

The FACTS Leaderboard: A Comprehensive Benchmark for Large Language Model Factuality

Aileen Cheng^{*,♦}, Alon Jacovi^{*,♦}, Amir Globerson^{*,♦}, Ben Golan^{*,♦}, Charles Kwong^{*,◊}, Chris Alberti^{*,◊}, Connie Tao^{*,◊}, Eyal Ben-David^{*,♦}, Gaurav Singh Tomar^{*,◊}, Lukas Haas^{*,◊}, Yonatan Bitton^{*,♦}, Adam Bloniarz[♦], Aijun Bai[◊], Andrew Wang[♦], Anfal Siddiqui[◊], Aravindan Raghuveer[◊], Arturo Bajuelos Castillo[◊], Aviel Atias[♦], Chang Liu[♦], Corey Fry[♦], Daniel Balle[◊], Deepanway Ghosal[◊], Doron Kukliansky[♦], Dror Marcus[♦], Elena Gribovskaya[◊], Eran Ofek[♦], Honglei Zhuang[◊], Itay Laish[♦], Jan Ackermann[◊], Lily Wang[◊], Meg Risdal[♦], Megan Barnes[◊], Michael Fink[♦], Mohamed Amin[♦], Moran Ambar[♦], Natan Potikha[♦], Nikita Gupta[◊], Nitzan Katz[♦], Noam Velan[◊], Ofir Roval[♦], Ori Ram[♦], Polina Zablotskaia[◊], Prathamesh Bang[♦], Priyanka Agrawal[◊], Rakesh Ghiya[◊], Sanjay Ganapathy[◊], Simon Baumgartner[◊], Sofia Erell[♦], Sushant Prakash[◊], Thibault Sellam[◊], Vikram Rao[◊], Xuanhui Wang[◊], Yaroslav Akulov[◊], Yulong Yang[♦], Zhen Yang[◊], Zhixin Lai[◊], Zhongru Wu[♦], Anca Dragan[◊], Avinatan Hassidim[♦], Fernando Pereira[◊], Slav Petrov[◊], Srinivasan Venkatachary[◊], Tulsee Doshi[◊], Yossi Matias[♦], Sasha Goldshtein[♦] and Dipanjan Das[◊]

^{*}Equal Contribution, [◊]Google DeepMind, [♦]Google Research, [◊]Google Cloud, [♦]Kaggle

We introduce The FACTS Leaderboard, an online leaderboard suite and associated set of benchmarks that comprehensively evaluates the ability of language models to generate factually accurate text across diverse scenarios. The suite provides a holistic measure of factuality by aggregating the performance of models on four distinct sub-leaderboards: (1) FACTS Multimodal, which measures the factuality of responses to image-based questions; (2) FACTS Parametric, which assesses models' world knowledge by answering closed-book factoid questions from internal parameters; (3) FACTS Search, which evaluates factuality in information-seeking scenarios, where the model must use a search API; and (4) FACTS Grounding (v2), which evaluates whether long-form responses are grounded in provided documents, featuring significantly improved judge models. Each sub-leaderboard employs automated judge models to score model responses, and the final suite score is an average of the four components, designed to provide a robust and balanced assessment of a model's overall factuality. The FACTS Leaderboard Suite will be actively maintained, containing both public and private splits to allow for external participation while guarding its integrity. It can be found at <https://www.kaggle.com/benchmarks/google/facts>.

1. Introduction

Large Language Models (LLMs) have improved dramatically in recent years, yet they continue to generate factually inaccurate information. Indeed, factuality remains one of the most critical and challenging aspects of LLMs. Research in this area can be broadly divided into two distinct scenarios: (1) factuality with respect to a given context, where a model's response must be fully grounded in the provided information (e.g., a document or an image; Honovich et al., 2022; Jacovi et al., 2025; Rashkin et al., 2023; Tang et al., 2024), and (2) factuality with respect to general world knowledge, where a model must accurately answer factoid queries using its internal parameters (Chen et al., 2023; Kwiatkowski et al., 2019; Lin et al., 2022) or by using external sources like the web (Mialon et al., 2023; Vu et al., 2023; Wei et al., 2024, 2025; Yang et al., 2024). Practical use cases would typically rely on both these capabilities (e.g., analyzing financial reports).

While previous works, including our own FACTS Grounding benchmark (Jacovi et al., 2025), each focused on specific capabilities, a comprehensive understanding of an LLM's factuality requires evaluating its performance across a wider spectrum of tasks. Models that excel at summarizing a

Table 1 | Main results on the FACTS benchmark suite. Values represent Accuracy with 95% confidence intervals, averaged over both the public and the private datasets. The *FACTS Score* is an aggregate metric representing the average performance across all the four subsets.

#	Model	FACTS Score	Grounding	Multimodal	Parametric	Search
1	Gemini 3 Pro	68.8	$69.0_{\pm 2.1}$	$46.1_{\pm 3.6}$	$76.4_{\pm 1.8}$	$83.8_{\pm 1.6}$
2	Gemini 2.5 Pro	62.1	$74.2_{\pm 2.0}$	$46.9_{\pm 3.6}$	$63.2_{\pm 2.0}$	$63.9_{\pm 2.0}$
3	GPT 5	61.8	$69.6_{\pm 2.1}$	$44.1_{\pm 3.5}$	$55.8_{\pm 2.1}$	$77.7_{\pm 1.6}$
4	Grok 4	53.6	$54.7_{\pm 2.3}$	$25.7_{\pm 3.1}$	$58.6_{\pm 2.1}$	$75.3_{\pm 1.8}$
5	GPT o3	52.0	$36.2_{\pm 2.2}$	$39.9_{\pm 3.5}$	$57.1_{\pm 2.1}$	$74.8_{\pm 2.0}$
6	Claude 4.5 Opus	51.3	$62.1_{\pm 3.3}$	$39.2_{\pm 3.5}$	$30.6_{\pm 2.0}$	$73.2_{\pm 1.7}$
7	GPT 4.1	50.5	$45.6_{\pm 2.3}$	$40.1_{\pm 3.5}$	$51.5_{\pm 2.1}$	$64.6_{\pm 2.2}$
8	Gemini 2.5 Flash	50.4	$69.9_{\pm 2.1}$	$41.0_{\pm 3.5}$	$30.7_{\pm 2.0}$	$60.0_{\pm 2.2}$
9	GPT 5.1	49.4	$50.0_{\pm 2.3}$	$41.8_{\pm 3.5}$	$43.2_{\pm 2.1}$	$62.4_{\pm 2.1}$
10	Claude 4.5 Sonnet <i>Thinking</i>	49.1	$61.8_{\pm 2.3}$	$31.1_{\pm 3.3}$	$29.0_{\pm 1.9}$	$74.5_{\pm 1.8}$
11	Claude 4.1 Opus	46.5	$54.8_{\pm 2.4}$	$33.1_{\pm 3.4}$	$33.2_{\pm 2.0}$	$65.0_{\pm 1.9}$
12	GPT 5 mini	45.9	$58.3_{\pm 2.3}$	$41.5_{\pm 3.5}$	$16.0_{\pm 1.5}$	$67.9_{\pm 1.9}$
13	Claude 4 Sonnet	42.8	$56.1_{\pm 2.3}$	$28.6_{\pm 3.2}$	$20.4_{\pm 1.8}$	$66.3_{\pm 1.9}$
14	GPT o4 mini	37.6	$29.3_{\pm 3.0}$	$34.5_{\pm 3.4}$	$20.5_{\pm 1.7}$	$66.2_{\pm 2.0}$
15	Grok 4 Fast	36.0	$43.1_{\pm 2.3}$	$17.7_{\pm 2.7}$	$15.8_{\pm 1.6}$	$67.3_{\pm 1.9}$

provided document may struggle to answer factoid questions from memory, and vice-versa. A robust factuality benchmark should therefore measure a model’s capabilities in multiple contexts, including its handling of different modalities (text, images), knowledge sources (provided context, internal parameters, external search), and response formats.

The FACTS Leaderboard introduced here is designed to address this need by providing a holistic evaluation suite. It aggregates performance across four specialized sub-leaderboards, each targeting a distinct dimension of factuality.

- **FACTS Multimodal** tests a model’s ability to combine visual grounding with world knowledge to answer questions about an image.
- **FACTS Parametric** measures the model’s ability to use its internal knowledge accurately in factoid question use-cases.
- **FACTS Search** evaluates the practical and increasingly common use case of generating factual responses by interacting with a search tool.
- **FACTS Grounding v2** is an updated version of FACTS Grounding, which tests grounding to a given document, with improved judges.

By combining these diverse evaluations into a single suite, we aim to offer a more comprehensive and robust measure of a model’s factual reliability, rather than focusing on a narrow set of tasks.

In what follows, we introduce the four pillars of the FACTS Leaderboard Suite, detail the methodology for each, and explain the aggregation process that yields a single, comprehensive factuality score. We believe this suite offers a nuanced and thorough tool for tracking progress in the ongoing challenge of building factually reliable LLMs.

2. The FACTS Leaderboard

The FACTS suite provides a systematic protocol for evaluating LLMs on diverse aspects of factuality. We maintain a live leaderboard tracking performance on the four FACTS benchmarks: Multimodal, Parametric, Search, and Grounding. The leaderboard will remain open to new model submissions.

To mitigate overfitting, only a subset of the prompts will be released publicly, and the remaining prompts will remain private. All model evaluation will be conducted by Kaggle.

Detailed metrics for each of the four tasks will be published on dedicated task-specific leaderboards. Since each task measures a different aspect of factuality, analyzing them separately yields the most complete insight. However, to facilitate comparison across benchmarks, the main leaderboard will feature a single holistic performance metric: the average accuracy across all four tasks, where accuracy per task is the average over the public and private sets. We refer to this metric as the “FACTS Score”.

Table 1 presents the evaluation of proprietary API-based models across the full FACTS benchmark suite (i.e., both private and public sets). For this overview, we report the FACTS Score, as well as the accuracy for each subset, with 95% confidence intervals.

In the following sections, we describe the four benchmarks, outlining the dataset construction, defining the tailored metrics, and analyzing benchmark-specific outcomes.

3. FACTS Multimodal

The *FACTS Multimodal* benchmark evaluates the ability of models to generate factually accurate text in response to image-based questions. This task requires the integration of visual grounding with parametric world knowledge, a critical capability for modern multimodal systems. The evaluation framework is designed not only to verify the correctness of claims but also to ensure that responses are sufficiently comprehensive to be helpful to the user.

3.1. Data

The evaluation set contains approximately 1,500 questions, divided and filtered into a 711-item *public set* and an 811-item *private set*. Questions were curated from various sources to reflect diverse real-world user queries and were filtered to focus on objective, information-seeking tasks. The benchmark covers a range of capabilities, including detailed visual description, data interpretation from charts and graphs, object recognition, and logical reasoning about visual scenes. Figure 1 presents the distribution of image and question categories in the public set.

3.2. Metrics

The evaluation produces two primary scores — a *Coverage* score and a *No-Contradiction* score — determined by an automated judge that evaluates each response against a human-authored, ground-truth rubric. This rubric-based methodology is central to our benchmark, as it provides a scalable and objective framework for evaluation.

Reference Rubric Creation. For each question, human annotators created a detailed rubric listing relevant facts. Facts that are critical for a complete and satisfactory answer are labeled as *Essential*, while other relevant, contextual facts are labeled as *Non-Essential*. This distinction allows for a nuanced evaluation of both correctness and substance.

Automated Evaluation. An automated judge, acting as a meticulous fact-checker, is used to verify the model’s response is factual using two boolean verdicts:

Table 2 | Examples from the FACTS Multimodal benchmark, illustrating the rubric-based evaluation. The autorater uses the rubric to generate two distinct verdicts for factuality and completeness.

User Request	Human-Collected Rubrics	Candidate Model Response	Coverage Verdict	Contradiction Verdict
What is this and what year was it introduced?	<p><u>Essential Facts</u></p> <ol style="list-style-type: none"> 1. This is an image of the FC Sonora-Baja California (SBC) 2203, a rare surviving example of an EMD FT diesel-electric locomotive. 2. The FC Sonora-Baja California (SBC) 2203 was introduced in 1949. 3. The image shows a train at the National Museum of Mexican Railways. 4. The train in the image has the numbers "2203" across its front. 5. The image shows an orange train on a sunny day. <p><u>Non-essential Facts</u></p>	<p>Based on the visual details, this is a Ferrocarril Sonora-Baja California (SBC) 2203, a rare surviving example of an EMD FT diesel-electric locomotive.</p> <p>Here are the key details:</p> <ul style="list-style-type: none"> • Model Introduced: The EMD FT model line was introduced in 1939. • Specific Unit History: This specific locomotive was built in 1945 and sold to the Mexican railroad SBC in 1964. <p>It is currently preserved at the National Museum of Mexican Railways in Puebla, Mexico.</p>	50% Complete Covers 1 of 2 essential facts. Fails to correctly identify the introduction year.	<p>final verdict: HAS CLEAR CONTRADICTION(S).</p> <p>reasoning: The response claims the model line was introduced in 1939 and the unit built in 1945, which directly contradicts Fact #2 (introduced in 1949).</p>
What genus does this animal belong to?	<p><u>Essential Facts</u></p> <ol style="list-style-type: none"> 1. The creature belongs to the genus <i>Racta</i>. 2. The scientific name for the creature is <i>Racta apella</i>. 3. The genus <i>Racta</i> is part of the butterfly family <i>Hesperiidae</i>. 4. The <i>apella</i> skipper is native to South America and Central America. 5. Skippers are known for their hooked antennae. <p><u>Non-essential Facts</u></p>	<p>Based on the visual features, this butterfly belongs to the genus Thymelicus.</p> <p>Here is a breakdown of why:</p> <ul style="list-style-type: none"> • Family: It is clearly a member of the Hesperiidae family, identifiable by the hooked antennae. • Genus Characteristics: The genus <i>Thymelicus</i> is characterized by this specific rusty-orange or golden-brown coloration. • Species differentiation: It is likely either a Small Skipper or an Essex Skipper. <p>Both species are common in Europe and parts of the Middle East.</p>	0% Complete Covers 0 of 1 essential facts. Fails to identify the correct genus.	<p>final verdict: HAS CLEAR CONTRADICTION(S).</p> <p>reasoning: The response identifies the genus as <i>Thymelicus</i>, which contradicts Essential Fact #1 (<i>Racta</i>). Consequently, it provides incorrect species and location information (Europe vs. South America).</p>
Write an elaborative description for the image.	<p><u>Essential Facts</u></p> <ol style="list-style-type: none"> 1. This image contains the first nine letters of the alphabet. 2. The letters 'H' and 'T' are painted in yellow, making the word 'hi'. 3. The letters 'ABCDEFGHI' are painted in white. 4. The painting is on a white brick wall. <p><u>Non-essential Facts</u></p> <ol style="list-style-type: none"> 5. At the bottom frame... are shrubs... 6. The large sidewalk... allows for space to view the mural... 7. The artist's name, Kay Rosen, appears in black text below the mural... 	<p>This eye-level, outdoor shot captures a large, contemporary mural painted on an exterior brick wall. The mural itself is a vibrant, sky-blue rectangle featuring the capital letters 'ABCDEFGHI' in a bold, sans-serif font. The letters 'A' through 'G' are rendered in stark white, while the letters 'H' and 'I' are distinctly painted in a vivid yellow, cleverly forming the word 'hi' and drawing attention to this friendly greeting. The artist's signature, '© Kay Rosen,' is visible in small white text in the bottom right corner of the blue mural.</p> <p>(... rest of the response)</p>	100% Complete Covers 4 of 4 essential facts.	<p>final verdict: HAS CLEAR CONTRADICTION(S).</p> <p>reasoning: The response states, "The artist's signature... is visible in small white text in the bottom right corner of the blue mural." According to Fact #7, the artist's name appears in "black text below the mural." The response incorrectly describes the color and location of the signature.</p>

- Coverage verdict: This boolean verdict verifies that the model response includes the essential facts specified in the ground-truth rubric.
- No-Contradiction verdict: This boolean verdict verifies that the model response does not include any claims that contradict either the ground-truth rubrics (essential and non-essential), common knowledge or the input image itself.

Accuracy score: only responses that both cover essential facts and do not include any contradictions are considered accurate. The overall accuracy score is the percentage of such responses in the set. Table 2 demonstrates this dual-verdict process, showcasing the detailed reasoning the autorater provides to justify its verdicts and the nuances of the errors it can detect.

Table 3 presents the results of our main FACTS Multimodal benchmark. The Gemini model family is more recall-oriented than other families, demonstrating high Coverage scores. Conversely, GPT models are more precision-oriented, achieving the highest No-Contradiction Scores.

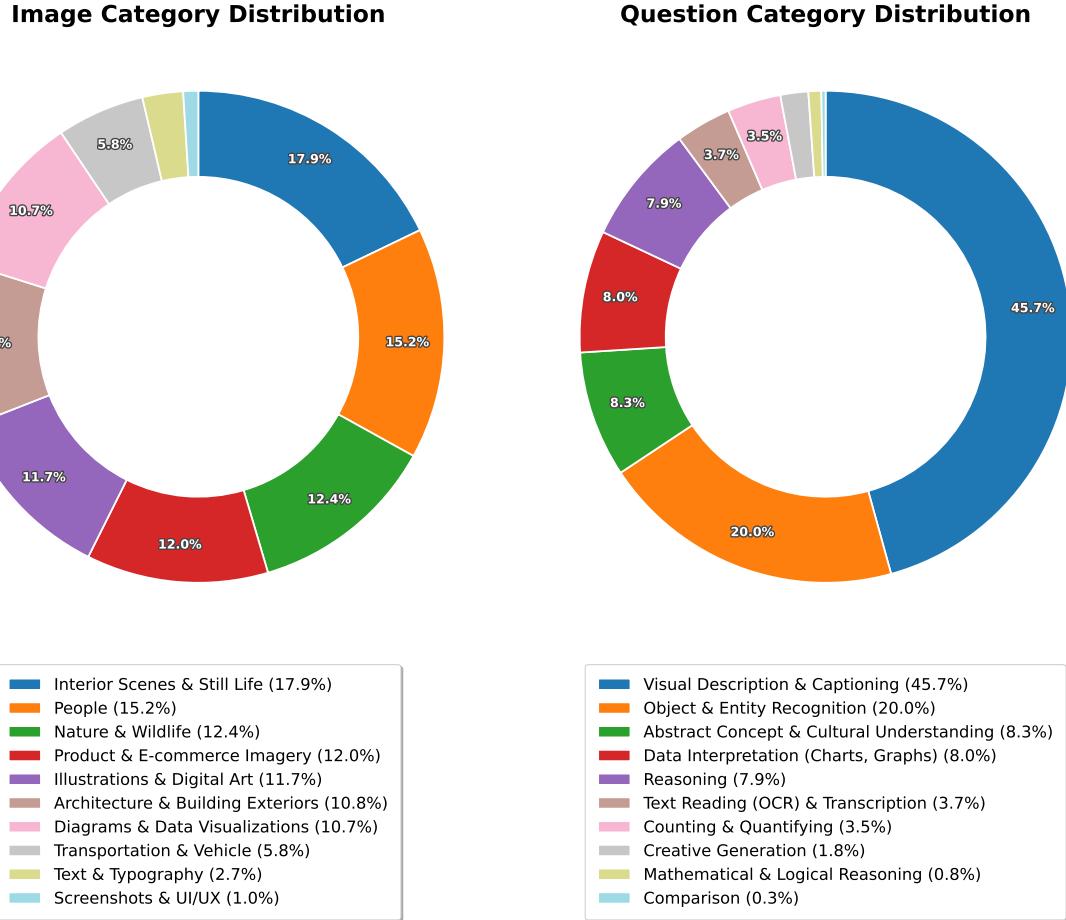


Figure 1 | Distributions of image and question categories in the FACTS Multimodal benchmark.

3.3. Autorater Validation.

The credibility of our automated metrics was established by validating them against human ground-truth annotations. For *Coverage*, the human annotation task mirrored the autorater’s function precisely: given a model response, human raters marked each essential fact from the rubric as either “supported” or “unsupported.” The autorater was tasked with the same objective. This allowed for a direct comparison of the final “percentage of essential facts supported” metric, on which our autorater achieved a high degree of reliability with a Spearman’s rank correlation of 0.64 with human judgments. After applying a threshold to convert this to a boolean outcome by using a threshold of 0.5 to ensure most facts are covered, we obtained a macro F1 score of 72.3.

For *No-Contradiction*, we validated the autorater against fine-grained human annotations. The process required a comprehensive review of the entire model response; as illustrated in Figure 2, annotators assessed the text sentence-by-sentence, using the interface to mark whether each sentence contained a contradiction. This validation achieved a macro F1 score of 78.2. Table 4 details these results, where the positive class denotes the absence of a contradiction.

Table 3 | Detailed results on the FACTS Multimodal benchmark. Numbers are reported on top of the private and the public set.

Model	Accuracy (%)	No-Contradiction (%)	Coverage (%)
Gemini 2.5 Pro	46.9	58.8	67.7
Gemini 3 Pro	46.1	57.3	68.4
GPT-5	44.1	64.7	59.9
GPT-5.1	41.8	65.0	58.3
GPT-5 mini	41.5	65.1	58.8
Gemini 2.5 Flash	41.0	53.9	64.8
GPT-4.1	40.1	62.2	57.3
o3	39.9	55.7	62.2
Claude 4.5 Opus	39.2	51.1	62.1
o4 mini	34.5	47.6	59.1
Claude 4.1 Opus	33.1	49.2	54.6
Claude 4.5 Sonnet <i>Thinking</i>	31.1	42.6	56.0
Claude 4 Sonnet	28.6	46.2	50.2
Grok 4	25.7	32.0	57.8
Grok 4 Fast	17.7	22.0	50.9

Table 4 | Performance of FACTS Multimodal Autoraters: Coverage and No-Contradiction.

Metric	Coverage	No-Contradiction
Macro F1	72.3	78.2
Negative F1	72.3	70.1
Negative Precision	64.6	63.4
Negative Recall	82.6	78.3
Positive F1	72.2	86.3
Positive Precision	82.4	90.7
Positive Recall	64.2	82.3

4. FACTS Parametric

The *FACTS Parametric* benchmark assesses the ability of models to accurately answer factual questions (§4.1.1) that users care about, without the aid of external tools. The questions are derived by user-interest, and their answers are confirmed to exist within Wikipedia, a source available for LLM pretraining. Each of the questions was found challenging through adversarial sampling with a suite of open-source models (see §4.1.2). The resulting benchmark measures how well models can recall challenging facts that users care about and that are supported by a primary source.

4.1. Data

FACTS Parametric consists of 2104 QA pairs, equally divided into a 1052-item public set and a 1052-item private set. We verified that the underlying intent of each question reflects widespread user interest. The questions were then subjected to adversarial sampling, and human verification to create a reliable and challenging benchmark (see §4.1.2).

The resulting queries span a broad range of topics, including politics, sports, and technology, while the answer types fall into diverse categories such as people, dates, and places. Figure 3 presents the breakdown of these topics and answer types in the public set, illustrating the distribution of challenging factoid queries as they appear in real-world traffic. Finally, Table 5 presents a set of examples from the public dataset.

User


Describe this image in details

Model
Sentence 5 out of 10

A medium shot of a soft, plush pink pig stuffed animal facing forward. The pig has short ears that are beginning to droop down. The snout is very small, and the eyes are black with small dark brown pupils. The stomach is a light tan. The underbelly of the pig is not visible. **The pig has black, rounded hooves on its front and back feet and a pink nose.** The pig is sitting on a light brown hard wood floor with a sage green wall behind it. The wall has a horizontal groove where the baseboard is. The pig is casting a shadow on the wall behind it, angled towards the top of the shot. Indoors. The lights are on.

Assessment

What is the relationship of the highlighted sentence with respect to the image?

Nothing to assess

Contradiction

Neutral

Entailment

Explanation

“The pig back hooves are pink, not black. The front hooves are not visible.”

Figure 2 | Example of the sentence-level annotation interface. Annotators are shown the image, the full model-generated paragraph, and a highlighted sentence. They assess its factuality by selecting a label (here, ‘Contradiction’) and providing a textual explanation for any inaccuracies observed.

4.1.1. FACTS Parametric properties

Questions Reflect User Traffic. One of our goals with FACTS Parametric is to evaluate how familiar models are with facts users genuinely care about. To do this, we collected questions that reflect interest shown by many users. However, since strictly following these guidelines tends to yield highly popular (and therefore easier) topics, we deliberately selected the least frequent topics from the eligible set. We then further refined the list of initial questions using adversarial sampling to ensure we only retained questions that remain challenging for models (see §4.1.2).

Answers Explicitly Supported by a Prevalent Source. To effectively measure the recall of parametric knowledge acquired during pre-training, we established a key criterion for FACTS Parametric: every answer must be explicitly supported by information found within Wikipedia documents. We selected Wikipedia as the mandatory source because its content is highly prevalent and widely assumed to be a significant component of the training corpora for all LLMs. This constraint helped ensure that the benchmark evaluates knowledge the model was likely exposed to during training, thereby allowing a clearer assessment of its ability to recall factual information learned during that phase, especially when running the assessment with no access to web-search tools.

Factoid Criteria. To facilitate straightforward and reliable model assessment, we designed FACTS Parametric with a few important properties in mind:

- **Single, Atomic Fact:** Each question targets exactly one piece of factual information, avoiding multi-part queries. This ensures that evaluation focuses on recalling one easily verifiable fact.
- **Unambiguous Answer:** Questions are designed to have only one distinct, correct answer, minimizing ambiguity during evaluation.
- **Clear Answer Specification:** The expected type of answer (e.g., person, location, date) or the required granularity is typically stated in the question itself, or is strongly implied.
- **Concise, Factual Answers:** The expected answer is a short entity (like a name, a number, or a specific term) rather than a simple “yes/no” response (which the model can ‘guess’) or a long detailed response. This simplifies matching model outputs to the ground truth.
- **Stable Facts:** The benchmark focuses on facts that are either static (e.g., “What is the capital of France?”) or explicitly time-anchored within the question (e.g., “Who was the US president in 1995?”), ensuring the stability and longevity of the ground truth answers.

7

Table 5 | Example questions from the *FACTS Parametric* benchmark (Public set).

Question	Answer	URL Suffix	Topic	Answer Type
kevin abstract hometown	Corpus Christi, Texas	Kevin_Abstract	People (artists)	Place
who played harmonica on the rockford files theme song	Tommy Morgan	Tommy_Morgan	People (artists)	Performer
cristian mijares record	59 wins, 9 losses, and 2 draws	Cristián_Mijares	People (sports)	Person
triphosgene boiling point (°C)	206 °C (acceptable range 203–206)	Triphosgene	Science	Number

4.1.2. Data Processing

To satisfy the structural requirements detailed in §4.1.1 and construct a challenging benchmark, we implemented a multi-stage filtering pipeline. First, we applied automatic LLM-based filters to the initial set of questions, which were collected to reflect user interest, to identify queries satisfying the factoid criteria above. Next, we utilized an adversarial sampling mechanism to isolate the most challenging examples. Finally, we conducted human verification to confirm adherence to all specified properties.

Adversarial Sampling with Open LLMs. While FACTS Parametric is grounded in questions frequently asked by users, a key goal is to ensure the benchmark remains challenging for frontier LLMs, avoiding saturation in the near future. To achieve this, we employ an adversarial sampling strategy during the data collection phase.

First, we generate preliminary “silver” labels for all questions using *Gemini-2.5-Pro* equipped with search tools. In this setup, the model returns both a generated answer and the specific search results used to derive it. Since we aim to ground our benchmark in verifiable sources, we filter this set to retain only the questions where the model’s answer is supported by a Wikipedia URL found in the search results.

Next, to identify the most challenging questions among this verified set, we collect responses from five strong open-weight models (i.e., models with publicly available weights that can be run locally). We specifically utilized open-weight architectures to decouple the adversarial selection process from the proprietary API models used in our evaluation, thereby ensuring unbiased filtering. Crucially, we query these models in a closed-book setting, without access to external search tools.

Finally, we retain only the questions that none of these open-weight models answered correctly.¹ As a final step, this refined list is sent to human annotation for rigorous verification (see below).

Human Annotation. To maintain data quality and validity, each question-answer (QA) pair was verified by three independent third-party human annotators. Annotators received the question, the LLM-generated candidate answer (silver label), and a supporting URL from the initial data collection. Their role was to confirm the factual accuracy of the QA pair and its compliance with the FACTS Parametric properties detailed in Section 4.1.1.

¹At this phase, “correctly” is defined as matching the silver answer provided by Gemini-2.5-Pro. Note that these are preliminary filters; the final labels are determined later.

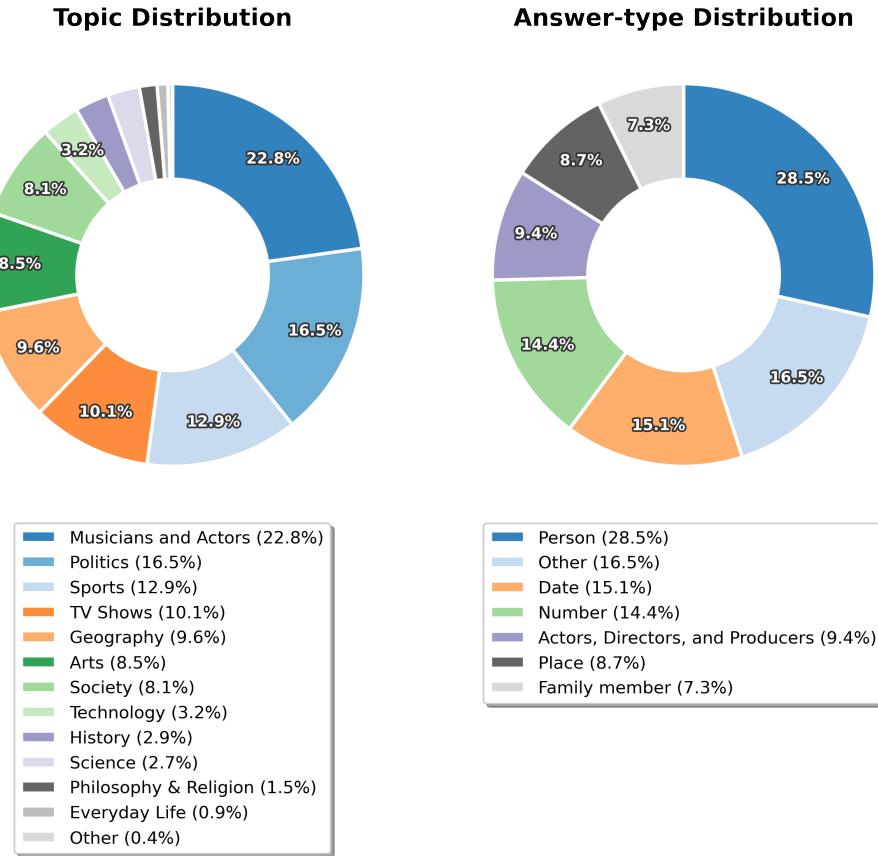


Figure 3 | Distributions of context domain and of answer type as a percent of the total set of questions in the FACTS Parametric benchmark.

Specifically, annotators performed the following evaluations for each example:

- **Factuality Assessment:** Determine the correctness of the provided answer for the given question. Annotators were explicitly instructed to base their judgment on information available on the web, and not solely on the provided URL. The possible verdicts were:
 - *Accurate*: The answer is verifiably correct based on accessible online information.
 - *Inaccurate*: The answer is deemed incorrect due to one of the following reasons: (a) reliable online sources provide contradictory information, (b) no supporting information could be found online, or (c) the topic is disputed, with both supporting and contradicting information found online.
- **Properties Compliance Check:** Verify whether the question and answer strictly adhere to all the dataset properties described in Section 4.1.1. The outcome was a simple True/False verdict.
- **Wikipedia Evidence Extraction:** Identify and provide a Wikipedia document that confirms the fact represented by the QA pair. The rater was given a Wikipedia URL that might contain the required information, but it had to be checked. When annotators did not find a supporting Wikipedia URL, the example was disqualified from appearing in the final dataset.
- **Correction Provision:** If a QA pair was initially rated as ‘Inaccurate’ or failed the ‘Properties Compliance Check’, but the issue could be resolved with a simple correction (to either the question, answer, or URL) without altering the original user intent, annotators were instructed to provide the corrected version. Otherwise, the QA pair was discarded. In the cases where a correction was proposed, a new annotator was instructed to finalize the examples with proposed correction. The extra annotator had the option to disqualify the examples from appearing in the final set.

Table 6 | Detailed results on the FACTS Parametric benchmark. Numbers are reported on top of the private and the public set.

Model	F1	Accuracy (%)	Attempted accuracy (%)	Hedging rate (%)
Gemini 3 Pro	77.0	76.4	77.6	1.4
Gemini 2.5 Pro	63.8	63.2	64.5	1.9
Grok 4	60.9	58.5	63.6	7.9
GPT-5	59.7	55.7	64.3	13.3
o3	57.6	57.0	58.2	1.9
GPT-4.1	52.5	51.5	53.6	3.8
GPT-5.1	45.8	43.2	48.7	11.3
Claude 4.1 Opus	39.0	33.2	47.4	29.9
Claude 4.5 Sonnet	34.3	28.9	42.2	31.3
Claude 4.5 Opus	32.8	30.5	35.5	14.0
Gemini 2.5 Flash	31.4	30.6	32.3	5.1
Claude 4 Sonnet	26.3	20.3	37.2	45.1
GPT-5 mini	24.2	16.0	49.6	67.6
o4 mini	21.6	20.4	22.9	10.5
Grok 4 Fast	16.9	15.7	18.4	14.4

This multifaceted annotation process, including verification, evidence extraction, and correction, was vital for constructing a high-quality benchmark.

4.2. Metrics

We follow the grading scheme proposed by [Wei et al. \(2024\)](#) with some modifications. First, we slightly change the examples given in the grader instruction prompt to better represent the scenarios we see in our data. Then, we introduce an additional grading label, *unknown*, which represents cases where the grader is unsure whether the gold answer and the model response are aligned. We find this label to improve the already high accuracy presented by the grader.

Accordingly, the resulting grader automatically grades each model response as either *correct*, *incorrect*, *not-attempted*, or *unknown*. Our primary metric is *accuracy*, which is measured by the percentage of correct responses. We also report three secondary metrics: *Hedging rate* (the percentage of not-attempted), *attempted-accuracy*, and *F1-score*, the harmonic mean of accuracy and accuracy given attempted.

To enhance grading reliability, we sample three grades from *Gemini-2.5-Pro* for each query, gold-answer, response triplet and average them to determine the final score. We observed that utilizing a powerful model for grading improves grading accuracy. We ultimately decided to standardize on *Gemini-2.5-Pro* as our sole judge to keep the benchmark simple and easy to maintain. We validated this choice by comparing it against a mixed-model panel (sampling once each from *Gemini-2.5-Pro*, *GPT-o3*, and *Grok-4*). The results confirmed that using *Gemini-2.5-Pro* alone preserves the same relative performance trends and rankings as the more complex ensemble.

Table 6 presents the main results for the FACTS-parametric benchmark. For all metrics, we report the average score derived from three sampled grades per model response. These results illuminate model behavior regarding uncertainty: while guessing is the optimal strategy for standard accuracy, the “attempted-accuracy” metric incentivizes hedging.

This trade-off is evident when comparing GPT-o3 and GPT-5. Although GPT-o3 achieves higher raw accuracy (57.0% vs. 55.7%), GPT-5 hedges significantly more often (13.3% of cases vs. 1.9%). Consequently, GPT-5 achieves superior attempted accuracy (64.3% vs. 58.2%) and F1 scores (59.7 vs. 57.6).

Table 7 | Examples from *FACTS Search* (Public set).

Slice	Question	Answer
Hard Tail	For the person who had the most followed Instagram account in 2017, how many solo studio albums did they release prior to this accomplishment?	Two
Wiki Two-Hop	In which month of 2017 did the European Medicines Agency grant orphan drug designation to the drug with DrugBank ID DB11978?	October
Wiki Multi-Doc	What is the sum of the birth years of the British boxer who defeated Vazik Kazarian at the 1960 Summer Olympics, the Moroccan boxer who also competed in the men's light welterweight event at those same Olympics, and the Danish boxer who competed in both the 1960 and 1964 Summer Olympics?	5821
KG Hops	Among all the films written by the creator of the TV Program The Sopranos, which one was released the earliest?	Grave of the Vampire

5. FACTS Search

The *FACTS Search* benchmark evaluates the ability of models to use web search. Indeed, most recent generative models are designed to make use of a search tool, and several benchmarks have been proposed to test this capability (e.g., see [Wei et al., 2025](#); [Yang et al., 2024](#)).

The key challenge when designing such benchmarks is to collect questions whose solution requires search tools. Generally, the only case where this is guaranteed is when the information sought by the question does not appear in the model training data, and does appear in the search results. However, this is hard to guarantee across models (whose training data and cutoff dates are different), and thus we do not pursue this direction. Instead, we focus on other aspects that are hard for models that do not have access to web search. These include tail-entities which are often not sufficiently encoded in parameters, and multi-hop queries (where it's less likely that all hops are encoded in the parameters).

5.1. Data

The evaluation set contains 1884 questions, which are split into public and private test sets, of sizes 890 and 994 respectively. The questions were collected from various sources, with the goal of capturing diverse aspects of queries that are likely to require search. Table 7 shows some examples from the dataset. The set of questions consists of four subsets, each generated using a different strategy. One subset was written by human raters, and the other three were synthetically generated. See details below.

- **Hard Tail** : A set of questions written by human raters as follows. The raters were instructed to write questions that require information that is challenging to extract with web search. Namely, that there is no single-step web search answer available on the first page, or the information is not readily available as a verbatim piece of text on the internet. Raters were also asked to verify that the Gemini model publicly available at the time (Gemini 1.5) could not solve these, even when using search.
- **Wiki Two-Hop** - A set of question-answer pairs, synthetically generated using Wikipedia as follows. An initial set of QA pairs was extracted from Wikipedia abstracts, and filtered to focus on tail entities. Next, each question was modified to be a harder, multi-step question via synthetic alteration. This was achieved by substituting the main entity of the question with a different description of this entity that is extracted from the Google Knowledge Graph. For example,

Table 8 | Detailed results on the FACTS Search benchmark. Numbers are reported on top of the private and the public set.

Model	F1	Accuracy (%)	Attempted accuracy (%)	Hedging rate (%)	Average searches
Gemini 3 Pro	85.6	83.8	87.6	4.4	3.39
GPT-5	81.4	77.7	85.5	9.1	4.28
Claude 4.5 Opus	80.0	73.2	88.2	16.9	3.98
Claude 4.5 Sonnet	78.5	69.8	89.7	22.2	4.02
o3	77.6	74.8	80.7	7.3	4.64
Grok 4	77.4	75.3	79.6	5.4	4.5
GPT-5 mini	76.5	67.9	87.5	22.5	3.97
Grok 4 Fast	75.1	67.3	85.0	20.9	4.74
Claude 4.1 Opus	75.1	65.0	88.9	26.9	4.66
Claude 4 Sonnet	74.7	66.3	85.7	22.6	4.49
GPT-4.1	65.8	64.6	67.1	3.7	3.17
o4 mini	68.7	66.2	71.4	7.2	4.66
Gemini 2.5 Pro	68.2	63.9	73.1	12.6	3.64
Gemini 2.5 Flash	67.2	60.0	76.5	21.5	3.38
GPT-5.1	66.7	62.4	71.7	13.0	3.34

“What is the birthplace of John Lennon” could be modified to “What is the birthplace of Yoko Ono’s spouse”. These questions were not adversarially filtered, but evaluation on Search-Off and Search-On Gemini models available at the time gave low accuracies of 30% and 38%.

- **Wiki Multi-Doc** - A set of question-answer pairs synthetically generated from multiple Wikipedia documents as follows. First, a seed document D_{seed} was sampled. Then, it was used to sample a set of similar documents $D_{related}$, and only the documents D_{torso} with rank that is neither too low nor too high are kept.² Finally, Gemini was prompted to synthesize a query-answer pair (Q, A) from the content of these n documents. The query Q was formulated to be answerable only by synthesizing information present in both the seed document D_{seed} and one or more documents from the D_{torso} subset. The prompt also encouraged Gemini to find interesting ways to connect information rather than relying on a simple direct chain or a combination of unrelated queries. Next, questions were filtered as follows. First, an automated critic model filtered out pairs where the question was not self-contained or the answer was not strictly grounded in the source documents. Second, a hardness filter was applied, discarding any queries that Gemini could correctly answer when utilizing standard web search tools.
- **KG Hops** - A set of question-answer pairs synthetically generated using multiple hops in the Google Knowledge Graph. To generate these, we first collected common path-queries, such as “films that actor X appeared in”. These queries were then concatenated and combined with other functions (e.g., max) to create more complex ones. For example, “films that actor X appeared in” and “publication date of film X” could be combined to create “publication date of the first film that actor X appeared in”.

The collection process above resulted in a set of questions with corresponding gold answers, either written by human raters or extracted automatically as part of the data generated process. In order to further check the quality of the answer for all questions, three independent human raters were asked to rate each question and answer according to the following criteria:

- **Correctness:** Use Google search to check that the provided answer is a correct response to the question.
- **Uniqueness:** Check whether there are entities different and distinct from the provided answer that could also be correct answers to the given question.
- **Immutability:** Identify whether the answer to the given query is likely to change in the next five years.

²This eliminates documents that are too similar or too distant from D_{seed} .

Table 9 | Examples from *FACTS Grounding v1* (Public set).

System Instruction	Context Document Description	Context Tokens	User Request
Answer the question using only the information provided in the context. Do not rely on external knowledge or sources.	The development and deployment of an autonomous robotic system designed to clean skyscraper windows, highlighting its technological advancements, safety implications, and potential impact on the window-washing industry.	~1.1k	My sister and her dog live in NYC. I've visited there and have always been fascinated with their tall buildings. Then I thought...someone has to clean those! Then next thing you know, window washing robots popped up on my feed. How do these robots work? Also what does this mean for the people who do those jobs?
Provide a response based solely on the information provided in the prompt. External sources and prior knowledge must not be used.	Legal interpretations and effects of the medical marijuana appropriations rider on federal marijuana prosecutions, focusing on differing circuit court approaches to determining compliance with state medical marijuana laws.	~1.6k	What did the first circuit conclude?
This task requires you to answer questions based solely on the information provided in the prompt. You are not allowed to use any external resources or prior knowledge. Present your answer in headed sections with an explanation for each section. Each explanation should be in bullet points with exactly three bullet points.	Comparison of different economic systems, including free market, command, and mixed economies, highlighting their key characteristics, advantages, and disadvantages.	~0.9k	which famous economists are mentioned?
Answer the question based solely on the information provided in the passage. Do not use any external knowledge or resources.	A study that investigates the correlation between advanced maternal age (40+) and increased risk of obstetric, fetal, and neonatal complications compared to women aged 25-35.	~2.1k	Researchers at Foch Hospital in France published this study of pregnancy outcomes in two groups of patients. Please summarize outcomes across the three kinds of complications that the researchers studied.

The dataset was filtered to include only questions that all three raters marked as correct, unique and immutable. Finally, all questions were filtered to exclude those that Gemini 2.5 Flash without search answered correctly, thus emphasizing the need for a search tool. The resulting final dataset sizes were 328, 932, 268, 356 for Hard Tail, Wiki Two-Hop, Wiki Multi-Doc, and KG Hops respectively.

5.2. Search Engine

The goal of the FACTS Search benchmark is to evaluate how well LLMs use search tools. Because performance relies heavily on the specific tool used, meaningful comparison requires that all models access the same search engine. The FACTS leaderboard evaluation uses the Brave Search API as the search tool. All evaluated models receive the same description of the tool. When an LLM triggers a tool call, the API is queried, and the output is appended to the LLM context.

5.3. Metrics

To evaluate the quality of model responses, we use a prompted auto-rater. Specifically, given a query, a model response, and a gold response, Gemini 2.0 Flash is prompted to assess if the response is correct, incorrect or does not attempt to answer the query.

Table 8 presents our main FACTS Search results. Notably, Gemini 3 Pro, the highest-performing model, conducts fewer searches on average than other top models, whereas the Grok model family searches the most. Consistent with observations in the FACTS Parametric slice, Claude appears to prioritize accuracy on attempted queries over overall accuracy.

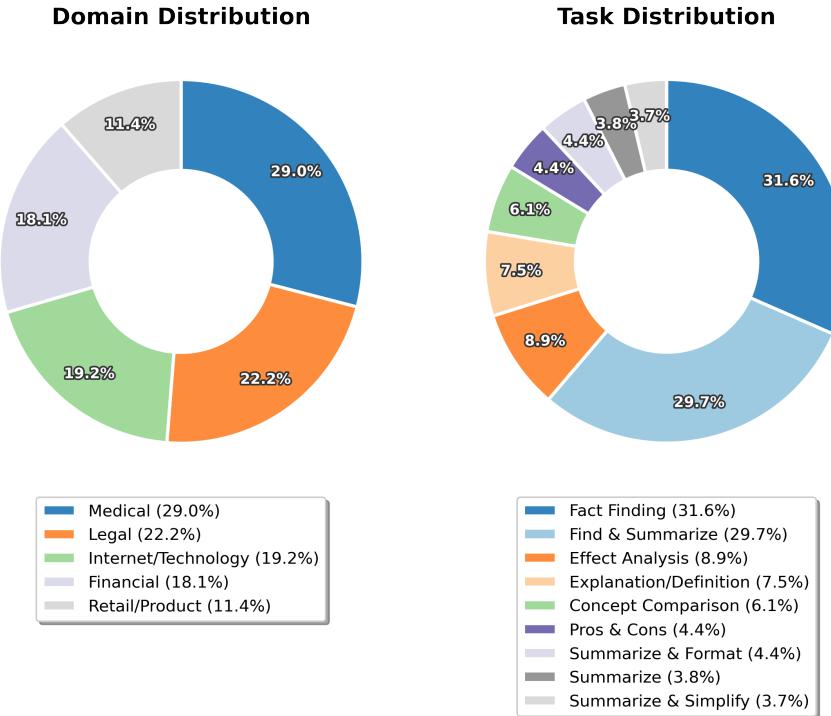


Figure 4 | Distributions of context domain and of task requested by the user as a percent of the total set of prompts in the benchmark.

6. FACTS Grounding v2

The *FACTS Grounding v2* benchmark extends the *FACTS Grounding* benchmark (referred to as *FACTS Grounding v1* below) previously introduced in [Jacovi et al. \(2025\)](#). *FACTS Grounding v1* evaluated whether a model response is consistent with the given context document and user query about that document. Here we briefly re-introduce *FACTS Grounding v1*, and then describe the new version, in which the judge models were updated. Please refer to [Jacovi et al. \(2025\)](#) for more details about the original *FACTS Grounding* leaderboard.

6.1. Data

The set of prompts for *FACTS Grounding v2* is the same as v1. We provide a description below for completeness. Third-party human raters were instructed to design prompts requiring the processing of long-form input and the writing of long-form output. These tasks include Q&A, summarization, and document rewriting. Each example within our evaluation set consists of a context, which is a document or set of reviews sourced from the web, paired with a non-trivial user request that can be addressed using the provided context, necessitating a long-form response. Additionally, each example includes a system instruction directing the model to generate its response exclusively from the given context, without incorporating external knowledge.

To ensure the diversity of the evaluation set, prompts were generated across a range of document lengths (up to 32k tokens) and various enterprise domains, including finance, technology, retail, medical, and legal. The annotation instructions were carefully designed to avoid prompts requiring creative responses, expert-level domain knowledge, mathematical or logical reasoning, or meta-analysis of the text, such as tone analysis or interpretation of author intent. Table 9 provides concrete examples of data instances in the collection. The specific distributions of enterprise domains and of tasks requested by users are shown in Figure 4.

Table 10 | Evaluation of different judge models and evaluation prompts on a private test-set (N=320, class ratio 79:19). Chosen judges via Macro-F₁ in **bold**.

Judge Model	Prompt	Macro-F ₁	F ₁ (+)	F ₁ (-)	Acc.	Prec.	Rec.
gemini-2.5-flash	v2	65.33	84.51	46.15	75.94	85.71	83.33
gpt-5-2025-08-07	v2	65.18	80.09	50.27	71.56	89.27	72.62
gemini-2.5-pro	v1	64.87	87.92	41.82	80.00	83.81	92.46
gemini-2.5-pro	v2	64.07	84.17	43.97	75.31	85.02	83.33
gemini-2.5-flash	v1	63.44	85.60	41.27	76.88	83.97	87.3
gemini-2.0-flash	v1	63.43	86.86	40.00	78.44	83.52	90.48
gpt-5-2025-08-07	v1	61.83	87.48	36.19	79.06	82.69	92.86
gpt-4o	v1	59.73	85.28	34.19	75.94	82.29	88.49
gemini-1.5-pro	v1	56.89	89.09	24.69	80.94	81.11	98.81
gemini-1.5-pro	v2	55.65	86.56	24.74	77.19	80.76	93.25
gemini-2.5-flash	v1	54.78	88.13	21.43	79.38	80.59	97.22
claude-3-5-sonnet-20241022	v1	53.15	79.36	26.95	67.81	80.16	78.57
gemini-2.0-flash	v2	51.25	86.59	15.91	76.88	79.67	94.84

6.2. Metrics

We keep the same evaluation approach as in FACTS Grounding v1 (Jacovi et al., 2025). Specifically, we first prompt multiple “judge” LLMs to determine if the response is grounded in the input. Finally, we find “ineligible” responses that do not sufficiently address the user request, and mark these as inaccurate, so only grounded and eligible responses are labeled as accurate. The main updates in FACTS Grounding v2 are the LLMs used as judges, and the prompt used. See details below.

Unadjusted Factuality Score. As in FACTS Grounding v1, the principal component of our evaluation process is an unadjusted *factuality score*, which is the initial score *before* the adjustment for ineligible responses (described later).

First, we utilize a language model judge to produce a binary classification label identifying whether a full model response is grounded in the user request and the context document given an instruction (see Table 9). A model response is marked with a positive label (“accurate”) if all the claims in the response are grounded in the contents of the prompt, or do not require grounding; the response is marked with a negative label (“not accurate”) if at least one claim that bears information is deemed to be not grounded in the contents of the prompt. We use two different judge models in order to reduce the bias of a particular judge model, as models have been shown to be biased towards favorably judging their own outputs (Wataoka et al., 2024). The judge models are *Gemini 2.5 Flash* (Comanici et al., 2025) and *GPT-5* (OpenAI, 2025). We note that *FACTS Grounding v1* used a different set of models: *Gemini 1.5 Pro* (Gemini Team: R. Anil et al., 2023), *GPT-4o* (Achiam et al., 2023), and *Claude 3.5 Sonnet* (Anthropic, 2024).

To evaluate the quality of the new judge models, we compared them to human-ratings on a held-out evaluation set (N=320). We also investigated changes to the judge prompt templates. Specifically, we considered two prompt variants: the one from FACTS Grounding v1, and a slightly modified version v2. Results are shown in Table 10, and demonstrate that the new models combined with the v2 prompt outperform other model-prompt combinations. Given the two judges, the individual factuality score for each judge is the percentage of accurate responses, and the unadjusted factuality score is the average of all judge scores.

Disqualifying Ineligible Responses. Metrics that are focused on evaluating the factuality of the generated text with respect to a context document can be “hacked” by ignoring user intent. Namely, by providing shorter responses that evade conveying adequately comprehensive information, even if such content was an important aspect of a user request, it is possible to achieve a high factuality score while not providing a helpful response. See illustrative examples in Table 11.

Table 11 | Examples of ineligible responses: these responses, while fully grounded in the context document, fail to address the user request meaningfully and are consequently considered ineligible.

Context Document Description	User Request	Ineligible Response	Rationale
A research report on renewable energy sources, including wind, solar, and hydroelectric power, with specific statistics and case studies.	Can you summarize the key advantages and disadvantages of wind energy from this document?	Wind energy is good because it is renewable and clean, but it has some challenges too.	(1) The response is extremely vague, failing to provide any specific points from the document, such as the cost-effectiveness, geographic limitations, or impacts on wildlife. (2) It does not engage with the query's focus on key advantages and disadvantages.
A company's annual financial report, discussing quarterly earnings, expenditures, future investments, and an analysis of the market environment.	Summarize the main reasons the company's revenue decreased in Q3.	The company faced challenges in Q3 that impacted its revenue.	(1) The response avoids specifying any reasons, such as market trends, increased competition, or operational setbacks, which would likely be in the document. (2) It does not demonstrate an attempt to engage with or extract relevant details.
A historical article on the causes and consequences of the Great Depression.	What were the main causes of the Great Depression as explained in the document?	The Great Depression was a difficult time in history with many causes and effects.	(1) The response provides no substantive information on the causes, such as stock market speculation, bank failures, or trade policies, which were discussed in the document. (2) It ignores the user's explicit focus on the "main causes."

We safeguard against such responses by using prompted judge LLMs to determine whether a given generated response sufficiently addresses the user's request. The judge LLM is asked to output a binary label indicating the eligibility of the response: either "eligible," signifying that it answers the user request, or "ineligible," otherwise. Ineligible responses are disqualified from factuality evaluation and the final factuality score is adjusted such that ineligible responses are deemed as *inaccurate*. The judge models used are the same two models used for checking grounding.

7. Conclusion

As LLMs improve, existing benchmarks become saturated. It is thus important to introduce benchmarks that challenge current models. Here, we present the FACTS suite, a benchmark where the top performing model has an average accuracy of only 69%, leaving considerable headroom for future progress. The FACTS leaderboard results report high-level metrics such as accuracy. However, it will be useful to obtain more fine-grained analysis, studying what affects the hardness of questions. For example, previous work (Kandpal et al., 2023) has shown that infrequent entities are harder to learn, and it would be interesting to check if this is reflected in FACTS Parametric. Similarly, it would be interesting to study notions of "tailness" for FACTS Search, where some facts might be harder to search for than others.

Each of the FACTS subsets requires different capabilities to solve, and can serve to drive research on these fronts. FACTS Multimodal requires integration of image understanding with knowledge not represented in the image; FACTS Parametric relies on representing broad factual knowledge in the model parameters, FACTS Search involves effective use of search tools, and FACTS Grounding focuses on the ability to ground the response to context. These are all facets of factuality, where the model relies on different sources of information to generate factual responses. Naturally, there are aspects of factuality not covered by FACTS, such as video understanding and fast-changing information. In addition, tool-use introduces new factuality challenges, for example when using knowledge-base calls as a tool. We hope FACTS will inspire additional benchmarks that address these areas and others.

8. Contributions and Acknowledgments

- **Leads:** Connie Tao, Dipanjan Das, Lukas Haas, Sasha Goldshtain.
- **Benchmark design:** Aileen Cheng, Alon Jacovi, Amir Globerson, Andrew Wang, Chang Liu, Chris Alberti, Eyal Ben-David, Gaurav Singh Tomar, Lukas Haas, Mohamed Amin, Ofir Roval, Prathamesh Bang, Yonatan Bitton, Yulong Yang, Zhongru Wu
- **FACTS Team:** Adam Bloniarz, Aijun Bai, Anfal Siddiqui, Aravindan Raghuvir, Arturo Bajuelos Castillo, Aviel Atias, Ben Golan, Charles Kwong, Corey Fry, Daniel Balle, Deepanway Ghosal, Doron Kukliansky, Dror Marcus, Elena Gribovskaya, Eran Ofek, Honglei Zhuang, Itay Laish, Jan Ackermann, Lily Wang, Meg Risdal, Megan Barnes, Michael Fink, Moran Ambar, Natan Potikha, Nikita Gupta, Nitzan Katz, Noam Velan, Ori Ram, Polina Zablotskaia, Priyanka Agrawal, Rakesh Ghiya, Sanjay Ganapathy, Simon Baumgartner, Sofia Erell, Sushant Prakash, Thibault Sellam, Vikram Rao, Xuanhui Wang, Yaroslav Akulov, Zhen Yang, Zhixin (Lucas) Lai
- **Sponsors:** Koray Kavukcuoglu, Anca Dragan, Avinatan Hassidim, Fernando Pereira, Slav Petrov, Srinivasan Venkatachary, Tulsee Doshi and Yossi Matias who both sponsored the effort and provided technical guidance.

We would also like to thank:

- **Gemini Team** for the support and model access.
- **Kaggle Team** for their expertise and releasing the leaderboard.
- **Expert data annotators** who helped to collect examples in the paper.
- **Our reviewers** Lakshman Yagati, John Blitzer, Phoebe Kirk, and Anand Rao for valuable feedback.

References

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

S. Chen, Y. Zhao, J. Zhang, I.-C. Chern, S. Gao, P. Liu, and J. He. Felm: Benchmarking factuality evaluation of large language models. *arXiv preprint arXiv:2310.00741*, 2023.

G. Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL <https://arxiv.org/abs/2507.06261>.

Gemini Team: R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

O. Honovich, R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansky, V. Cohen, T. Scialom, I. Szpektor, A. Hassidim, and Y. Matias. TRUE: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*, 2022.

A. Jacovi, A. Wang, C. Alberti, C. Tao, J. Lipovetz, K. Olszewska, L. Haas, M. Liu, N. Keating, A. Bloniarz, C. Saroufim, C. Fry, D. Marcus, D. Kukliansky, G. S. Tomar, J. Swirhun, J. Xing,

L. Wang, M. Gurumurthy, M. Aaron, M. Ambar, R. Fellinger, R. Wang, Z. Zhang, S. Goldshtein, and D. Das. The FACTS grounding leaderboard: Benchmarking llms’ ability to ground responses to long-form input, 2025. URL <https://arxiv.org/abs/2501.03200>.

N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel. Large language models struggle to learn long-tail knowledge. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR, 2023. URL <https://proceedings.mlr.press/v202/kandpal23a.html>.

T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2022.

G. Mialon, C. Fourrier, C. Swift, T. Wolf, Y. LeCun, and T. Scialom. GAIA: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023.

OpenAI. GPT-5 System Card. <https://openai.com/research/index/publication/>, August 2025. Accessed on September 12, 2025.

H. Rashkin, V. Nikolaev, M. Lamm, L. Aroyo, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023.

L. Tang, P. Laban, and G. Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. *arXiv preprint arXiv:2404.10774*, 2024.

T. Vu, M. Iyyer, X. Wang, N. Constant, J. Wei, J. Wei, C. Tar, Y.-H. Sung, D. Zhou, Q. Le, and T. Luong. FreshLLMs: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*, 2023.

K. Wataoka, T. Takahashi, and R. Ri. Self-preference bias in LLM-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

J. Wei, N. Karina, H. W. Chung, Y. J. Jiao, S. Papay, A. Glaese, J. Schulman, and W. Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

J. Wei, Z. Sun, S. Papay, S. McKinney, J. Han, I. Fulford, H. W. Chung, A. T. Passos, W. Fedus, and A. Glaese. BrowseComp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

X. Yang, K. Sun, H. Xin, Y. Sun, N. Bhalla, X. Chen, S. Choudhary, R. D. Gui, Z. W. Jiang, Z. Jiang, et al. CRAG-comprehensive RAG benchmark. *Advances in Neural Information Processing Systems*, 37:10470–10490, 2024.