ReFACT: A Benchmark for Scientific Confabulation Detection with Positional Error Annotations

Yindong Wang* **Martin Preiß*** **Margarita Bugueño**
Jan Vincent Hoffbauer **Abdullatif Ghajar** **Tolga Buz** **Gerard de Melo**
Hasso Plattner Institute / University of Potsdam, Germany
{yindong.wang,margarita.bugueno,tolga.buz,gerard.demelo}@hpi.de
{martin.preiss, janvincent.hoffbauer, abdullatif.ghajar}@student.hpi.uni-potsdam.de
*Equal contribution

Abstract

The mechanisms underlying scientific confabulation in Large Language Models (LLMs) remain poorly understood. We introduce **Reddit False And Correct Texts** (ReFACT), a benchmark of 1,001 expert-annotated question–answer pairs with **span-level error annotations** derived from Reddit's *r/AskScience*. Evaluating 9 state-of-the-art LLMs reveals two critical limitations. First, models exhibit a dominant **salient distractor** failure mode: 61% of incorrect span predictions are semantically unrelated to actual errors. Crucially, this pattern persists across all model scales (1B to 70B), indicating a fundamental semantic grounding deficit that **scaling alone fails to resolve**. Second, we find that **comparative judgment** is paradoxically harder than independent detection—even GPT-4o's F_1 score drops from 0.67 to 0.53 when comparing answers side-by-side. These findings directly challenge the reliability of LLM-as-Judge paradigms for scientific factuality. Code and data are released at https://github.com/ddz5431/ReFACT.

1 Introduction

While Large Language Models (LLMs) demonstrate impressive fluency, they remain prone to scientific confabulation—generating content that sounds plausible to non-experts but is fundamentally factually incorrect. Unlike obvious factual errors that can be easily verified, scientific misinformation is often subtle and domain-specific. Such inaccuracies may appear credible and require expert knowledge to detect, presenting a critical challenge for reliable knowledge dissemination in online discourse.

We adopt the term *scientific confabulation* to describe fluent, plausible yet factually incorrect scientific text, which are often also informally referred to as hallucinations. Following clinical usage, confabulation refers to fabricated but coherent content that appears convincing (Sui et al., 2024b). Scientific confabulations are particularly insidious due to their surface fluency and domain-specific language, often requiring expert knowledge to detect (e.g., describing DNA replication using RNA mechanisms; see Figure 1).

To address this gap in the evaluation landscape, we introduce **Reddit False And Correct Texts** (ReFACT), the first benchmark specifically designed to evaluate LLMs' ability to detect, localize, and correct confabulations. ReFACT is constructed from authentic, human-authored scientific discourse in r/AskScience—a community with over 23 million members and rigorous moderation standards. Our contributions include:

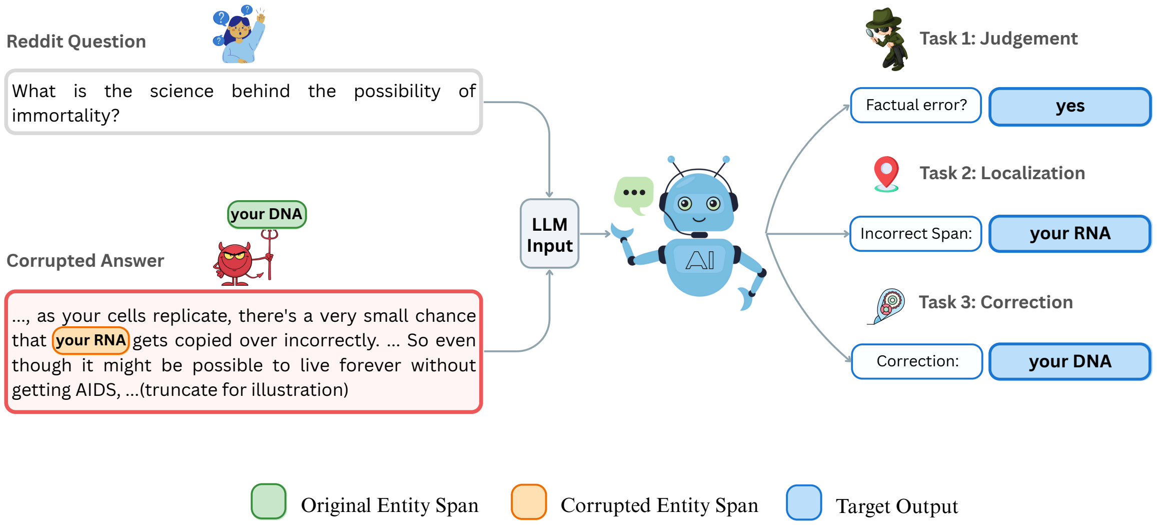
- **Scientific Confabulation Benchmark**: 1,001 question–answer pairs across 10 scientific domains with fine-grained span-level confabulation annotations and error types.
- **Three-Task Evaluation**: (1) binary confabulation detection judgment, (2) span-level localization, and (3) confabulation correction generation.
- **Human-Verified Pipeline**: LLM-assisted corruption of authentic answers with multi-annotator verification ensuring high-quality confabulations.
- **The Salient Distractor Phenomenon**: We identify a scale-invariant failure mode where models fixate on contextually salient terms rather than factual errors. This limitation persists from 1B to 70B parameters, providing empirical evidence that model scaling alone is insufficient for scientific factuality.

2 Related Work

2.1 Factuality, Hallucination & Confabulation

A range of prior work has attempted to define and categorize errors produced by LLMs, but terms such as *hallucination*, *factuality*, and *faithfulness*

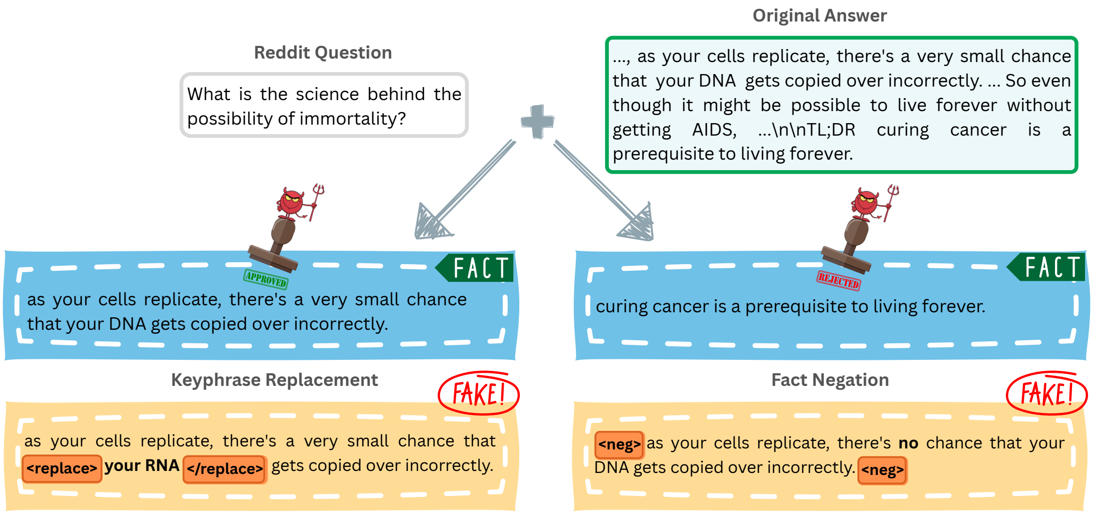
Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8174–8187
March 24-29, 2026 ©2026 Association for Computational Linguistics



What is the science behind the possibility of immortality?

..., as your cells replicate, there's a very small chance that your RNA gets copied over incorrectly. ... So even though it might be possible to live forever without getting AIDS, ...(truncate for illustration)

Figure 1: **Overview of the ReFACT Evaluation Pipeline.** Illustrated via an entity replacement example ("your DNA" → "your RNA"), the benchmark evaluates three capabilities: (1) **Judgment** – detecting the confabulation, (2) **Span Localization** – identifying the corrupted span, and (3) **Correction** – recovering the original entity.



What is the science behind the possibility of immortality?

..., as your cells replicate, there's a very small chance that your DNA gets copied over incorrectly. ... So even though it might be possible to live forever without getting AIDS, ...\n\nTL;DR curing cancer is a prerequisite to living forever.

as your cells replicate, there's a very small chance that your DNA gets copied over incorrectly.

curing cancer is a prerequisite to living forever.

as your cells replicate, there's a very small chance that <replace> your RNA </replace> gets copied over incorrectly.

<neg> as your cells replicate, there's **no** chance that your DNA gets copied over incorrectly. <neg>

Figure 2: **Confabulation Generation Pipeline.** Illustrated via the negation and replacement strategies, the process transforms factual claims through: (1) **Extraction** and **Selection** of salient facts; (2) **Transformation** via logical *negation* or domain-specific *entity replacement* (e.g., "your DNA" → "your RNA"); and (3) **Annotation** of precise error spans.

are often used inconsistently (Ji et al., 2023). *Factuality* typically refers to agreement with external or real-world facts, while *hallucination*, adapted from psychology, describes model outputs that introduce information not grounded in reality. This phenomenon extends beyond simple textual errors (Ji et al., 2023), manifesting in multimodal inconsistencies (Wang et al., 2025) or failures in uncertainty calibration (Cohen et al., 2024). Conversely, *faithfulness* denotes strict consistency with the source input, independent of real-world factual correctness (Maynez et al., 2020; Augenstein et al., 2023).

Recently, the term *confabulation* has gained traction to describe a distinct error category characterized by fluent, coherent, and contextually plausible content that nevertheless contains subtle factual inaccuracies (Sui et al., 2024a). For instance, a model might claim, "*mitochondria produce glucose through cellular respiration*"—a statement plausible to non-experts but is biologically inaccurate. While prior work has explored confabulation in general domains, domain-specific epistemic failures remain largely overlooked. To address this gap, we introduce a benchmark specifically designed for fine-grained evaluation of scientific confabulation.

2.2 Factuality and Hallucination Benchmarks

Research on misinformation has shifted from external social media cues toward detecting hallucinations intrinsic to LLM-generated text (Chen and Shu, 2023). Initial benchmarks prioritized **binary classification**: TruthfulQA (Lin et al., 2022) employs a GPT-3 judge, complemented by prompt-based (Lee et al., 2023) and agentic approaches (Li et al., 2024b). However, the reliability of LLMs as evaluators remains contested. While some studies suggest LLMs excel at verification (Guan et al., 2024; Huang and Sun, 2024), Hu et al. (2024) demonstrate that specialized BERT-based architectures still outperform general-purpose LLMs on binary detection tasks.

To enable fine-grained analysis, recent benchmarks advance toward **span-level localization**. FactScore (Min et al., 2023) assesses atomized statements via retrieval-based scoring. FAVABENCH (Mishra et al., 2024) and HALoGEN (Ravichander et al., 2025) introduce detailed error taxonomies (e.g., contradictions, intent), yet rely on model-generated outputs rather than human authorship. Similarly, HaluEval 2.0 (Li et al., 2024a) and HalluLens (Bang et al., 2025) are limited to synthetic or general-domain content, lacking the complexity of authentic scientific discourse.

3 ReFACT Creation Methodology

Our dataset construction prioritizes **real-world grounding** over synthetic generation. Unlike benchmarks relying on LLM-generated QA pairs (Mishra et al., 2024; Ravichander et al., 2025), we source factual content from Reddit's r/AskScience [1], a community-moderated platform where domain experts provide peer-validated answers to scientific questions. This ensures a verifiable factual basis, while subsequent transformations introduce subtle confabulations that preserve contextual plausibility. Such a design is critical for evaluating a model's ability to identify nuanced factual inaccuracies rather than the obvious artifacts typical of synthetic data.

3.1 Data Collection and Quality Control

Reddit is a popular social media platform organized into topic-specific communities called subreddits. Among them, r/AskScience is strictly moderated and dedicated to scientific inquiry. Community users post questions and receive multiple, threaded responses, which can be up- or downvoted, producing a score generally reflective of perceived content quality. However, Reddit's social dynamics (e.g., humor or topicality) can introduce noise.

Filtering To mitigate this, and following Hoffbauer et al. (2024), we select only top-rated questions (score ≥ 4) and their best-scoring answers. We further retain pairs with 500–1,000 characters to avoid non-descriptive or overly verbose content, yielding 10,282 pairs (statistics in Appendix B).

3.2 Confabulation Generation

We systematically corrupt factual answers using two strategies: **negation** (logical reversal) and **entity replacement** (term substitution). To ensure the generated confabulations remain fluent and contextually plausible while being factually incorrect, we employ `Gemma-2-27B-it` (Team et al., 2024) within a multi-stage prompt pipeline (see Appendix A and Figure 2). Crucially, for entity replacement, we integrate an NLI classifier (Laurer et al., 2022) to filter out synonym substitutions, guaranteeing that the generated entities represent distinct semantic errors rather than mere paraphrases.

3.3 Human Annotation Procedure

Using the `doccano` platform (Nakayama et al., 2018), annotators evaluated each generated sample, considering the triplet $(q_i, a_i, t(a_i))$–where q_i is the question, a_i the original answer, and $t(a_i)$ the transformed answer–against strict validity criteria. To be considered valid, a transformed answer a_i must be: (1) **coherent** and contextually relevant; (2) **factually** incorrect relative to scientific consensus; and (3) **precisely tagged** (e.g., `<replace>`). This rigorous process filters out sarcasm or low-quality generations, ensuring that only subtle, plausible confabulations are preserved. Detailed guidelines are available in our repository [2].

3.4 Results

Transformation Success Rate Defined as the proportion of generations that pass human validation for plausibility and falsity, our pipeline achieved success rates of 58% for negation and 57% for replacement on a batch of size 400. The released dataset is constructed by aggregating multiple batches of generations, each of which under-

[1] https://www.reddit.com/r/AskScience

[2] https://github.com/ddz5431/ReFACT

Benchmark	**Source**	**Samples**	**Length**	**Binary**	**Span**	**Error Type**	**Human?**
SimpleQA (Wei et al., 2024)	Curated questions	4,326	1–3	✓	×	×	✓
TruthfulQA (Lin et al., 2022)	Curated questions	817	10–30	✓	×	×	✓
HaluEval (Li et al., 2023)	Mixed (real/synth)	35,000	~37	✓	×	×	Mixed
HaluEval 2.0 (Li et al., 2024a)	Mixed (real/synth)	8,770	~37	✓	×	×	Mixed
FavaBench (Mishra et al., 2024)	Model outputs	1,000	~30	×	✓	✓	Mixed
HALoGEN (Ravichander et al., 2025)	Model outputs	10,923	~23	×	✓	✓	×
FELM (Chen et al., 2023)	Model outputs	847	~89	×	✓	✓	×
HalluLens (Bang et al., 2025)	Wikipedia	5,000	~40	✓	✓	×	Mixed
ReFACT (This work)	**Reddit**	**1,001**	**130+**	✓	✓	✓	✓

Table 1: **Benchmark Comparison**. ReFACT is the only benchmark with fully human-verified span-level and error-type annotations on long-form scientific QA. Columns: **Samples** (dataset size), **Length** (answer words), **Binary** (binary labels), **Span** (span localization), **Error Type** (error categorization), **Human** (human verified; *Mixed* = partial).

goes strict human filtering and therefore comprises exclusively verified samples.

Annotator Agreement To assess the reliability of our validity criteria, we measured pairwise inter-annotator agreement among three annotators on a random subset of 100 samples in which all three annotators overlapped. The resulting agreement rate is 72.56%, suggesting consistent application of the annotation guidelines.

Resulting Dataset The final dataset comprises 1,001 samples, balanced between negation ($N = 527$) and entity replacement ($N = 474$). Each entry consists of the original question–answer pairs aligned with its. Further statistics on sample length and the distribution of knowledge domains can be found in Appendix B.

4 Benchmarking LLMs with ReFACT

4.1 Models and Implementation

We evaluate instruction fine-tuned versions of `Gemma-3` (1B, 4B, 12B, 27B) (Team et al., 2025) and Llama-3 (1B, 3B, 70B) (Grattafiori et al., 2024) alongside GPT-4o (OpenAI, 2024) as a proprietary reference. All experiments exmploy zero-shot prompting with task-specific templates (details in Appendix C).

4.2 Evaluation Tasks and Metrics

We proposed a structured evaluation framework to assess LLM's capabilities in handling scientific confabulation across three key dimensions: Judgment, Localization, and Correction. Figure 1 illustrates the evaluation pipeline using an entity replacement example.

Task 1: Confabulation Judgment
Input: A pair (q, a) (Independent) or a triplet (q, a_1, a_2) (Comparative).
Output: A predicted character span $\hat{s}$ identifying the error (entity-level for replacement, sentence-level for negation).
Metrics: We report Accuracy and F_1 score for the confabulated class.

Task 2: Confabulation Localization
Input: A question q and a confabulated answer a_{conf}.
Output: A predicted character span $\hat{s}$ identifying the error (entity-level for replacement, sentence-level for negation).
Metrics: We compute the Intersection-over-Union (IoU) between the predicted span $\hat{s}$ and the gold span s^*. Localization is considered accurate if the overlap exceeds 50%:

$$\mathrm{Acc}_{\mathrm{loc}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\frac{|\hat{s}_i \cap s_i^*|}{|\hat{s}_i \cup s_i^*|} \geq 0.5\right] \tag{1}$$

Task 3: Entity Correction
Input: a tuple (q, a_{conf}), where the error span in a_{conf} is explicitly marked (e.g., via `<replace>`).
Output: a corrected entity string $\hat{e}$ **Metric:** Exact Match (EM) accuracy on whitespace-normalized strings:

$$\mathrm{Acc}_{\mathrm{corr}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\mathrm{norm}(\hat{e}_i) = \mathrm{norm}(e_i^*)] \tag{2}$$

We exclusively evaluate entity replacement, as preliminary experiments indicated that reversing logical negation is a trivial task for current models (near-perfect performance).

Aggregate Performance. We report the macro-average accuracy ($\mathrm{Acc}_{\mathrm{avg}}$) across the five sub-tasks: Independent/Comparative Judgment ($\mathrm{J}_{\mathrm{ind}}, \mathrm{J}_{\mathrm{comp}}$), Negation/Entity Localization ($\mathrm{L}_{\mathrm{neg}}, \mathrm{L}_{\mathrm{ent}}$), and Entity Correction ($\mathrm{C}_{\mathrm{ent}}$):

$$\begin{aligned}\mathrm{Acc}_{\mathrm{avg}} = \frac{1}{5}\Big(&\mathrm{Acc}_J^{\mathrm{ind}} + \mathrm{Acc}_J^{\mathrm{comp}} + \mathrm{Acc}_L^{\mathrm{neg}} \\ &+ \mathrm{Acc}_L^{\mathrm{ent}} + \mathrm{Acc}_C^{\mathrm{ent}}\Big)\end{aligned} \tag{3}$$

Metric Selection. For **Judgment**, we report F_1 to balance precision and recall given the class imbalance. For **Localization**, we adopt the standard Intersection-over-Union metric (IoU $\geq$ 0.5) following extractive QA benchmark (Rajpurkar et al., 2016). Crucially, for **Correction**, we enforce **Exact Match (EM)**. Our preliminary analysis showed that semantic similarity metrics (e.g., BERTScore) fail to penalize subtle scientific errors, assigning high scores ($>$ 0.85) to plausible yet factually inverted confabulations.

5 Results and Analysis

General Results Table 2 reveals a prevasive struggle across state-of-the-art LLMs on ReFACT. While `GPT-4o` secures the top performance, its modest aggregate score ($\mathrm{Acc}_{\mathrm{avg}}$ = 0.54) highlights the difficulty of detecting subtle scientific confabulations. Among open-weight models, `Gemma-3-27B` (0.52) remarkably outperforms the much larger `Llama-3.3-70B` (0.43), suggesting that for scientific factuality, data quality or architectural efficiency may outweigh sheer parameter scale. Critically, all models fail to reliably identify specific errors: even GPT-4o achieves less than 50% accuracy on entity localization. To explain this systematic deficienty, we investigate the underlying error patterns below.

5.1 Error Analysis

To diagnose failure mechanisms, we categorized 5,826 false positive predictions from the entity localization task relative to the gold span s^*. We distinguish seven distinct distractor types:

- **adjacent_span**: Spatially proximal (overlapping or within 20 chars of s^*).
- **partial_overlap**: Intersects with s^* but misses key tokens.
- **word_overlap**: Shares surface tokens without semantic alignment.
- **number_date**: Type mismatch (numerical/temporal) against a non-numerical s^*.
- **common_noun**: Generalization error (generic noun vs. specific entity).
- **long_fragment**: Over-extraction ($>$ 10 tokens) containing s^* plus noise.
- **unrelated**: Semantically disconnected from s^*.

The **unrelated** category is further subdivided into: (1) stopwords, (2) named entities, (3) technical terms, and (4) general content.

5.1.1 The Salient Distractor Problem

Semantic Disconnect. Our analysis of 5,826 false positive predictions exposes a critical failure mode: **51.8–67.3%** of localization errors are semantically unrelated to the actual factual contradiction (Figure 3). Instead of identifying errors, models systematically fixate on **salient distractors**-contextually plausible technical terms (12.5%), named entities (8.8%), or generic scientific phrases (35.1%). Crucially, this is not a boundary detection issue but a foundamental grounding failure. IoU analysis reveals a stark **bimodal distribution**: 90.6% of these false positives share *exactly zero* character overlap with the gold span. Models are not imprecisely locating the error; they are confidently pointing to entirely wrong regions driven by surface-level keyword prominence.

Invariance to Scale and Prompting. This semantic deficit proves remarkably resistant to standard scaling and prompting interventions:

- **Scaling Limitation:** While unrelated errors decrease slightly from 1B (67%) to 70B (61%) models, even `Llama-3.3-70B` remains dominated by this failure mode. Larger models merely become better at selecting "scientific-sounding" distractors without improving factual discrimination.
- **Prompting Failure:** As shown in Table 3, advanced strategies fail to mitigate this issue. **Zero-shot prompting consistently outperforms** Few-Shot and Chain-of-Thought (CoT) approaches

Model	Ind. Judg. (Acc / F_1)	Comp. Judg. (Acc / F_1)	Neg. Loc. (Acc / IoU)	Ent. Loc. (Acc / IoU)	Ent. Corr. (Acc)	Avg. Acc.
Gemma-3-1B	0.51 / 0.53	0.53 / 0.34	0.13 / 0.13	0.10 / 0.10	0.00	0.25
Gemma-3-4B	0.58 / 0.60	0.52 / 0.33	0.36 / 0.33	0.38 / 0.25	0.10	0.39
Gemma-3-12B	0.65 / 0.63	**0.71** / 0.48	0.44 / 0.44	0.46 / 0.26	0.19	0.49
Gemma-3-27B	**0.71** / 0.72	0.56 / 0.54	0.61 / 0.46	0.46 / 0.29	0.24	0.52
Llama-3.2-1B	0.49 / 0.39	0.50 / 0.27	0.00 / 0.00	0.08 / 0.04	0.01	0.22
Llama-3.2-3B	0.52 / 0.34	0.48 / 0.29	0.02 / 0.02	0.04 / 0.08	0.02	0.22
Llama-3.3-70B	0.67 / **0.73**	0.50 / 0.39	**0.69** / **0.61**	0.13 / 0.24	0.16	0.43
GPT-4o-mini	0.62 / 0.52	0.59 / **0.55**	0.59 / 0.54	0.07 / 0.20	0.21	0.42
GPT-4o	0.67 / 0.67	0.60 / 0.53	0.66 / 0.57	**0.47** / **0.38**	**0.28**	**0.54**

Table 2: Metrics are reported as Accuracy / F_1 or Accuracy / IoU, depending on the task. The final column shows the average of **accuracy** scores only (excluding F_1 and IoU). **Bold** indicates the best-performing model in each metric column.

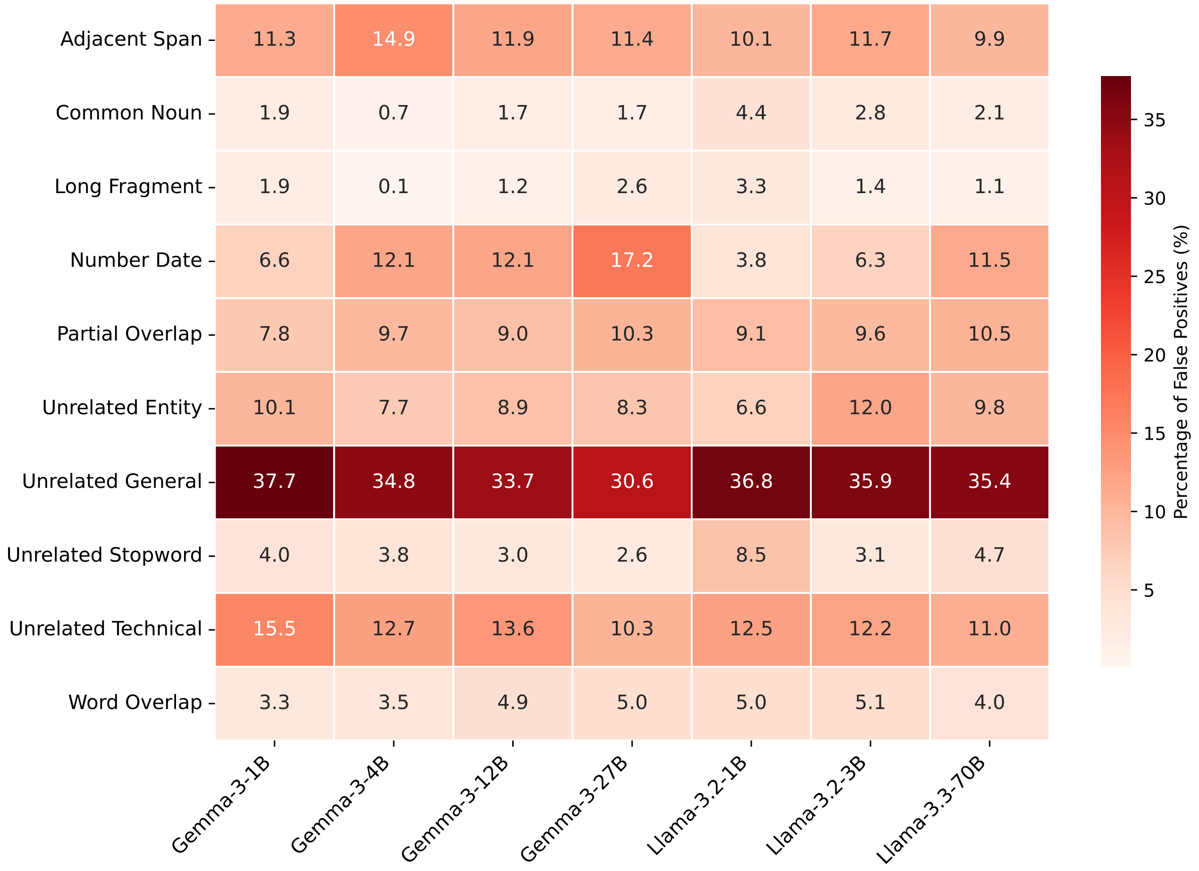


Figure 3: **The Salient Distractor Phenomenon.** Analysis of 5,826 false positives reveals a systematic failure in semantic grounding: **61 % of localization errors** focus on contextually salient but semantically unrelated terms, rather than actual factual errors. Crucially, this pattern is **scale-invariant**, persisting from 1B to 70B models. This confirms that current LLMs rely on surface-level saliency heuristics rather than true factuality detection. For instance, when models fail to detect the error in “DNA replication”, they often wrongly highlight “RNA” simply because it is a biological term, ignoring the actual semantic contradiction.

across all scales (e.g., `Gemma-3-27B` drops from 0.71 F_1 zero-shot to 0.64 with CoT).

This counter-intuitive finding suggests that the bottleneck is not instruction adherence or reasoning depth, but the underlying representation of scientific truth. Models cannot "reason" their way out of a knowledge deficit.

Transformation Agnosticism. Finally, error patterns remain consistent across transformation types

(Negation: 62.6% unrelated vs. Entity Swap: 61.0%), indicating that the failure is agnostic to the logical complexity of the error. However, subtle nuances exist: Entity Swaps elicit more unrelated entity predictions (10.3% vs. 6.4%) as models grasp for alternative nouns, while Negations show higher partial overlap (20.1%), suggesting models can vaguely sense the region of logical reversal but lack the precision to pinpoint the operator.

Model	Zero-Shot	Few-Shot	Few-Shot + CoT
Gemma-3-1B	0.504	0.455	0.432
Gemma-3-4B	0.584	0.561	0.517
Gemma-3-12B	0.653	0.629	0.612
Gemma-3-27B	0.707	0.704	0.640
Llama-3.2-1B	0.475	0.416	0.463
Llama-3.2-3B	0.481	0.412	0.487
Llama-3.3-70B	0.651	0.653	0.644

Table 3: **Prompting Strategy Ablation for Independent Judgment.** F_1 scores across prompting techniques. Zero-shot consistently matches or outperforms few-shot and chain-of-thought prompting, indicating the task difficulty stems from semantic grounding limitations rather than instruction-following challenges.

5.2 Task-Specific Performance Analysis

Confabulation Judgment Detecting confabulations remains challenging even for state-of-the-art models. As shown in Table 2, smaller models (Llama-3.2-1B: 0.49, Gemma-3-1B: 0.51) perform near random chance (0.50) on Independent Judgment. Larger models show improvement: Llama-3.3-70B achieves 0.67/0.73 (Acc/F_1) and Gemma-3-27B reaches 0.71/0.72, but still fall short of reliable detection.

Contrary to expectations, Comparative Judgment appears to be even more difficult for most models. Across the board, models exhibit lower F_1 scores on Comparative Judgment than on Independent Judgment, despite the task involving a relative comparison between two answers. For example, Llama-3.3-70B drops from 0.73 to 0.39 in F_1, and GPT-4o drops from 0.67 to 0.53. This suggests that when both answers are plausible yet subtly different, models often struggle to discern which one contains a confabulation, potentially due to shallow comparative heuristics or overconfidence in surface-level coherence. Detailed precision, recall, and F_1 scores on original and confabulated answers are given in Table 4.

Implications for LLM-as-Judge Paradigms These findings challenge the widespread use of LLMs as evaluators in research. Many recent benchmarks employ GPT-4o or similar models to assess other systems, assuming reliable error detection capabilities. However, our results show that even GPT-4o achieves only 0.60 accuracy on Comparative Judgment, barely above random for binary classification. This poor performance, particularly in scientific contexts requiring domain expertise, suggests LLM-as-Judge paradigms may produce unreliable evaluations for factuality and reasoning tasks. Researchers should reconsider this evaluation methodology, or, at a minimum, report confidence intervals and baseline comparisons.

Confabulation Localization The localization results show that LLMs struggle to accurately identify the span that contains a confabulation, particularly in the case of entity replacements. Across all models, performance on negation localization is substantially higher than on entity localization. For example, GPT-4o achieves 0.66 accuracy and 0.57 IoU for negation, whereas its performance drops to 0.47/0.38 for entity replacements. Similarly, Llama-3.3-70B achieves a respectable 0.69/0.61 on negation but only 0.13/0.24 on entity localization, with smaller models often scoring to zero.

This disparity highlights a key challenge: While negated claims may be easier to catch due to their syntactic structures, identifying subtle entity substitutions, especially technical or domain-specific ones, requires deeper semantic understanding that most models seem to lack. Among open-weight models, Gemma-3-27B shows the strongest performance on entity localization (0.46/0.29), yet still falls short of robust detection.

Correction The correction task remains the most challenging task for all evaluated models. Even the best-performing model, GPT-4o, achieves only 0.28 accuracy, far from indicating reliable correction ability. Among open-weight models, Gemma-3-27B model shows the strongest performance at 0.44, while Llama-3.3-70B falls behind at 0.23. Smaller models perform drastically worse, for instance, Llama-3.2-1B achieves only 0.05, and Gemma-3-1B drops to 0.03.

Although we computed BERTScore values, we omit them from our primary results. This metric often yields inflated scores for factually incorrect predictions—e.g., BERTScore exceeds 0.85 for multiple models despite very low EM accuracy, making it unsuitable for evaluating factual correction.

6 Conclusion

We identify **salient distractor selection** as a dominant, scale-invariant failure mode in LLM confabulation detection. Across 5,826 false positives, 61% target semantically unrelated spans—90.6% with zero character overlap to actual errors. This pattern persists from 1B to 70B parameters (67% → 61% unrelated), across transformation types (negation: 62.6%, entity-swap: 61.0%), and resists prompting interventions, exposing a fundamental semantic grounding deficit that scaling alone cannot resolve.

Comparative judgment proves paradoxically harder than independent detection: GPT-4o's F_1 drops from 0.67 to 0.53 when evaluating answers side-by-side. Entity errors pose particular difficulty—localization accuracy falls to 0.47 versus 0.66 for negations, and correction reaches only 0.28 even for GPT-4o. These results directly challenge LLM-as-Judge paradigms for scientific factuality.

Future work should target these failure modes through contrastive training with hard negatives, hierarchical localization objectives, and explicit entity grounding mechanisms. We aim to extend ReFACT across domains to deepen understanding of LLM factuality.

Limitations

Data Constraints & Scope. Our benchmark relies on community-curated scientific discourse (r/AskScience). While we mitigate noise by filtering for high-consensus answers, latent crowd-sourced inaccuracies may remain. Furthermore, we strictly target fine-grained, *span-level* confabulations; this design offers precise error localization but does not address broader, document-level thematic contradictions.

Annotation Ambiguity. Distinguishing subtle confabulations from facts entails inherent subjectivity. We address this by enforcing a three-way annotator consensus protocol, prioritizing high inter-annotator agreement over dataset scale.

Evaluation Bias & Metrics. We acknowledge potential *self-preference bias* when models evaluate their own generations. To minimize this, we maximize architectural diversity (e.g., Llama, GPT, Gemma) between generation and evaluation stages. Finally, as automatic metrics remain imperfect proxies for human judgment, we interpret results as relative performance trends rather than absolute measures of truth.

Ethics Statement

Selection Bias & Content Quality. We acknowledge inherent selection biases in sourcing from Reddit (r/AskScience), where visibility correlates with popularity rather than purely scientific merit. To mitigate this, our annotation guidelines strictly filter for objective scientific validity, excluding anecdotal, humorous, or institutionally-biased responses. While we employ automated and human filtering to remove toxic or inappropriate language, the dataset reflects the linguistic characteristics of public internet discourse.

Privacy & Data Compliance. Data collection utilized the public Pushshift API, consistent with Reddit's Terms of Service. As individual consent is infeasible at scale, we adhere to a strict non-identification policy: all user identifiers were removed, and content was screened to eliminate Personal Identifiable Information (PII). The released dataset contains exclusively scientific explanations, posing minimal risk to original content creators.

Acknowledgments

This research was funded in part by the German Federal Ministry of Research, Technology and Space under the funding code "KI-Servicezentrum Berlin-Brandenburg" 16IS22092. Responsibility for the content remains with the authors.

References

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2023. Factuality challenges in the era of large language models. *Preprint*, arXiv:2310.05189.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM hallucination benchmark. *Preprint*, arXiv:2504.17550.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of LLMs: Opportunities and challenges. *Preprint*, arXiv:2311.05656.

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. *Preprint*, arXiv:2310.00741.

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. 2024. I don't know: Explicit modeling of uncertainty with an [IDK] token. In *Proceedings of 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. *Preprint*, arXiv:2310.14564.

Jan Hoffbauer, Sylwester Sawicki, Marc Ulrich, Tolga Buz, Konstantin Dobler, Moritz Schneider, and Gerard De Melo. 2024. Knowledge acquisition through continued pretraining is difficult: A case study on r/AskHistorians. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 96–108, Bangkok, Thailand. Association for Computational Linguistics.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:22105–22113.

Yue Huang and Lichao Sun. 2024. FakeGPT: Fake news generation, explanation and detection of large language models. *Preprint*, arXiv:2310.05046.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Factuality enhanced language models for open-ended text generation. *Preprint*, arXiv:2206.04624.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024a. The dawn after the dark: An empirical study on factuality hallucination in large language models. *Preprint*, arXiv:2401.03205.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Xinyi Li, Yongfeng Zhang, and Edward C. Malthouse. 2024b. Large language model agent for fake news detection. *Preprint*, arXiv:2405.01593.

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

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1919.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

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.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

OpenAI. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. 2025. HALoGEN: Fantastic LLM hallucinations and where to find them. *Preprint*, arXiv:2501.08292.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard So. 2024a. Confabulation: The surprising value of large language model hallucinations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14274–14284, Bangkok, Thailand. Association for Computational Linguistics.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard Jean So. 2024b. Confabulation: The surprising value of large language model hallucinations. *Preprint*, arXiv:2406.04175.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Weixing Wang, Zifeng Ding, Jindong Gu, Rui Cao, Christoph Meinel, Gerard de Melo, and Haojin Yang. 2025. Image tokens matter: Mitigating hallucination in discrete tokenizer-based large vision-language models via latent editing. In *Advances in Neural Information Processing Systems (NeurIPS) 2025*.

Jason Wei, Karina Nguyen, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

A Transformation Prompts

For reproducibility, we list the prompts used for the transformations described in Section 3.

A.1 Facts Extraction

Extract all facts mentioned in the following text that are not opinions or beliefs. Be as concise as possible. If there are no factual statements, return “<bad-example>” Otherwise, return the facts enclosed by “<fact>” and “</fact>”. Don’t add any explanations

A.2 Negation

Given a paragraph and a specific sentence from the paragraph, negate that sentence. The negation shall not contradict the paragraph but be factually wrong. Negate ONLY the specific sentence. Output ONLY the negated sentence. No Explanation.

A.3 Replacement: Select Keyphrase

Return a key phrase from the input text, which is not part of the question. This key phrase should:

- Contain factual information or be a technical term.
- Be hard to detect by individuals not educated in the domain, preferably specialized terms or phrases less familiar to the general public.
- Not be inside of the question.

Only return the original key phrase from the input text in list form. Do not provide any explanation.

Example 1:
Question: Where is the Reichstag?
Input Text: The Reichstag is in Berlin, Germany.
Output: [“Berlin”]

Example 2:
Question: Could you name me a deadly virus?
Input Text: Covid-19 is a deadly virus. To the public, it is more commonly known as Corona or SARS-CoV-2.
Output: [“SARS-CoV-2”]

Example 3:
Question: What is the Hasso-Plattner-Institute in Potsdam called?
Input Text: It is called Hasso-Plattner-Institute. The Hasso-Plattner-Institute was founded by Hasso Plattner.
Output: [“Hasso Plattner”]

Question:{question_context}
Input Text: {text}
Output:

A.4 Replacement: Coreference Resolution

You are a Coreference Resolution detector. Detect where the exact keyphrase is referred to as the same entity in the input text, excluding pronouns. Instructions:

- Include only exact matches or direct references to the keyphrase within the context.
- Do not include unrelated mentions or vague references.
- Output should list parts of the text where the entity is mentioned again, maintaining the same context.
- Do not include any pronouns.

Output: List format without explanations.

Example 1
Input Text: Covid 19 is a deadly virus. To the public it is more known as Corona or Sars Covid.
Input Keyphrase: Corona
Output: [“Covid 19” , “Sars Covid”]

Example 2
Input Text: The Hasso Plattner Institute was founded by Hasso Plattner. He also founded SAP.
Input Keyphrase: Hasso Plattner Institute
Output: [’Hasso Plattner Institute’]

Example 3
Input Text: Alcohol is dangerous for humans.
Input Keyphrase: Alcohol
Output: []

Example 4
Input Text: The Amazon rainforest is known for its biodiversity.
Input Keyphrase: biodiversity
Output: []

Example 5
Input Text: My head hurts because of a headache. That is bad.
Input Keyphrase: headache
Output: []

Input Text: text
Input Keyphrase: keyphrase
Output:

A.5 Replacement: Suggest new Keyphrases

Create false information by changing the tagged substrings. Ensure the rewritten answer to be coherent, convincing and false. For preprocessing reasons only return the altered substrings in list form. No explanation.

Example 1:
Input: The <replace>Hasso Plattner Institute</replace> for Digital Engineering gGmbH (German: <replace>Hasso-Plattner-Institut</replace> fuer Digital Engineering gGmbH; <replace>HPI</replace>) is an information technology non-profit company affiliated with the University of Potsdam in Potsdam. The teaching and research of <replace>HPI</replace> are focused on ’IT-Systems Engineering’. <replace>HPI</replace> was founded in 1998, and is the first, and as of 2018, the only entirely privately funded faculty[3] in Germany. It is financed entirely through private funds donated by billionaire <replace>Hasso Plattner</replace>,[4] who co-founded the software company <replace>SAP</replace>, and is currently the chairman of <replace>SAP</replace>’s supervisory board. In addition to Hasso Plattner and Christoph Meinel, the management of <replace>HPI</replace> was expanded to include other board members on 2019.
Tagged Substrings: [’Hasso Plattner Institute’, ’Hasso-Plattner-Institut’, ’HPI’, ’HPI’, ’HPI’, ’Hasso Plattner’, ’SAP’, ’SAP’, ’HPI’]

	Independent Judgment				**Comparative Judgment**			
Model	**Acc**	**Prec**	**Rec**	**F$_1$ (C/O)**	**Acc**	**Prec**	**Rec**	**F$_1$ (C/O)**
Gemma-3-1B	0.51	0.51	0.56	0.53 / 0.48	0.53	0.36	0.35	0.34 / 0.34
Gemma-3-4B	0.58	0.58	0.62	0.60 / 0.57	0.52	0.37	0.36	0.33 / 0.33
Gemma-3-12B	0.65	**0.68**	0.58	0.63 / **0.68**	**0.71**	0.48	0.48	0.48 / 0.48
Gemma-3-27B	**0.71**	0.69	0.75	0.72 / 0.69	0.56	**0.55**	**0.55**	0.54 / 0.54
Llama-3.2-1B	0.49	0.48	0.33	0.39 / 0.56	0.50	0.32	0.33	0.27 / 0.27
Llama-3.2-3B	0.52	0.54	0.25	0.34 / 0.62	0.48	0.33	0.32	0.29 / 0.29
Llama-3.3-70B	0.67	0.62	**0.90**	**0.73** / 0.57	0.50	0.52	0.50	0.39 / 0.39
GPT-4o-mini	0.62	0.70	0.41	0.52 / 0.68	0.59	0.65	0.59	**0.55** / **0.55**
GPT-4o	0.67	0.67	0.66	0.67 / 0.67	0.60	**0.73**	0.59	0.53 / 0.53

Table 4: **Detailed Judgment Metrics.** Evaluation of Independent vs. Comparative Judgment tasks. We report Accuracy (Acc), Precision (Prec), Recall (Rec), and F$_1$ scores. For F$_1$, we report values for both the Confabulated (C) and Original (O) classes (format: C / O). **Bold** indicates the best performance per column.

Output: ['Mark Zuckerberg Institute', 'Mark-Zuckerberg-Institut', 'MZI', 'MZI', 'MZI', 'Mark Zuckerberg', 'Facebook', 'Meta', 'MZI']

Example 2:
Input: Same reason you can feel someone else's pain when they get hurt, e.g. Guys can feel another guy's pain when he gets kicked in the testes. We have <replace>mirror neurons</replace> in our brain. When you do a certain activity or feel a certain emotion, a network of neurons in your brain fire. Now when you see someone who is doing this activity or going through this emotional experience, a subset of those neurons fire in your brain! This probably evolved so that we can feel empathy and relate to what other people are going through, since this is important for living in large societies like humans do.
Tagged Substrings: ['mirror neurons']
Output: ['echo neurons']

Example 3:
Input: The circle (it's not a spiral) only occurs above the <replace>two rotational poles</replace>. That circle is caused by the Earth rotating. Those star tracks each occupy about 50 degrees or so, so we can estimate that the photo was taken using about a 3.5 hour exposure. You should be able to get a similar photo anywhere on the planet, although the closer to the <replace>equator</replace> you are, the closer the rotation centre will be to the horizon.
Tagged Substrings: ['two rotational poles', 'equator']
Output: ['equator', 'two rotational poles']

Example 4:
Input: I'm not quite sure what you mean here. However, the most common way of transmitting COVID-19 is through <replace>the air</replace>.
Tagged Substrings: ['the air']
Output: ['contaminated surfaces']

Example 5:
Input: Because the stability and binding of the nucleus depends on the neutrons just as much as it does the protons. If you have too many or too few neutrons for a given number of protons, you'll no longer have a bound system. The nucleus will break apart on timescales characteristic of the strong force (10**(-22) seconds). The boundaries between bound nuclei and unbound nuclei described above are called the *<replace>driplines</replace>*.
Tagged Substrings: ['driplines']
Output: ['stabilitylines']

Input: {question_context + tagged_answer}
Tagged Substrings:{keyphrases}
Output:

A.6 Most Convincing

You will get a list of {len(texts)} paragraphs manipulated to be factually incorrect. Return the index of the most convincing and coherent paragraph, which is still factually incorrect. No explanation.

B Dataset Statistics

The figures in this section illustrate key dataset statistics, providing insight into its structure and diversity. Figure 4 visualizes the distribution of dataset domains, highlighting its thematic range. Figure 5 and Figure 6 plot the distribution of word and character counts before and after the transformation, calculated using the NLTK library for tokenization (Bird et al., 2009). Lastly, Figure 7 presents statistics for the r/AskScience subset that was used to create the benchmark.

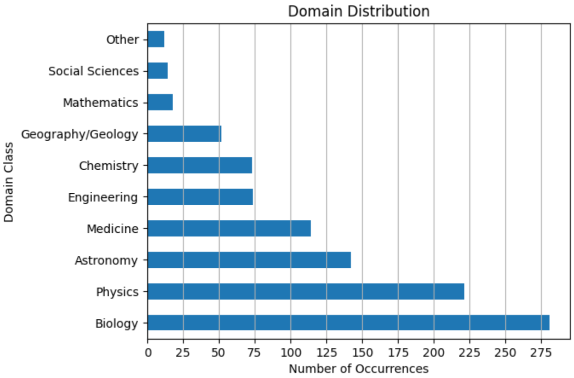


Figure 4: Domains of the Dataset

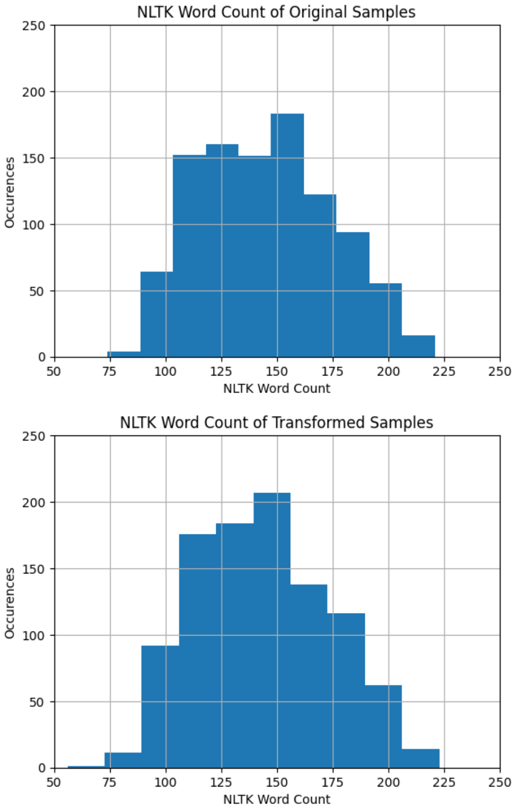


Figure 5: Word Count of Dataset Samples

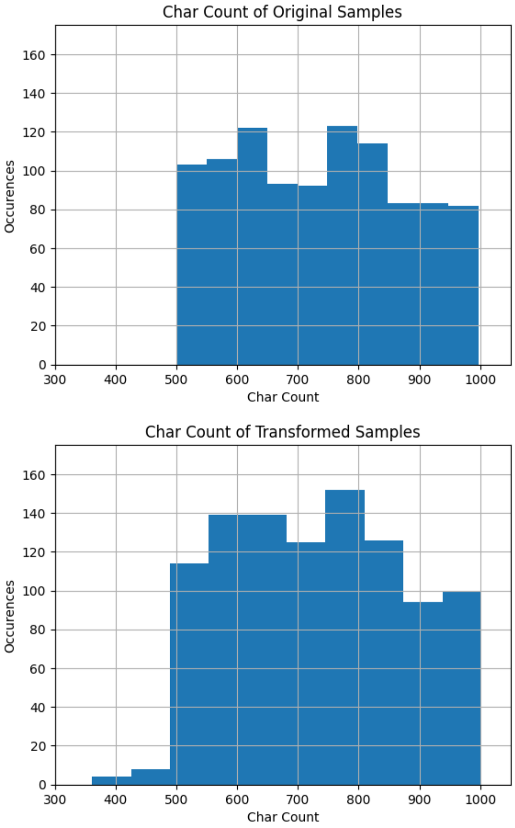


Figure 6: Character Count of Dataset Samples

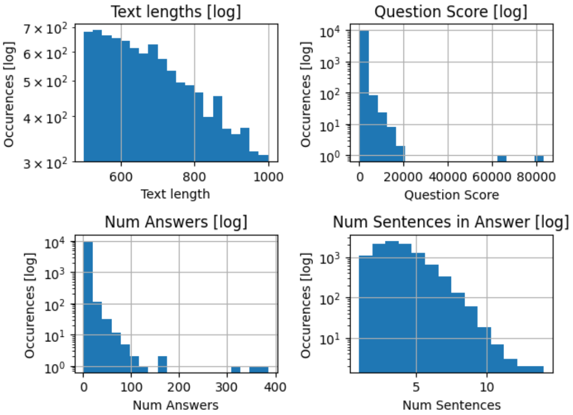


Figure 7: Statistics of the r/AskScience sample that was used as a basis for our dataset.

C Experiment Prompts

In the following, we list the prompts used in the experiments described in Section 4.

C.1 Independent Confabulation Judgment

System Prompt:

Your task is to evaluate the factual correctness of a given answer to a question. Read the entire answer and question carefully, then analyze its factual accuracy step by step. After your analysis, provide a final verdict of either “True” if the entire answer is factually correct, or “False” if any part of the answer contains inaccuracies or hallucinations.

User Prompt:

```
Task:
Question: {question}
Answer: {answer}
Final Verdict:
```

C.2 Comparative Confabulation Judgment

System Prompt

Your task is to return the factually correct answer out of the two given answers (A and B) to a question. Read both entire answers and the question carefully, then analyze the factual accuracy of both answers within the context. After your analysis, provide a final verdict of either answer A or answer B is factually correct.

User Prompt

```
Question: {question}
Answer A: {answer_a}
Answer B: {answer_b}
Final Verdict:
```

C.3 Negation Confabulation Localization

System Prompt

You will get an answer to a question with one factually wrong sentence inside the answer, which was changed beforehand. Your task is to locate the factually wrong sentence of the fake answer to the question. Read the entire answer with the factually wrong sentence and the corresponding question carefully. Then analyze the factual accuracy of every part in the given answer. After your analysis, return only the whole sentence without changes.

User Prompt

```
Question: {question}
Answer: {transformed_answer}
Wrong Sentence:
```

C.4 Entity Confabulation Localization

System Prompt

You will get an answer to a question with factually wrong entities inside the answer, which were changed beforehand. An entity can be a single word or multiple words of any type. Your task is to locate the factually wrong entities of the fake answer to the question. Read the entire answer with factually wrong entities and the corresponding question carefully. Then analyze the factual accuracy of every part in the given answer with the focus on factually wrong entities. After your analysis, return the factually wrong entities separated with newlines without changes.

User Prompt

```
Question: {question}
Answer: {transformed_answer}
Wrong Entities:
```

C.5 Entity Confabulation Correction

System Prompt

Your task is to return replacements for the <mask> tags inside an answer to a question. Read the entire answer and question carefully, then analyze the answer and think about possible replacements. After your analysis, return only the list of replacements in the order they appear separated by new line.

User Prompt

```
Task:
Question: {question}
Answer: {answer}
{number_of_masks} Replacements expected
Replacements:
```