families. If we assume LLMs hallucinate differently—due to variations in training methods, datasets, and architectures, as well as changes in behavior as models scale up—then conclusions drawn from such benchmarks are incomplete, capturing only specific types of hallucinations.

To address this gap, the industry and research community need a hallucination benchmark that includes modern LLMs across diverse model families, along with human-annotated ground truth for more reliable evaluation. This paper presents FaithBench, a summarization hallucination benchmark built on top of Vectara's Hallucination Leaderboard which is popular in the community (Hong et al., 2024; Merrer and Tredan, 2024) because it contains summaries generated by dozens of modern LLMs. We add human annotations, including justifications at the level of individual text spans, to summaries from 10 LLMs belonging to 8 LLM families. To make the best use of our annotators' time, we focus on labeling challenging samples where hallucination detectors disagree the most, as obvious hallucinations can be reliably detected automatically. The majority of our annotators are experts in the field of hallucination detection, with half of them having published hallucination-related papers at major NLP conferences.

FaithBench allows us to evaluate both the hallucination rates of LLMs and the accuracy of hallucination detection models. To the best of our knowledge, this is the first evaluation of hallucina-

[*]Equal contribution to this work.

[*]https://huggingface.co/spaces/vectara/leaderboard

[*]https://www.rungalileo.io/hallucinationindex

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

tions across 10 LLMs and 8 LLM families using human-annotated ground truth. GPT-4o has the lowest hallucination rate, followed by GPT-3.5-Turbo, Gemini-1.5-Flash, and Llama-3-70B. All hallucination detectors are found to correlate poorly with human-annotated ground truth, with the best balanced accuracy and F1-macro score at 62% and 57% respectively. This highlights our limited understanding of hallucinations and the challenges ahead.

We hope that FaithBench can catalyze research into detecting and mitigating hallucinations in LLMs. In contrast with existing benchmarks, FaithBench 1) covers a wide array of LLM families and diverse hallucination characteristics, 2) factors the subjectivity of hallucination perception, by expanding binary consistent vs. unfaithful labels to include two new "gray-area" labels: "questionable" and "benign", 3) includes only challenging hallucination samples. The repo is `https://github.com/vectara/FaithBench`

2 The Benchmark

2.1 Definition of hallucinations

The word "hallucinating" has two meanings in the context of LLMs. It could mean either "non-factual" (Mishra et al., 2024; Ji et al., 2024, 2023; Deng et al., 2024; Li et al., 2024; Chen et al., 2023), when the LLM-generated text is not supported by the world knowledge, or "unfaithful" or "inconsistent" (Tang et al., 2023; Niu et al., 2024; Tang et al., 2024b) when the LLM-generated text does not adhere to its input. This paper focuses on the latter case, wherein an LLM is expected to fulfill a task, often generating a summary or answering a question, based on a given passage or reference. Such scenarios are common in applications such as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). By this definition, a statement can be simultaneously factual yet unfaithful. For example, if the passage states that "water has a smell", then the statement "water is odorless" is a hallucination despite being factual according to common world knowledge.

2.2 Hallucination Taxonomy

While hallucinations draw a great deal of attention in NLP because they are often harmful and misleading, recent research argues that not all hallucinations are necessarily bad (Ramprasad et al., 2024). In fact, users often value the enrichment LLMs provide through reasoning, creativity, and factual knowledge. Hence, we separate hallucinations into *benign* and *unwanted* categories.

Given that some hallucinations are disputed even among human annotators, this paper categorizes hallucinations into three types:

- **Questionable**: not clearly a hallucination, classification may differ depending on whom you ask.
- **Benign**: clearly a hallucination, but supported by world knowledge, common sense, or logical reasoning, such that a reader finds it acceptable or welcomed.
- **Unwanted**: A clear hallucination that is not benign. This category is further subdivided into two categories:
  - **Intrinsic**: Contradicted by the passage, either in part or in whole.
  - **Extrinsic**: neither supported by the passage, nor inferrable from it, nor factual.

2.3 Data Sampling

Sourcing the data We utilize Vectara's hallucination leaderboard, which already contains summaries generated by dozens of LLMs and is frequently cited in the community. In the leaderboard dataset, the passages for summarization come from various Natural Language Inference (NLI), fact-checking, or summarization datasets. Some passages are specifically crafted to 'trick' LLMs into hallucinating (Appendix G), such as by combining information about two unrelated individuals in the same profession within one passage to induce a coreference error. A *sample* is defined as a pair consisting of a source passage and an LLM-generated summary.

Filtering samples by LLM To balance annotator effort with our goal of LLM diversity, we restrict the benchmark to eight of the most anecdotally popular LLM families: GPT, Llama, Gemini, Mistral, Phi, Claude, Command-R, and Qwen. For each family, we then selected the smallest version in its latest generation. The exceptions are the GPT and Llama series from which we select two each. For GPT, we select GPT-4o and GPT-3.5-Turbo as they are cost efficient. For Llama, we select Llama-3.1-70B and -8B in order to assess the impact of model size. Our preference towards small and affordable models aims to maximize the value of our work

to the community as these models are used more widely than their larger counterparts.

Filtering samples by consensus of detectors Human annotation of obvious hallucinations is of limited value, as they can be easily detected by automatic systems; the real value lies in annotating challenging samples where popular detection models disagree. This will provide a valuable calibration for the community, highlighting areas where detectors struggle and guiding future improvements. Based on their popularity (Mickus et al., 2024; Sansford et al., 2024), the following hallucination detectors are chosen to identify challenging samples: Google's True-NLI (Honovich et al., 2022) and TrueTeacher (Gekhman et al., 2023), Vectara's HHEM-2.1-Open (Bao et al., 2024), and GPT-{4o, 3.5-Turbo}-as-a-judge (Liu et al., 2023; Luo et al., 2023).

Sample groups In this paper, our samples are divided into groups of ten which share one common source passage but contain outputs from 10 different LLMs. This allows us to compare the performance of each LLM while controlling for the characteristics of the source text.

We then rank groups by the number of challenging summaries in each group. The top 115 groups containing at least 7 challenging summaries each are moved to the next step.

2.4 Human Annotation

Annotators The hallucination ground truth is added by 11 human annotators. The super majority of them are experts in the field of hallucination detection, with half of them having published hallucination-related papers at top-tier NLP conferences. About half of them are graduate students from three US/Canadian universities, and the other half are machine learning engineers. The diverse yet professional backgrounds of the annotators helps to ensure the quality of the annotations. Three annotators are native speakers of English. All annotators are aware that the data they created will be made open source to the public.

The pilot run A pilot run of 30 random samples pertaining to 30 different passages was conducted to ensure annotators are in agreement on the definition and categorization of hallucinations.

The pilot run revealed two issues. First, many sports-related samples required specific knowledge of European sports terminology, which posed a challenge for our annotators who are not familiar with these sports. Second, many source passages are not self-consistent due to noise introduced in their construction. Based on these observations, we visually inspected all passages and removed corresponding samples, leaving us with 800 samples.

The samples were then divided into 16 batches of 50 samples each (8 passages $\times$ 10 LLM-generated summaries). All batches were annotated by two annotators with most also having a third annotator to provide an additional opinion. In the process of post-pilot annotation, we found more samples with noisy passages including image captions or advertisements. They are then excluded from the benchmark. The final benchmark totals at 750 samples (75 passages $\times$ 10 LLMs).

Semantic-assisted cross-checking Given a text span in the summary, finding corresponding spans in the passage that support or refute it is often difficult because modern LLMs are very abstractive, limiting the benefit of exact string matching. Thus, we developed an in-browser annotation tool that highlights sentences in the passage that are semantically similar to a selected text span in the summary. With the benefit of this annotation tool, annotators are asked to select all spans in the summary that are hallucinations or suspected hallucinations. For each selected span, they are asked to assign a label (§ 2.2) and add a note explaining their reasoning. If the span is related to one in the passage, they are encouraged to link the summary span and the passage span.

3 Results

3.1 Annotation quality

Following the common practices in the field, the annotation quality is measured by inter-annotator agreement (IAA) using Krippendorff's alpha (Krippendorff, 2018) at the sample level.

Different spans in a summary maybe assigned different labels by the same annotator. To compute IAA, each sample's span-level labels are "worst-pooled" into one sample-level label using the worst label among all spans assigned by the annotator. The severity of hallucinations is ordered as: consistent (best) $\succ$ benign $\succ$ questionable $\succ$ unwanted (worst).

The IAA for the "consistent" and "unwanted" classes is 0.749. Undoubtedly, the IAA for the other two classes, "questionable" and "benign",

will be low. The IAA for tenary classification consistent vs. benign vs. unwanted, and ternary classification consistent + benign vs. questionable vs. unwanted, are 0.679 and 0.582, respectively. The much lower IAA after considering the "questionable" and "benign" labels indicates the high subjectivity on borderline hallucinations and justifies the necessity of introducing them in our benchmark.

Annotations are done in two rounds. In the first round, annotators work independently. In the second round, they discuss and resolve disagreements. Annotators are encouraged to hold their ground if they are confident in their annotations rather than being forced to converge with other annotators. IAA for the first round can be as low as 0 while the second round significantly boost the IAA. This reflects the challenge in annotating hallucinations that even experience professionals can miss them.

3.2 Ranking LLMs by Hallucinations

Figure 1 shows the distribution of "worst-pooled" (§ 3.1), sample-level labels per LLM. GPT-3.5-Turbo produces the highest percentage (38.67%) of fully consistent summaries. GPT-4o, Llama-3.1-70B and Gemini-1.5-Flash rank 2nd, 3rd, and 4th, respectively, with nearly 1/3 of the summaries produced by them are fully consistent. Claude-3.5-Sonnet produces a great amount (21.33%) of summaries that contain benign hallucinations.

Using the "worst-pooled", sample-level labels, we can compute the rate of hallucinations of LLMs and rank them (Table 1). The rankings according to FaithBench (first three columns) generally align well with the ranking in Vectara's Hallucination Leaderboard (rightmost column). It slightly differs from Galileo's Hallucination Index, which ranks Claude-3.5-Sonnet as the best proprietary LLM.

LLM	Unwanted	U+Q	U+Q+B	VHL
GPT-4o	40.00 (1)	53.33 (1)	66.67 (2)	1
GPT-3.5-Turbo	44.00 (2)	53.33 (1)	61.33 (1)	2
Llama-3.1-70B	48.00 (3)	54.67 (3)	68.00 (3)	3
Gemini-1.5-Flash	56.00 (6)	64.00 (5)	69.33 (4)	4
Llama-3.1-8B	53.33 (5)	66.67 (6)	77.33 (5)	5
Claude-3.5-Sonnet	48.00 (3)	61.33 (4)	82.67 (7)	6
Qwen2.5-7B	73.33 (10)	78.67 (9)	85.33 (9)	7
Phi-3-mini-4k	65.33 (7)	74.67 (7)	80.00 (6)	8
Command-R	68.00 (8)	84.0 (10)	92.00 (10)	9
Mistral-7B	69.33 (9)	77.33 (8)	84.00 (8)	10

Table 1: Hallucination rates (%) and LLM rankings (between parenthesis) based on three levels: Unwanted only (U), U + Questionable (U+Q), and U+Q+Benign (U+Q+B). Column VHL is the ranking of LLMs in Vectara's Hallucination Leaderboard.

Figure 2 presents, for each LLM, the ratios of unwated, questionable, and benign annotations (span-level) to all hallucination annotations. When interpreting all results above, it is important to keep in mind that they are only true for the challenging samples. It may not be true for all samples.

3.3 Ranking Hallucination Detectors

Table 2 shows the balanced accuracy (BA) and F1-Macro (F1-M) score of several hallucination detectors against the ground truth in FaithBench at the sample level. Here *a sample is hallucinated if it is unwanted or questionable*. Because of the popularity of LLM-as-a-judge, we extensively evaluated different OpenAI LLMs (GPT-4-Turbo, GPT-4o, o1-mini, and o3-mini) with two styles of prompts: non-reasoning, zero-shot (Luo et al., 2023) and chain-of-thought used in Google FACTS Grounding dataset (Jacovi et al., 2025). The two prompts are denoted as "simple zero-shot" and "FACTS CoT" in Table 2.

It turns out that 62.31% is the highest balanced accuracy for the binary classification problem where a random guess a has 50% chance to be correct, indicating the rigor of FaithBench and the need for a challenging benchmark like FaithBench in our battle against hallucinations. Reasoning-enhanced OpenAI LLMs, namely o1-mini and o3-mini, perform better than their non-reasoning counterparts, namely GPT-4-Turbo and GPT-4o.

Surprisingly, the CoT-style prompt used in FACTS (Jacovi et al., 2025) consistently underperformances the simple, zero-shot prompt used in (Luo et al., 2023) across all OpenAI LLMs (GPT-4-Turbo, GPT-4o, o1-mini, and o3-mini) in the LLM-as-a-judge fashion. Our hypothesis is that the state-of-the-art LLMs may hallucinate when reasoning (at least in the CoT fashion) and mislead themselves – although CoT is supposed to improve the reasoning capability of LLMs.

The two approaches that break down a summary into sentences or claims before hallucination detection, namely RAGAS and TruLens, achieve higher accuracy than the remaining approaches that treat the summary as a whole. RAGAS and TrueLens using GPT-4o outperforms GPT-4o-as-a-judge using the simple, zero-shot prompt (Luo et al., 2023) and the FACTS CoT prompt (Jacovi et al., 2025) by 6 to 10 percentage points.

Figure 3 presents the error distribution of hallucination detectors. For any detector, the most undetected hallucinations belonged to the "unwanted"

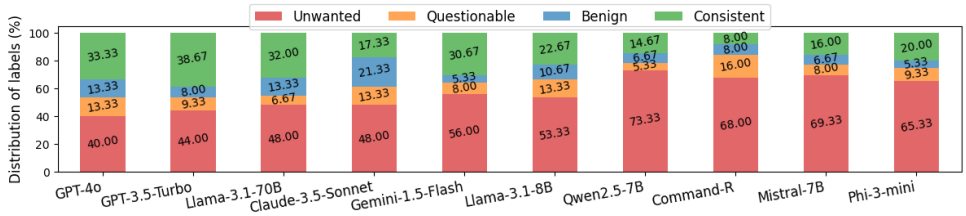

Figure 1: Sample-level distribution of annotations per "worst-pooling" (using the most severe hallucination label given by human annotators as the label of the sample) per LLM.

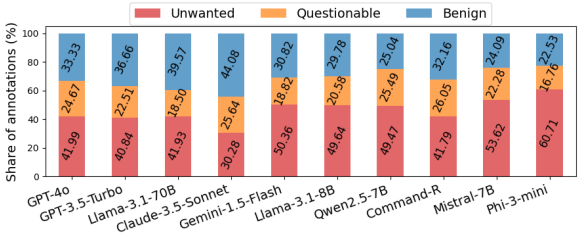

Figure 2: Span-level distribution of hallucinations by occurrence frequency, per LLM.

Hallucination Detector		BA (%)	F1-M (%)
HHEM-2.1 (Mendelevitch et al., 2024)		55.27	40.30
HHEM-2.1-Open (Bao et al., 2024)		51.98	33.03
HHEM-1		48.70	42.37
AlignScore-base (Zha et al., 2023)		51.31	44.92
AlignScore-large (Zha et al., 2023)		51.96	36.77
True-Teacher (Gekhman et al., 2023)		52.87	37.60
True-NLI (Honovich et al., 2022)		50.99	28.52
GPT-4-Turbo	w/ simple, zero-shot prompt (Luo et al., 2023)	55.96	42.16
GPT-4o		56.18	39.93
o1-mini		61.17	48.22
o3-mini		58.87	44.52
GPT-4-Turbo	w/ FACTS CoT prompt (Jacovi et al., 2025)	53.59	32.56
GPT-4o		52.19	30.35
o1-mini		58.67	45.27
o3-mini		58.18	42.44
MiniCheck-Roberta-large (Tang et al., 2024a)		52.04	51.21
MiniCheck-Deberta-large		55.21	55.19
MiniCheck-Flan-T5-large		50.14	49.17
RAGAS (Es et al., 2024)	w/ GPT-4o	62.31	57.06
TruLens (TruLens, 2024)		61.14	51.94

Table 2: Sample-level performance of hallucination detectors. The negative class is unwanted + questionable whereas the positive class is benign + consistent.

category, which is also the worst form of hallucination. This pattern indicates a universally low recall in detecting unwanted hallucinations. Specifically, for nine out of 13 detectors, over 70% of the misclassification were due to misclassifying "unwanted" hallucinations as "consistent." In contrast, this proportion is significantly lower for MiniCheck models, such as 42% for MiniCheck-Deberta-Large. Additionally, MiniCheck models exhibit a more cautious approach, enhancing recall at the cost of precision, with 24-30% of errors arising from misclassifying consistent samples as inconsistent.

4 Conclusion

This paper introduces FaithBench, a benchmark for summarization hallucinations, featuring human-annotated hallucinations in summaries generated by 10 modern LLMs across 8 different model families. To account for the subjective nature of hallucination perception, we introduced two gray-area labels—*questionable* and *benign*—in addition to the common binary labels of *consistent* and *hallucinated*. The human annotation is fine-grained at the span level and most annotations are accompanied by reasons for better explainability. With FaithBench, we are able to rank the state-of-the-art LLMs and hallucination detectors. While the ranking of LLMs largely aligns with a popular hallucination leaderboard, most state-of-the-art approaches only achieve around 50% accuracy on FaithBench. In summary, the creation and curation of FaithBench mark a crucial step in the long journey towards effectively addressing hallucinations.

Limitations

Although a primary goal of FaithBench is the diversity of hallucinations in various characteristics, as a short paper, it cannot cover a lot.

FaithBench covers only summarization. There are many other tasks where hallucination detection

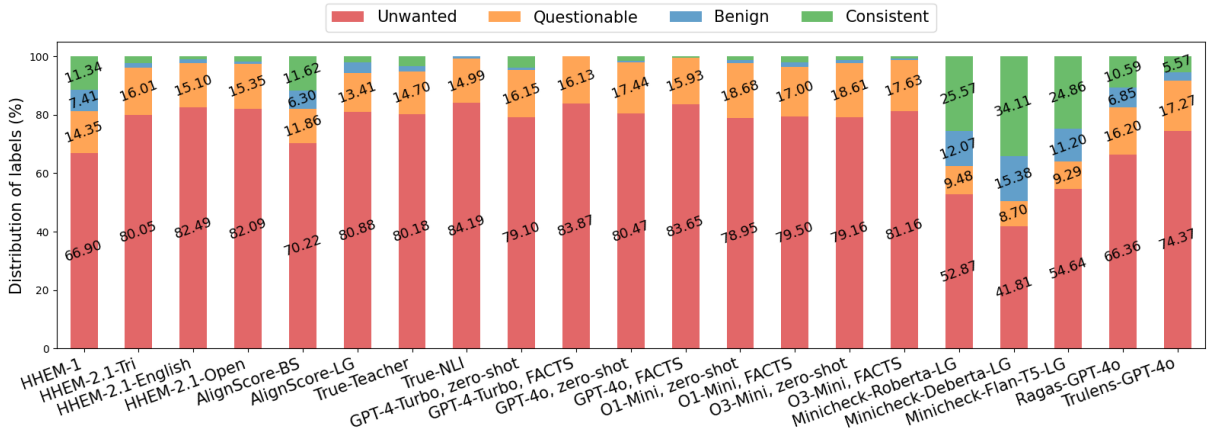

Figure 3: Error distribution of hallucination detectors. Only categories representing more than 4% are labeled in the figure.

is needed such as question answer.

Due to the composition of the foundation dataset, most passages are between 106 (1st quartile) to 380 (3rd quartile) English words in length (Appendix C). This translates to roughly 137 to 494 tokens. This means that FaithBench only measure short-context hallucinations for LLMs. We will extend it to include samples of longer contexts, such as using those in RAGTruth as the passages. But that will raise the human annotation difficulties and cost.

Due to the tremendous amount of labor needed in human annotation, we are not able to cover models of various sizes in the same family. This limits our ability to study the impact model sizes in hallucination.

The spans and reasoning collected in FaithBench are not used in evaluating LLMs and hallucination detectors.

Because FaithBench only contains challenging samples, our ranking to LLMs and hallucination detectors does not reflect their rankings on all samples. When interpreting all results above, it is important to keep this in mind.

Lastly, although FaithBench makes the effort to factor in subjectivity in labeling questionable and benign hallucinations, the inter-annotator agreements on the two gray-area hallucinations are low. We will need to develop a better taxonomy of hallucinations after taking a closer look such annotations/samples.

References

Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. 2024. HHEM-2.1-Open.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. Felm: benchmarking factuality evaluation of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 44502–44523.

Xiang Chen, Duanzheng Song, Honghao Gui, Chenxi Wang, Ningyu Zhang, Yong Jiang, Fei Huang, Chengfei Lyu, Dan Zhang, and Huajun Chen. 2024. Factchd: Benchmarking fact-conflicting hallucination detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6216–6224. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Kunquan Deng, Zeyu Huang, Chen Li, Chenghua Lin, Min Gao, and Wenge Rong. 2024. Pfme: A modular approach for fine-grained hallucination detection and editing of large language models. *Preprint*, arXiv:2407.00488.

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 R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, Clémentine Fourrier, and Pasquale Minervini. 2024. The hallucinations leaderboard - an open effort to measure hallucinations in large language models. *CoRR*, abs/2404.05904.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

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 preprint arXiv:2501.03200*.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. ANAH: Analytical annotation of hallucinations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8135–8158, Bangkok, Thailand. Association for Computational Linguistics.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

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.

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. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation. *Preprint*, arXiv:2406.07070.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization. *Preprint*, arXiv:2303.15621.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan Thomas 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.

Ofer Mendelevitch, Forrest Sheng Bao, Miaoran Li, and Rogger Luo. 2024. HHEM 2.1: A better hallucination detection model and a new leaderboard (blog post).

Erwan Le Merrer and Gilles Tredan. 2024. Llms hallucinate graphs too: a structural perspective. *Preprint*, arXiv:2409.00159.

Timothee Mickus, Elaine Zosa, Raul Vazquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. SemEval-2024 task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1979–1993, Mexico City, Mexico. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *Preprint*, arXiv:2401.06855.

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.

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.

Sanjana Ramprasad, Elisa Ferracane, and Zachary Lipton. 2024. Analyzing LLM behavior in dialogue summarization: Unveiling circumstantial hallucination trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12549–12561, Bangkok, Thailand. Association for Computational Linguistics.

Hannah Sansford, Nicholas Richardson, Hermina Maretic, and Juba Saada. 2024. Grapheval: A knowledge-graph based llm hallucination evaluation framework. In *KiL'24: Workshop on Knowledge-infused Learning co-located with 30th ACM KDD Conference*.

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.

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*. 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.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

TruLens. 2024. Moving to trulens v1: Reliable and modular logging and evaluation.

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.

Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, and Yejin Choi. 2024. Wildhallucinations: Evaluating long-form factuality in llms with real-world entity queries. *Preprint*, arXiv:2407.17468.

A Sentence-level performance of hallucination detectors

We further analyze the performance of hallucination detectors at the sentence level. For Ragas and Trulens, both frameworks first decompose the input text into claims pr statements for verification, make judgments on each unit, and then integrate these judgments into a final prediction. We use their intermediate judgments as sentence-level predictions. If a sentence in the summary is not explicitly checked by the framework, we assume it to be consistent. For other methods, we generate sentence-level inputs by first using GPT-4o to split summaries into sentences, ensuring that no sentence is excessively short (i.e., fewer than five words). If a sentence is too short, we manually merge it with its neighboring sentence. We then use regex to determine the start and end indices of each sentence. The sentence-level human labels are obtained in a manner similar to sample-level labeling. In our analysis, we use "worst-pooled" human labels as ground truth.

Table 3 presents the balanced accuracy (BA) and F1-Macro scores of hallucination detectors at the sentence level. A sentence is considered hallucinated if it is either unwanted or questionable. Compared to the sample-level results in Table 2, we observe an improvement in performance for most detectors, suggesting that detectors may be more effective with shorter inputs and can be distracted by longer inputs. However, Ragas and Trulens exhibit a significant drop in performance, indicating that while they excel at making overall judgments on summaries, they may overlook individual statements that require verification.

Figure 4 presents the sentence-level error distribution of hallucination detectors. Compared to

Hallucination Detector	BA (%)	F1-Macro (%)
HHEM-2.1 (Mendelevitch et al., 2024)	54.15	50.36
HHEM-2.1-Open (Bao et al., 2024)	54.36	50.78
HHEM-1	49.96	49.02
AlignScore-base (Zha et al., 2023)	53.30	52.77
AlignScore-large (Zha et al., 2023)	55.96	55.84
True-Teacher (Gekhman et al., 2023)	51.38	48.62
True-NLI (Honovich et al., 2022)	50.89	48.62
GPT-4-Turbo, zero-shot	53.10	51.65
GPT-4o, zero-shot	52.47	50.19
O1-Mini, zero-shot	53.54	51.73
O3-Mini, zero-shot	54.70	52.07
MiniCheck-Roberta-large (Tang et al., 2024a)	56.68	56.67
MiniCheck-Deberta-large	58.39	58.49
MiniCheck-Flan-T5-large	55.90	55.77
RAGAS w/ GPT-4o (Es et al., 2024)	49.96	46.25
TruLens w/GPT-4o (TruLens, 2024)	50.08	44.24

Table 3: Sentence-level performance of hallucination detectors.

the sample-level error distribution, we observe that detectors tend to be more cautious at the sentence level, with a higher percentage of errors arising from misclassifying non-hallucinated sentences as hallucinations. This suggests that detectors be more risk-averse when evaluating individual sentences, potentially leading to an increased tendency to flag accurate content as hallucinated.

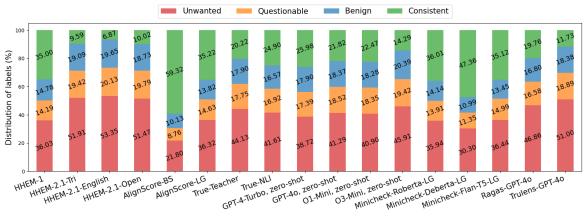

Figure 4: Sentence-level error distribution of hallucination detectors.

B Hallucinations vs. lengths

Here we study the relationship between hallucinations and passage length. When interpreting the results, please factor in the length distribution of passages (Appendix C). Points beyond 400 words are covered very sparsely.

Figure 5 shows the relationship between hallucination rates (considering only unwanted hallucinations) and the length of the passage. Contrary to the expectation that longer passages lead to more hallucinations, some models exhibit higher hallucination rates with shorter passages. Upon examining randomly sampled hallucinations for short passages, we found that LLMs often add extra information not present in the source, which is also difficult to validate even with external knowledge.

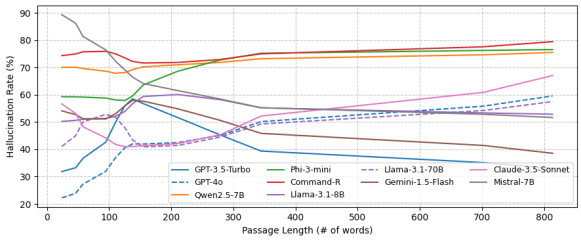

Figure 5: Hallucination rates vs. passage length

We further study the percentage of hallucination types relative to source passage length. As shown in Figure 6, most LLMs exhibit a decrease in the ratio of unwanted hallucinations as the passage length increases. The ratios of questionable and benign hallucinations show mixed trends across models, indicating that the relationship between hallucination types and passage length is inconsistent and model-specific.

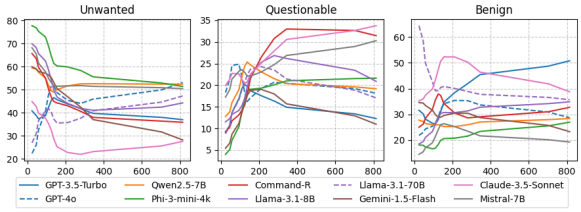

Figure 6: Ratio (%) of hallucination vs. passage length

Studying the relationship between the hallucination rates and the length of the summary is a bit hard because different LLMs yield summaries of different lengths. Despite that, we manage to get Figure 7.

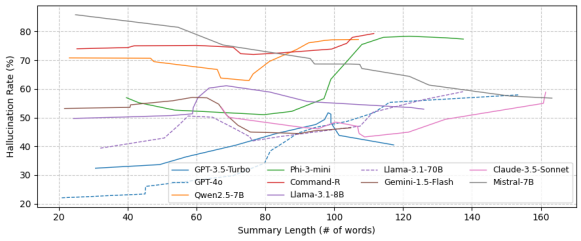

Figure 7: Hallucination rates vs. summary length

C Data Source details

The mean, median, and standard deviation of the lengths of passages are 300, 184, and 277 respectively. The 1st, 2nd, 3rd, and 4th 5-quantiles of passage lengths fall onto 87, 133, 282, 593 words.

Composition of Vectara's Hallucination Leaderboard is given in Table 4. Some samples are created with the intention to trick LLMs into hallucinating.

dataset	Percentage
XSum-Factuality (Maynez et al., 2020)	27.34
FEVER, dev (Thorne et al., 2018)	25.85
Polytope, test (Laban et al., 2022)	18.79
VitaminC, dev (Schuster et al., 2021)	11.23
SummEval, valid (Fabbri et al., 2020)	9.94
Frank, valid (Pagnoni et al., 2021)	6.86

Table 4: Composition of Vectara's Hallucination Leaderboard

D Annotator instructions and the annotation tool

Instruction to Annotators

The task is to label how faithful the output of an LLMs is to the input given to it.

In a RAG system, text retrieved based on a user query is called the "context". The context forms part of the input to an LLM to produce a summary that answers the user query.

Please select any text span in the summary that is not faithful to or supported by the context, and categorize it to one or multiple types of hallucination. If there is any text span in the context that is related to the summary span, please select it and link it with the summary span.

A faithful response can be contradictory to the world or your knowledge as long as such knowledge is in the context too. Do not confuse "faithful" with "factual".

{{Hallucination Taxonomy }}

{{Hallucination Examples }}

Annotation tool The semantic cross-checking feature of our annotation tool is given in Figure 8. Figure 9 shows that a pair of text spans, one in the passage and the other in the summary, are selected and their labels are being added in the pop-up bubble.

E Hallucination Taxonomy and examples

Short examples are:

- Questionable
  - Last August
    –> the August of last year
  - The train was late by 2 hours 45 minutes
    –> The train was late by almost 3 hours.
  - I ate a lot for lunch.
    –> Overeating causes obesity.
  - Tesla's Model S is sold for $79k.
    –> Model S is made by Tesla.
    (Common sense tells us that Tesla is not a person and thus not an owner but a manufacturer here.)
  - President Biden visited Japan today
    –> Joe Biden was in Japan today.
    (The first name of Biden is not mentioned in the passage. But we Chauvinistically assume that most people in the world know the first name of the current US president.)
  - At the University of Mississippi, about 55 percent of its undergraduates and 60 percent overall come from Mississippi, and 23 percent are minorities; international students come from 90 nations
    –> The University of Mississippi has a diverse student body.
    (This is hallucination because the passage does not assess diversity. But it is reasonable to infer. Hence, benign hallucination.)
- Unwanted
  - I ordered a pizza from downstairs.
    –> The pizza is yummy.
    (This is an extrinsic hallucination.)
  - I ate the pizza
    –> I tossed away the pizza.
    (This is an intrinsic hallucination because the summary cannot be true when the passage is also true.)
  - Goldfish weigh 1 pound and can grow up to 30 cm while koi weigh up to 2 pounds and are as long as 2 meters.
    –> Koi weigh 1 pound and can grow up to 2 meters.
    (This kind of hallucinations are often referred to as discourse hallucinations where pieces of information are stitched together wrongly.)
  - The Earth was believed flat.
    –> The Earth was flat.
  - Penguins cannot fly.
    –> No birds can fly.
  - Company X employees 50,000 people
    –> Company Y employees 50,000 programmers.

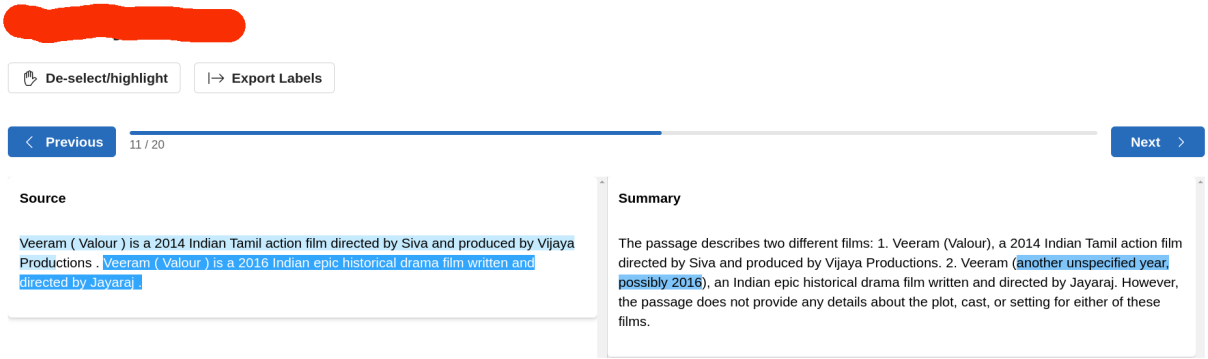

Figure 8: Semantic highlighting for easy cross-checking in our annotation tool. The selected summary span is embedded when selected. Then its dot-product distance to sentences, whose embeddings are precomputed during ingestion, in the passage are computed. Finally, sentences in the passage are highlighted with different color intensity proportional to their semantic distances.

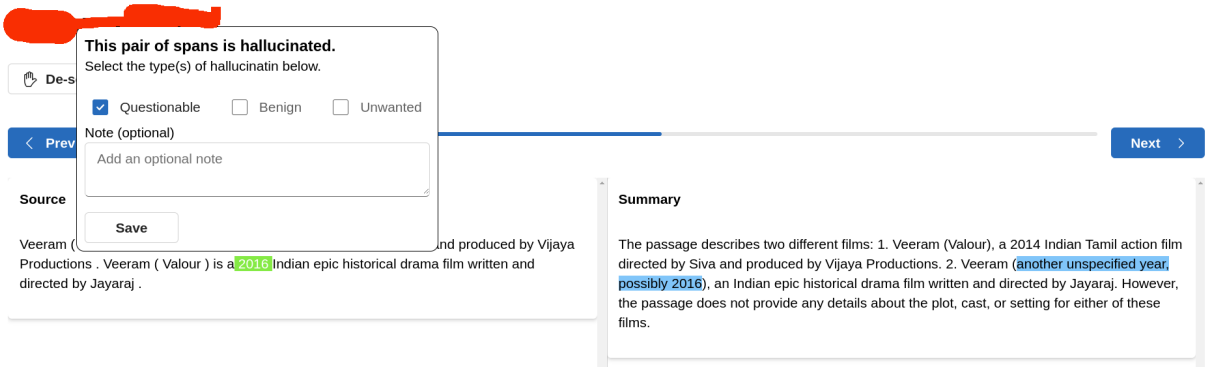

Figure 9: Annotating a pair of selected spans.

Long examples are shown in Figure 10.

F More related work

Table 5 shows the LLM families covered by different benchmarks. In all benchmarks, GPT family is covered. Llama models are also widely explored, covered in 5 benchmarks. Many of the benchmarks in Table 5 are for factuality rather than faithfulness in this paper, or do have have human ground truth.

A team from University of Edinburgh (Hong et al., 2024) evaluates LLMs' ability to serve as hallucination detectors, i.e., LLM-as-a-judge, on various tasks. The data may be human-written, LLM-generated, or rule-synthesized. In contrast, our work here focus on hallucinations in LLM-generated content. Another difference between this work and our work is that this work checks both faithfulness and factuality while our work focuses on faithfulness because we believe that LLMs's value is more in their ability to respond to variable inputs than in answering questions using knowledge parametrized, or memorized.

G Samples where LLMs hallucinate

We found many samples that LLMs (are successfully tricked to) hallucinate. Here are a couple:

- By GPT-4o:
  "Hourglass" is a song by British electronic duo Disclosure. Hourglass is singer-songwriter James Taylor's fourteenth studio album.
  –> The passage describes that "Hourglass" is a song by the British electronic duo Disclosure. This song is featured on singer-songwriter James Taylor's fourteenth studio album.
  (The passage does not say whether the song was in the album. But the LLM linked them together.)

- By Mistral-7B:
  Brad Wilk. He is best known as the drummer of the rock bands Rage Against the Machine (1991–2000; 2007–11) and Audioslave (2001–07). Wilk started his career as a drummer for Greta in 1990, and helped co-found Rage with Tom Morello and Zack de la Rocha in August 1991. As of 2016, he is playing in the band Prophets of Rage, with Commerford, Morello, Chuck D, B-Real and DJ Lord. Rage is a German heavy metal band, formed in 1984 by Peter "Peavy" Wagner.
  –> Brad Wilk is a drummer who gained prominence as a member of the rock bands Rage Against the Machine (1991-2000, 2007-2011) and Audioslave (2001-2007). He began his career with the band Greta in 1990 and co-founded Rage with Tom Morello and Zack de la Rocha in August 1991. As of 2016, he is playing in the band Prophets of Rage. Rage Against the Machine is not a German heavy metal band, as stated in the passage, but rather an American rock band. The passage incorrectly mentions that Rage is a German heavy metal band, formed in 1984 by Peter "Peavy" Wagner.
  (The LLM gets confused and thinks the info of the German band, which is factual and faithful, is a hallucination.)

- By Qwen
  The "black box" of the Su-24 jet was officially opened in Moscow on Friday in front of journalists and diplomats. Nikolai Primak, head of the Russian investigation, said flight information appeared to be missing.
  -> The black box from the Su-24 jet was opened in Moscow, revealing potentially missing flight information.

H AI assistant usage

We used AI assistants in generating analytics code and revising the paper occasionally.

Benchmark	Model Families
FELM (Chen et al., 2023)	GPT
FactCHD (Chen et al., 2024)	GPT
FavaBench (Mishra et al., 2024)	Alpacab, Vicuna, Llama2
ANAH (Ji et al., 2024)	GPT, InternLM
RAGTruth (Niu et al., 2024)	GPT, Mistral, Llama
TofuEval (Tang et al., 2024b)	GPT, Vicuna, WizardLM
HaluEval-2.0 (Li et al., 2024)	GPT, Claude, Vicuna , YuLan, Llama-2, Alpaca
HalluDial (Luo et al., 2024)	Mistral, Vicuna, Llama-2, and GPT
WildHallucination (Zhao et al., 2024)	GPT, Gemini, Claude, Command R, Perplexity, Llama, Mistral/Mixtral

Table 5: Number of LLM families covered by different benchmarks. Many of the benchmarks are for factuality rather than faithfulness in this paper, or they do have have human ground truth.

Source: The incident occurred on the A9 north of Berriedale in Caithness at about 14:00. Several other people received minor injuries in the crash. The road was closed and diversions were put in place via the A897 Helmsdale to Melvich road. Police Scotland said inquiries into the incident were ongoing.

Summary: The passage describes an incident that occurred on the A9 north of Berriedale in Caithness at around 14:00. This incident resulted in a crash, and several people received minor injuries. The road was closed, and diversions were set up via the A897 Helmsdale to Melvich road. Police Scotland is currently conducting ongoing inquiries into the incident.

Explanation: The source describes a past event, but "is currently conducting" in the summary implies present action, causing confusion.

Human Label: Questionable

Source: Sheryl Lee. In 2016, she appeared in Cafe Society, and also completed the Showtime revival of Twin Peaks (2017), reprising her role of Laura Palmer. Sheryl Lee Ralph. In 2016 she played Madame Morrible in Wicked on Broadway.

Summary: Sheryl Lee appeared in the film Cafe Society in 2016 and reprised her role as Laura Palmer in the Showtime revival of Twin Peaks in 2017. Sheryl Lee Ralph played Madame Morrible in the Broadway production of Wicked in 2016.

Explanation: The source indicates that 2017 was the release year for the Showtime revival of Twin Peaks, while Sheryl Lee completed her role in 2016. The summary misrepresents this timeline, causing a distortion of the original information.

Human Label: Unwanted.Intrinsic

Source: As of 22 February 2020 , 77,984 cases have been confirmed , including in all provinces of China and more than two dozen other countries .

LLM Summary: As of 22 February 2020, there were 77,984 confirmed cases of the disease, present in all Chinese provinces and over two dozen other countries.

Explanation: The source does not specify that the cases are related to a disease, and this information cannot be inferred with certainty.

Human Label: Unwanted.Extrinsic

Source: Albert Lawrence Brooks (born Albert Lawrence Einstein; July 22, 1947) is an American actor, filmmaker, author and comedian. Finding Dory. Directed by Andrew Stanton with co-direction by Angus MacLane, the screenplay was written by Stanton and Victoria Strouse.

Summary: The passage provides information about Albert Lawrence Brooks, an American actor, filmmaker, author, and comedian. It also mentions the film “Finding Dory,” which was directed by Andrew Stanton with co-direction by Angus MacLane. The screenplay for the film was written by Stanton and Victoria Strouse.

Explanation: The source does not explicitly state that “Finding Dory” is a film, but this can be reasonably inferred from the context or verified with general knowledge.

Human Label: Benign

Figure 10: Examples of each hallucination type