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

Anonymous ACL submission

Abstract

Large Language Models (LLMs) frequently confabulate scientific facts, but the mechanisms underlying these failures 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**, enabling fine-grained analysis of confabulation detection, localization, and correction. Evaluating 9 state-of-the-art LLMs reveals two fundamental limitations. First, models exhibit a dominant **salient distractor** failure mode: 61% of incorrect span predictions are semantically unrelated to actual errors, indicating models fixate on contextually prominent terms rather than true error locations. This pattern persists across model scales (1B to 70B), suggesting scaling alone is insufficient to address this limitation. Second, **comparative judgment** (selecting which of two answers contains factual confabulations) proves fundamentally harder than **independent judgment** (classifying a single answer as confabulated or not). For example, even GPT-4o's F_1 score drops from 0.67 to 0.53 when comparing factual versus confabulated answers side-by-side. This dramatic performance degradation challenges the reliability of LLM-as-Judge paradigms widely adopted in current benchmarks. Code and data are released on <anonymized>[1].

1 Introduction

As Large Language Models (LLMs) increasingly mediate access to scientific and medical information, their tendency to generate plausible but false content poses significant risks. 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. 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's 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 span-level annotations and error types.
- **Three-Task Evaluation**: (1) binary judgment, (2) span-level localization, and (3) correction generation.
- **Human-Verified Pipeline**: LLM-assisted corruption of authentic answers with multi-annotator verification ensuring high-quality confabulations.
- **Analysis of Model Failures**: Benchmarking 9 LLMs reveals best performance at ∼54% accuracy. Fine-grained analysis of 5,826 false positive localization predictions show 61% are semantically unrelated to gold error spans, indicating a limitation persisting across model scales and error types.

[1] We provide an anonymized link to the dataset for the reviewers: https://storage.googleapis.com/dataset-refact/refact-dataset.jsonl

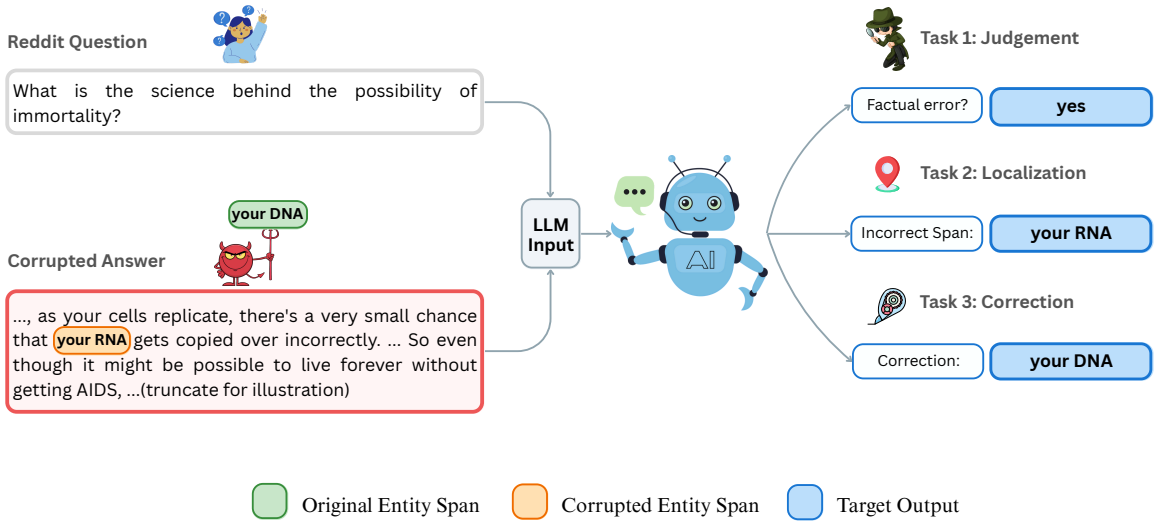

Figure 1: **Overview of the ReFACT Evaluation Pipeline** (Entity Replacement Example). Given a factual Reddit answer and a minimally transformed counterpart containing a subtle instance of *scientific confabulation* ("your DNA" → "your RNA"), the model is evaluated on: (1) **Judgment** – detecting confabulation, (2) **Span Localization** – identifying the corrupted entity span, and (3) **Correction** – recovering the original entity span. The figure illustrates a representative case from the entity replacement subset of ReFACT.

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* 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 or that contradict the input. In contrast, *faithfulness* is defined as strict consistency with the source input: a model is unfaithful if it generates content not directly supported by the input, regardless of factual correctness (Maynez et al., 2020; Augenstein et al., 2023).

Recently, *confabulations* have gained traction describing an error subset of fluent, coherent, and contextually plausible content that nevertheless contains subtle factual inaccuracies (Sui et al., 2024a). For example, a model might claim, "*mitochondria produce glucose through cellular respiration*", which sounds plausible to non-experts but is biologically inaccurate. Notably, most prior work on confabulation has focused on general patterns, with limited attention to domain-specific epistemic failures in scientific contexts. To address this gap, we introduce a benchmark specifically designed for fine-grained evaluation of scientific confabulation.

2.2 Factuality and Hallucination Benchmarks

The emergence of hallucinations in LLM-generated text has shifted misinformation detection research beyond traditional settings focused on fake news and social media diffusion, toward methods that operate without reliance on external cues (Chen and Shu, 2023). While early approaches relied on binary classification of texts as factual or not, real-world scenarios increasingly demand fine-grained detection in complex, user-generated content.

Early benchmarks focused on binary factuality evaluation. TruthfulQA (Lin et al., 2022) provides curated questions to evaluate LLM factuality using a fine-tuned GPT-3 judge, while recent work explores prompt-based detection (Lee et al., 2023) and agent-based workflows (Li et al., 2024b). While some studies argue that LLMs can detect misinformation (Guan et al., 2024; Huang and Sun, 2024), Hu et al. (2024) demonstrate that trained BERT-based architectures outperform current LLMs on binary detection tasks.

Recent benchmarks increasingly provide span-level localization and error categorization to enable fine-grained analysis. *FactScore* (Min et al., 2023) decomposes text into individual statements for separate factuality assessment, combining LLM outputs with retrieval-based scoring. *FAVABENCH* (Mishra et al., 2024) categorizes factual errors into contradictions, unverifiable claims, and fabrications with span-level annotations, while

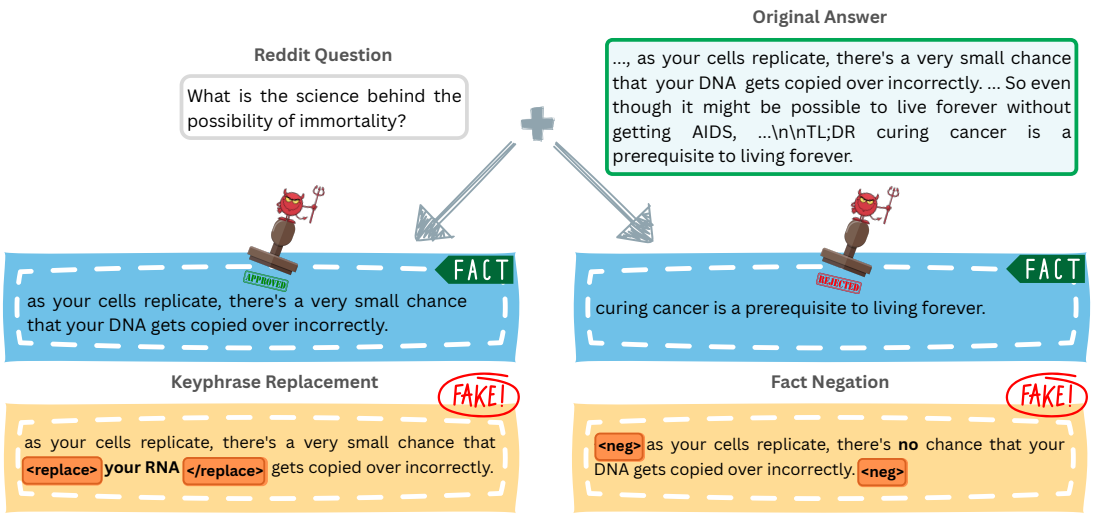

Figure 2: **Confabulation Generation Pipeline** (Data Creation Process). Given a filtered Reddit question–answer pair, factual statements are extracted and systematically corrupted through: (1) **Fact Extraction** – identifying factual claims from the original answer; (2) **Selection** – selecting the most convincing fact for negation or a keyphrase for entity replacement; (3) **Transformation** – applying either *negation* (flipping the factual assertion) or *entity replacement* (substituting technical terms, e.g., "your DNA" → "your RNA"); (4) **Annotation** – marking all generated confabulations with precise span-level error tags. The figure shows the systematic pipeline used to generate span-annotated confabulated answers for the ReFACT benchmark.

HALoGEN (Ravichander et al., 2025) analyzes hallucination intent and verifiability. However, both FAVABENCH and HALoGEN rely on model-generated outputs rather than authentic human-authored content, potentially limiting their ability to capture realistic error patterns. Similarly, *HaluEval 2.0* (Li et al., 2024a) and *HalluLens* (Bang et al., 2025) provide span-level annotations but use synthetic or Wikipedia-based content rather than domain-specific scientific discourse.

3 ReFACT Creation Methodology

Our dataset construction prioritizes **real-world grounding** over synthetic generation. Unlike benchmarks that prompt LLMs to generate question–answer pairs (Mishra et al., 2024; Ravichander et al., 2025), we source factual content from r/AskScience, a community-moderated platform where domain experts provide peer-validated answers to scientific questions. This approach ensures our base data reflects authentic scientific discourse with verifiable factual grounding, while subsequent transformations introduce subtle confabulations that preserve contextual plausibility. This is critical for evaluating models' ability to identify nuanced factual inaccuracies rather than obvious synthetic artifacts.

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[2] is particularly popular. It is dedicated to scientific inquiries, where users post questions and receive multiple, threaded responses. Additionally, contributions are up- or downvoted, producing a score generally reflective of perceived content quality. However, although r/AskScience maintains high moderation standards, given Reddit's social dynamics, factors like humor or topicality may influence voting.

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

3.2 Confabulation Generation

We systematically corrupt factual answers using two strategies: **negation** (reversing logical polarity) and **entity replacement** (substituting key terms). Transformations preserve contextual plausibility, structural similarity, and internal consistency while introducing factual errors mimicking subtle misinformation that resists surface-level detection.

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

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).

Using Gemma-2-27B-it (Gemma-Team and Deepmind, 2024), we apply multi-step prompted transformations (see Appendix A). Each transformation produces a modified answer with tagged error spans. Figure 2 illustrates the pipeline. Both methods follow: (1) extract factual sentences, (2) identify target content, (3) generate corruption, and (4) select the most plausible variant. For entity replacement, we additionally resolve coreferences and filter synonyms using an NLI classifier (Laurer et al., 2022) to ensure semantic incorrectness.

3.3 Human Annotation Procedure

Annotators reviewed each sample using the doccano platform (Nakayama et al., 2018), considering the question q_i, the original answer a_i, and the transformed answer $t(a_i)$. High-quality annotation requires carefully crafted guidelines, which we found critical for consistent labeling and effective filtering of low-quality or sarcastic responses.

A sample is considered valid if all transformations are explicitly tagged (e.g., <replace>...</replace>) and the resulting answer is coherent, relevant to the question, yet factually incorrect. We provide detailed annotation guidelines and examples online[3].

3.4 Results

Transformation Success Rate We define *transformation success rate* as the proportion of generated samples meeting the validity criteria. Following iterative prompt refinement, the final batch of generation (with $n = 400$ samples) achieved an overall success rate of 58% for negation and 57% for replacement transformations.

Annotator Agreement To assess the reliability of our validity criteria, we measured the pairwise inter-annotator agreement across three annotators. The resulting agreement rate is 72.56%, suggesting consistent application of the annotation guidelines.

Resulting Dataset We carefully curated a total of 1,001 valid samples, with 527 negation-based transformations and 474 entity replacements. Each sample contains the original question and answer, combined with an incorrect one, with every change tagged. 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 (Team et al., 2025) (1B, 4B, 12B, 27B) and Llama-3 (Touvron et al., 2023) (1B, 3B, 70B) plus GPT-4o (et al., 2024) via OpenAI API. All tasks use zero-shot prompting with task-specific templates (Appendix C) and are evaluated independently.

4.2 Evaluation Tasks and Metrics

We propose a structured evaluation framework to assess LLMs' capabilities in handling scientific confabulation across three key dimensions: Judgment, Localization, and Correction, each targeting a distinct capability. Figure 1 illustrates the evaluation pipeline using an entity replacement example.

[3] <anonymized annotation guidelines>

For each task, we formally define inputs, outputs, and metrics.

Task 1: Confabulation Judgment where the **input** corresponds to a question q and an answer a (Independent Judgment), or a question q with two answers a_1, a_2 (Comparative Judgment). The **output** is a binary label $\hat{y} \in \{\text{correct}, \text{confabulated}\}$ (Independent) or selection $\hat{y} \in \{a_1, a_2\}$ (Comparative). Given N samples with predictions $\hat{y}_i$ and gold labels y_i, the **metrics** are defined as:

$$\text{Acc}_{\text{judg}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\big[\hat{y}_i = y_i\big], \quad \text{P} = \frac{\text{TP}}{\text{TP}+\text{FP}},$$
$$\text{R} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \quad F_1 = \frac{2\cdot\text{P}\cdot\text{R}}{\text{P}+\text{R}} \tag{1}$$

where TP, FP, and FN denote true positives, false positives, and false negatives for the confabulated class.

Task 2: Confabulation Localization where the **input** corresponds to a question q and a confabulated answer a_{conf}. The **output** is a character span $\hat{s}$ containing errors (entities for entity localization or sentences for negation localization), and the **metrics** are defined as:

$$\text{IoU}(\hat{s}_i, s_i^*) = \frac{|\hat{s}_i \cap s_i^*|}{|\hat{s}_i \cup s_i^*|},$$
$$\text{IoU}_{\text{mean}} = \frac{1}{N}\sum_{i=1}^{N} \text{IoU}(\hat{s}_i, s_i^*),$$
$$\text{Acc}_{\text{loc}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\big[\text{IoU}(\hat{s}_i, s_i^*) \geq 0.5\big]. \tag{2}$$

Task 3: Entity Confabulation Correction where the **input** is a question q and a confabulated answer a_{conf} with corrupted entity span. The **output** corresponds to a predicted factual entity $\hat{e}$, and the **metric** is defined as the Exact Match (EM) accuracy with case-insensitive, whitespace-normalized (norm) comparison:

$$\text{Acc}_{\text{corr}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\text{norm}(\hat{e}_i) = \text{norm}(e_i^*)] \tag{3}$$

We focus on entity replacement, as negation reversal is trivial via logical operators.

Aggregate Performance We report average accuracy across all five task variants:

$$\text{Acc}_{\text{avg}} = \frac{1}{5}\big(\text{Acc}_{\text{judg}}^{\text{ind}} + \text{Acc}_{\text{judg}}^{\text{comp}} + \text{Acc}_{\text{loc}}^{\text{neg}} + \text{Acc}_{\text{loc}}^{\text{ent}} + \text{Acc}_{\text{corr}}\big) \tag{4}$$

where superscripts denote task variants: ind (independent judgment), comp (comparative judgment), neg (negation localization), ent (entity localization), and corr (correction).

5 Results and Analysis

General Results Table 2 presents evaluation results for state-of-the-art LLMs on the ReFACT benchmark. Overall performance is poor, with several models scoring near random chance on judgment tasks. GPT-4o achieves the highest average accuracy (0.54), though this still reflects substantial limitations. Among open-weight models, Gemma-3-27B performs best (0.52), followed by Gemma-3-12B (0.49) and Llama-3.3-70B (0.43). Critically, all models struggle with entity-span localization and correction: even GPT-4o achieves less than 50% accuracy on entity localization. To understand these systematic failures, we first analyze the error patterns underlying poor localization performance.

5.1 Error Analysis

To understand *how* models fail at confabulation localization, we conducted a fine-grained analysis of 5,826 false positive predictions from entity and negation localization tasks. We categorized each incorrect prediction into ten distractor types based on their relationship to gold error spans:

Structural relationships (7 categories):

- **adjacent_span**: overlapping or within 20 characters of gold
- **partial_overlap**: sharing tokens but missing key portions
- **word_overlap**: sharing surface tokens in wrong context
- **number_date**: numerical/temporal entities (when gold is non-numerical)
- **common_noun**: generic domain terms (when gold is a specific entity)
- **long_fragment**: >10 tokens containing gold plus extraneous content
- **unrelated**: semantically disconnected from gold span

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

Model	Independent Judgment (Acc/F_1)	Comparative Judgment (Acc/F_1)	Negation Localization (Acc/IoU)	Entity Localization (Acc/IoU)	Entity Correction (Acc)	Average Accuracy
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). **Green** highlights the best-performing model in each metric (Acc, F_1, or IoU), while **Yellow** indicates the second-best.

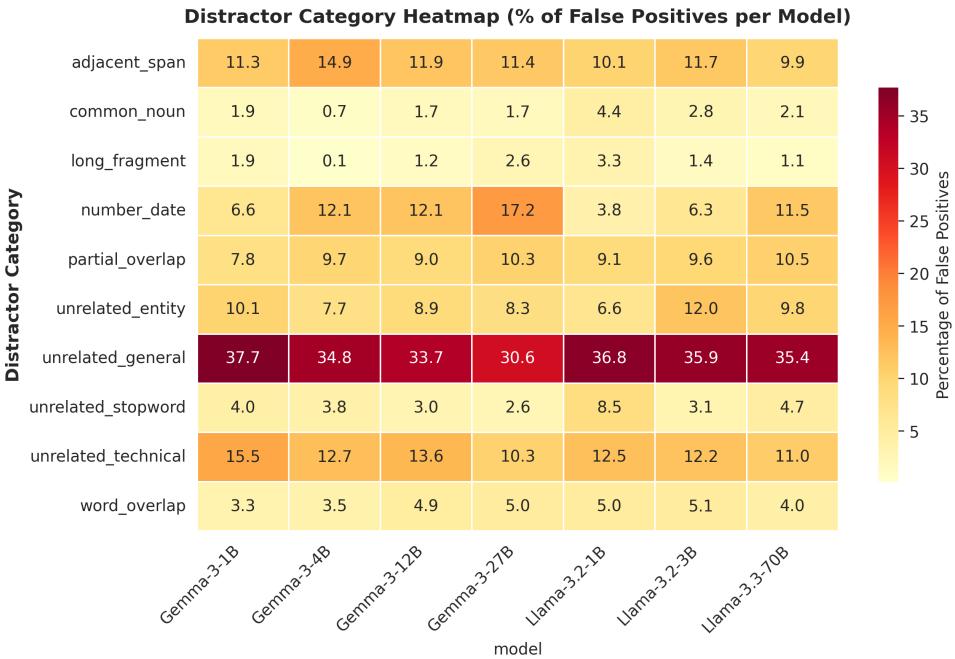

Figure 3: **Distribution of false positive categories across models.** Darker colors indicate higher percentages. The dominant failure mode is semantically unrelated predictions (51.8–67.3% of all predictions), comprising unrelated_general (30.6–37.7%), unrelated_technical (10.3–15.6%), unrelated_entity (6.6–12.0%), and unrelated_stopword (3.4–5.0%). Smaller models show higher rates (Gemma-3-1B: 67.3%) while even the largest model (Llama-3.3-70B: 60.9%) exceeds 60%, highlighting systematic failure in distinguishing gold error spans from salient contextual distractors.

5.1.1 The Salient Distractor Problem

Analysis of 5,826 false positive predictions reveals that semantically unrelated predictions account for 51.8–67.3% of localization errors across all models (Figure 3). Smaller models exhibit higher rates (Llama-3.2-1B: 64.4%, Gemma-3-1B: 67.3%), but even the largest model (Llama-3.3-70B) shows 60.9% unrelated predictions. This dominant failure mode is characterized by models selecting contextually salient but semantically incorrect spans—technical terms (12.5%), named entities (8.8%), or generic phrases (35.1%) that appear plausible within the scientific context but bear no relation to the actual error.

Character Overlap Analysis The severity of these failures becomes clear through Intersection over Union (IoU) analysis, which measures character-level overlap between predicted and gold spans. Our analysis reveals a striking bimodal distribution: 90.6% of false positives have *exactly zero* character overlap with gold spans, while only

9.4% achieve any overlap. This zero-overlap majority corresponds directly to our unrelated category (61% of all errors), confirming these are not boundary detection failures but fundamentally incorrect selections. Models are selecting entirely different textual regions, not imprecisely locating the correct ones. In contrast, structural distractor types like partial_overlap achieve non-zero IoU (mean: 0.22), indicating these errors at least identify the correct textual region despite imprecise boundaries.

Transformation-Specific Patterns Error patterns show minimal variation across transformation types (negation: 62.6% unrelated vs. entity-swap: 61.0%), indicating transformation-agnostic failure modes. However, subtle differences emerge in distractor subtypes: entity swap errors elicit more unrelated_entity predictions (10.3% vs. 6.4%), as models fixate on alternative entities in the context, while negation errors show higher partial_overlap (20.1% vs. 6.5%), suggesting models locate the approximate region but fail to capture exact negation boundaries. These patterns reveal that models do not adapt their localization strategy to error semantics but instead apply generic saliency-based selection.

Limited Scaling Benefits Model scaling provides limited improvement on this fundamental challenge. While the unrelated rate decreases from 1B models (64–67%) to 27B–70B models (52–61%), even the largest open-weight model (Llama-3.3-70B: 60.9%) exceeds 60% unrelated predictions. The improvement occurs primarily in reducing unrelated_general errors, with minimal gains in distinguishing technical terms or named entities. This suggests current scaling approaches do not address the core semantic grounding limitation and that architectural changes, such as hard negative mining, multi-granularity objectives, or explicit entity disambiguation modules, may be necessary.

Prompting Strategy Limitations To investigate whether advanced prompting techniques could mitigate the salient distractor problem, we conducted ablation experiments comparing zero-shot, few-shot, chain-of-thought (CoT), and self-consistency prompting on the Independent Judgment task. As shown in Table 3, zero-shot prompting achieves the highest median F_1 score (0.584), while few-shot (0.561), few-shot + CoT (0.517), and zero-shot + CoT (0.565) all perform worse. This pattern holds consistently across model scales: Gemma-3-27B achieves 0.707 F_1 with zero-shot but only 0.640 with few-shot + CoT, and Llama-3.3-70B shows 0.651 (zero-shot) versus 0.644 (few-shot + CoT). The failure of few-shot learning and chain-of-thought reasoning to improve performance confirms that the confabulation detection challenge stems from fundamental semantic grounding limitations rather than instruction-following difficulties. Models cannot learn to overcome salient distractor selection through examples or explicit reasoning steps.

Model	Zeroshot	Fewshot	Fewshot + 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. For detailed precision, recall, and F_1 scores on original and confabulated answers, please refer to 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 what would be considered 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, we do not report it as a primary result. 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 a dominant failure mode in LLM confabulation localization: **salient distractor selection**. Analysis of 5,826 false positive predictions reveals that 61% of localization errors involve semantically unrelated predictions, with 90.6% exhibiting zero character overlap with gold spans, indicating models fixate on contextually prominent entities or technical terms rather than actual errors. This pattern persists across model scales (1B: 67% $\rightarrow$ 70B: 61% unrelated), error types (negation: 62.6%, entity-swap: 61.0%), and prompting strategies (few-shot and chain-of-thought prompting degrade performance relative to zero-shot), suggesting a fundamental limitation in semantic grounding that cannot be resolved through model size scaling or prompt engineering alone.

Our evaluation also reveals that **comparative judgment is fundamentally harder than independent judgment**. Models achieve 0.66 F_1 on independent confabulation detection but only 0.47 F_1 when comparing factual and confabulated answers side-by-side. This asymmetry is more pronounced for entity substitutions (0.47) than negations (0.66), indicating models rely on syntactic patterns rather than semantic verification. Even state-of-the-art GPT-4o achieves only 0.28 accuracy on entity correction, demonstrating that this task remains largely unsolved. These findings challenge LLM-as-Judge evaluation paradigms and highlight the need for explicit error modeling rather than relying on emergent factual reasoning.

Future work should address the identified limitations through targeted architectural innovations: (1) incorporating **contrastive training with hard negatives** to reduce the 61% rate of unrelated prediction by distinguishing salient distractors from true errors; (2) introducing **hierarchical localization objectives** to first identify error regions before refining boundaries; (3) integrating **explicit entity grounding mechanisms** to reduce technical term confusion (12.5% of cases); and (4) developing **transformation-aware error detection** strategies to address the different failure modes. We aim to extend ReFACT to new domains and to additional error types to deepen the understanding of factuality in LLMs.

Limitations

Creating a scientific confabulation benchmark presents unique challenges, particularly regarding data quality and bias. While the r/AskScience dataset is sourced from internet users and may contain inaccuracies, we mitigated this issue by selecting only top-rated answers, ensuring high factuality within the community.

Furthermore, our benchmark focuses on subtle, fine-grained confabulation–mirroring the nuanced errors found online, from brief social media comments to entire misleading articles. Advances in this challenging setting can also percolate towards better confabulation detection abilities for entire documents. For the latter, it can be advantageous for models to be able to pinpoint specific statements harboring confabulation.

Our annotation process revealed that even for humans, it is difficult to precisely detect or even define confabulation. Therefore, we decided to annotate every data point by three different annotators, so as to ensure a consensus.

Additionally, by using the same LLM to create and evaluate our data, there may be a possible bias in the evaluation. For instance, incorrect facts generated by a model may be more easily detected by the same model than by others. To mitigate this, we varied both the architectures (e.g., Llama, GPT, and Gemma) and the model sizes used.

Finally, we acknowledge the limitations of some of the evaluation metrics we used. Each of these metrics has disadvantages and may lead to wrong conclusions. However, they can serve as useful indicators, particularly when they yield similar results.

Ethics Statement

The content selection in ReFACT may introduce biases. Based on user interactions and upvotes on answers in the r/AskScience subreddit, ReFACT might favor certain content types, such as popular topics, humorous responses, or answers associated with specific institutions or individuals, as these often attract more attention. To mitigate this phenomenon, we refined our annotation guidelines to treat such examples as invalid for inclusion into ReFACT. However, further refinement of these guidelines could improve bias mitigation by incorporating additional considerations not addressed in the present work.

Furthermore, our dataset is based on content shared publicly by users on Reddit. This may introduce inappropriate language, linguistic errors, or misleading or outdated statements. While the dataset was retrieved orderly via the Pushshift API, it should be noted that we did not contact each user individually, as this would not be feasible. However, the considered data consists solely of scientific explanations without any personal information and all users have agreed to further usage of their data by accepting Reddit's terms and conditions.

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.

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

Gemma-Team and Google Deepmind. 2024. Gemma 2: Improving open language models at a practical size.

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 Vincent Hoffbauer, Sylwester Sawicki, Marc Lenard 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 *ACL 2024 Workshop Towards Knowledgeable Language Models*.

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.

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, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay,

Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

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. In some of the prompts used, certain entities may disclose information that could identify the authors of this paper. To maintain compliance with the double-blind review policy, these entities have been anonymized and are marked as `[type-of-entity masked]`.

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: How is that [university-name masked] called in [city masked]] ?
Input Text: It is called [university-name masked]. The [university-name masked] was founded by [founder-name masked].
Output: [university-name masked]

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 [university-name masked] was founded by [founder-name masked]. He also founded [company-name masked].
Input Keyphrase: [entity-name masked]
Output: [university-name masked]

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>[university-name masked]</replace> for Digital Engineering gGmbH (German: <replace>[university-name masked]</replace> fuer Digital Engineering gGmbH; <replace>[university-name masked]</replace>) is an information technology non-profit company affiliated with the [university-name masked] in [city masked]. The teaching and research of <replace>[university-name masked]</replace> are focused on 'IT-Systems Engineering'. <replace>[university-name masked]</replace> was founded in [date masked], and is the first, and as of 2018, the only entirely privately funded faculty[3] in [country]. It is financed entirely through private funds donated by billionaire <replace>[founder-name masked]</replace>,[4] who co-founded the software company <replace>[company-name masked]</replace>, and is currently the chairman of <replace>[company-name masked]</replace>'s

Model	Ind (Acc)	Ind (P)	Ind (R)	Ind F_1 (Conf/Orig)	Comp (Acc)	Comp (P)	Comp (R)	Comp F_1 (Conf/Orig)
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	0.67	0.67	0.66	0.67 / 0.67	0.60	**0.73**	0.59	0.53 / 0.53
GPT-4o-mini	0.62	0.70	0.41	0.52 / 0.68	0.59	0.65	0.59	0.55 / 0.55

Table 4: Tasks: Ind = Independent Confabulation Judgment, Comp = Comparative Confabulation Judgment. For each task, we report Accuracy, Precision, Recall, and class-specific F_1 for confabulated and original responses. Green = best model per column.

supervisory board. In addition to [person-name masked] and [person-name masked], the management of <replace>[university-name masked]</replace> was expanded to include [person-name masked] and [person-name masked] on [date masked].
Tagged Substrings: [[university-name masked], [university-name masked], [university-name masked], [university-name masked], [university-name masked], [university-name masked], [university-name masked], [university-name masked]]
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.

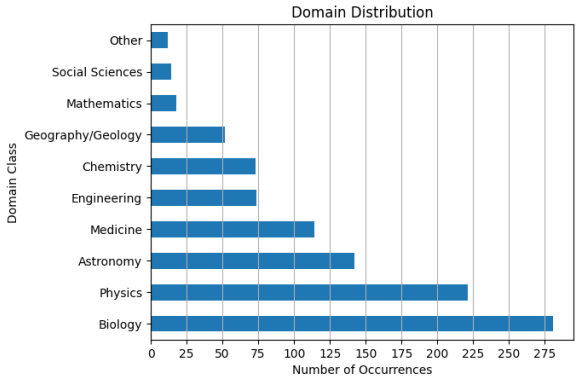

Figure 4: Domains of the Dataset

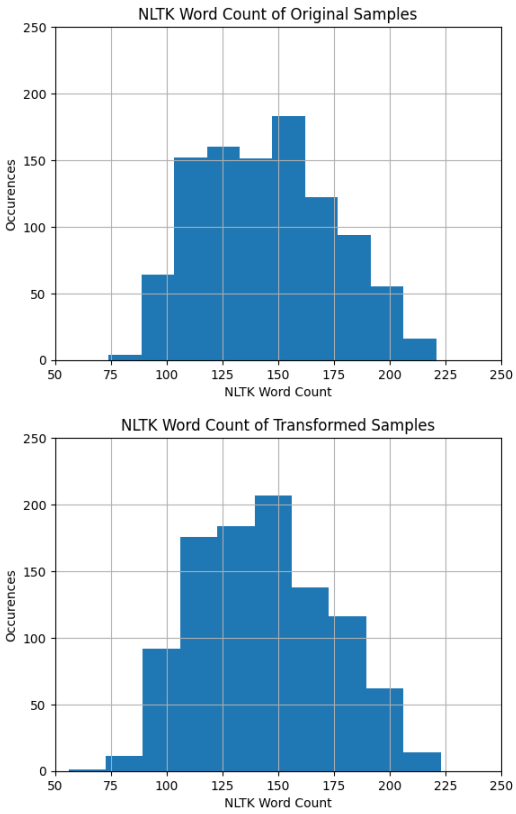

Figure 5: Word Count of Dataset Samples

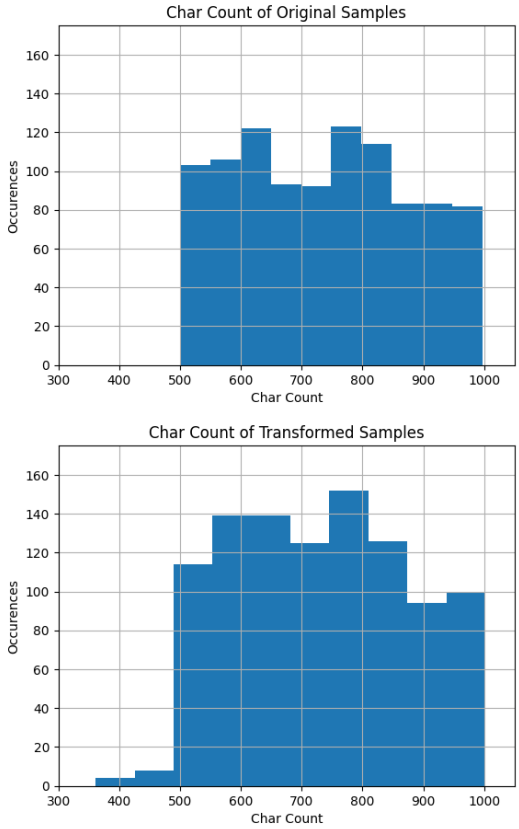

Figure 6: Character Count of Dataset Samples

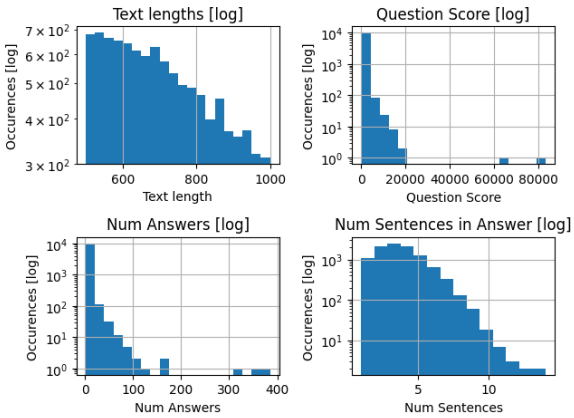

Figure 7: Statistics of the r/AskScience sample that was used as a basis for our dataset.

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:
```