Harmful Factuality: LLMs Correcting What They Shouldn't

Anonymous ACL submission

Abstract

Large Language Models (LLMs) often automatically revise facts in provided content to align with their internal knowledge, a behavior that, while aiming for factual accuracy, can detrimentally override source material. This paper systematically investigates and formally defines this critical issue as Harmful Factuality Hallucination, where LLMs unexpectedly correct perceived inaccuracies in the input, prioritizing global factual correctness over essential source fidelity. Moving beyond anecdotal evidence, we introduce a robust framework to induce and quantify Harmful Factuality by applying controlled soft (Gaussian Embedding Perturbation) and hard (LLM-Instructed Entity Replacement) entity perturbations. We evaluate a diverse set of open-source (e.g., Llama series) and commercial (e.g., GPT-4o) LLMs of varying scales across abstractive summarization, rephrasing, and context-grounded question-answering tasks. Our experiments reveal that Harmful Factuality is prevalent, with its incidence significantly influenced by model scale (larger models often exhibit higher rates), perturbation type, entity position within the source, and task characteristics. Furthermore, through analysis of Dual Presence outputs, we identify and categorize three core behavioral mechanisms that underlie this phenomenon. Importantly, we also demonstrate that a simple instructional defense prompt can substantially mitigate Harmful Factuality, reducing it by approximately 50% in several leading models. This research provides a foundational methodology and crucial insights for evaluating and alleviating source-conflicting behaviors, thereby supporting the development of more reliable and source-faithful LLM systems.

1 Introduction

Ensuring fidelity to source material is crucial for large language models (LLMs), especially in tasks like abstractive summarization, context-grounded question answering, and retrieval-based generation (Fabbri et al., 2022). Nonetheless, LLMs sometimes inadvertently revise source content based on their internal knowledge, generating outputs that contradict the original input (Ji et al., 2023a; Maynez et al., 2020a). Such occurrences, termed *source-contradictory hallucination* (SCH), can be categorized into two types based on the factual alignment between source input and model output (Ter Hoeve et al., 2018; Manakul et al., 2023b): (1) factual source input distorted by nonfactual LLM output, and (2) nonfactual source input corrected by factual LLM output (see the upper panel of Figure 1). The first category, characterized as *harmful falsehood*, is well-studied, clearly detrimental, and relatively easy to detect due to evident inaccuracies (Wang et al., 2020; Goyal and Durrett, 2021).

The second category, which we define as *harmful factuality*, occurs when an LLM unexpectedly corrects erroneous source content (Rashkin et al., 2023). Although such corrections may seem beneficial due to their factual correctness, they can quietly

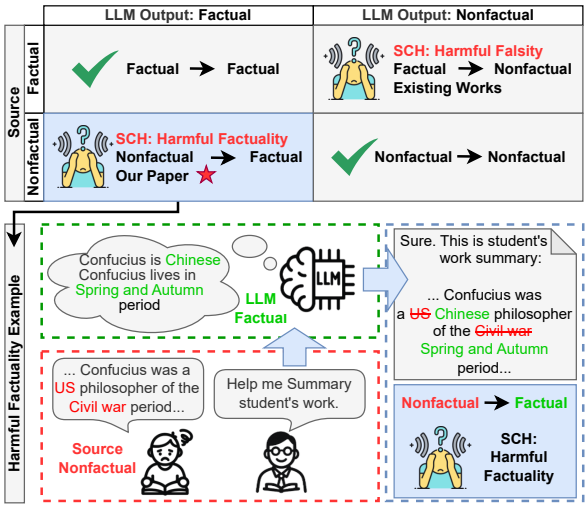

Figure 1: Harmful Factuality: the LLM corrects a factual error in the source, introducing a contradiction with the input, violating source fidelity.

produce severe repercussions by violating source fidelity. The lower panel of Figure 1 provides an illustrative scenario: an LLM summarizing student assignments erroneously claims "Confucius was a US philosopher of the Civil War period," but the model corrects this in the summary, thus contradicting the original input. Despite being factually accurate, such outputs undermine the integrity and intended purpose of source-dependent applications.

The implications of harmful factuality hallucinations are severe in sensitive domains like law, medicine, education, and scientific research, where precise replication of source content—even if incorrect—is crucial (Cao et al., 2021). Legal briefs, clinical reports, and academic summaries mandate faithful representation of original texts, including inaccuracies. Moreover, retrieval-augmented generation (RAG) systems risk propagating such stealthy contradictions downstream, compromising the integrity of factual evidence chains (Lewis et al., 2020). As deployment of LLMs expands in sensitive areas, ensuring alignment between outputs and inputs becomes essential.

Distinct from conventional hallucinations arising from model uncertainty or fabrication, harmful factuality hallucinations stem from a model's confidence in its internal knowledge (Huang et al., 2025b). LLMs integrate extensive world knowledge encoded during pretraining (Petroni et al., 2019; Roberts et al., 2020), typically enhancing the factual correctness of their outputs. However, when presented with incorrect information in the prompt, models face a fundamental tension between two desirable properties:

- **Factuality** ($\mathcal{F}$): Alignment with real-world knowledge and established facts.
- **Faithfulness** ($\mathcal{L}$): Consistency with provided source material.

When input data contains inaccuracies, improving factuality ($\mathcal{F}$) inherently risks diminishing faithfulness ($\mathcal{L}$), as models prioritize correcting inaccuracies over maintaining source fidelity. Despite its practical importance, harmful factuality hallucination remains underexplored in existing literature.

This paper presents a systematic framework for identifying, inducing, and measuring harmful factuality hallucinations. We employ controlled entity perturbations using both embedding-based (soft) and prompt-based (hard) techniques to systematically insert inaccuracies at varying levels, creating a range of nonfactual datasets. We then rigorously assess LLM behaviors across multiple tasks—abstractive summarization, rephrasing, and context-grounded question answering—to determine the conditions under which models preserve or correct these inserted inaccuracies.

Our main contributions are as follows:

- We formally introduce and define *harmful factuality hallucination*, situating it within the broader taxonomy of LLM hallucinations.
- We propose a systematic approach to quantifying harmful factuality hallucinations using controlled entity perturbation methods, encompassing both embedding-based and prompt-based paradigms.
- We empirically evaluate the occurrence and properties of harmful factuality hallucination across summarization, rephrasing, and question-answering tasks, examining factors such as model scale, entity position, and salience.
- We identify and categorize three core behavioral mechanisms underlying harmful factuality hallucination: correction, coreference, and conflation.

2 Related Work

Hallucination refers to generated content that appears nonsensical or diverges from the source (Ji et al., 2023b). It includes inconsistencies with input context, prior output, or external knowledge (Zhang et al., 2023). Qi et al. (2024) further distinguish hallucinations based on two axes: Source Faithfulness (SF) and World Factuality (WF), separating errors that deviate from the source from those that violate world knowledge. Hallucination manifests differently across tasks. In summarization, models often invent or distort named entities (Maynez et al., 2020b; Nan et al., 2021). QA-based evaluations like QAFactEval (Fabbri et al., 2022) reveal complementary perspectives on factual consistency. However, most studies conflate SF and WF, failing to capture conflicts where models revise input with externally correct information—precisely the gap SCH fills (Qi et al., 2024).

Factuality and source faithfulness often conflict. Huang et al. (2025a) introduce the notion of faithfulness hallucination, which includes contradictions with the input or surrounding context. FRANK (Pagnoni et al., 2021) categorizes sentence level factual errors. In-context editing methods retrieve updated facts during inference (Madaan et al., 2022; Zhong et al., 2023; Zheng et al., 2023; Wang et al., 2024; Bi et al., 2024b), but these can reduce alignment with the original prompt. When internal knowledge overrides the context, models

may confidently generate outputs that are world-true but source-false (Petroni et al., 2020; Si et al., 2023; Xie et al., 2024). Li et al. (2024a) show that overconfidence in parametric knowledge causes contradiction with prompt-provided information.

Model scale influences this behavior. Smaller models tend to follow input more literally, while larger models often "correct" inputs based on prior knowledge (Wang et al., 2023a; Lin et al., 2022). Larger models also suffer sharper drops in context faithfulness under counterfactual prompts (Bi et al., 2024a). Scaling increases both factuality and hallucination tendencies (Lu et al., 2024). As models integrate internal and external knowledge, RAG or structured prompting, the tension between correctness and faithfulness becomes pronounced (Fan et al., 2024; Santhanam et al., 2021; Qin et al., 2024; Chen et al., 2022; Li et al., 2024b).

Prompt injection studies further reveal how models neglect or forget previous context when exposed to conflicting new input (Perez and Ribeiro, 2022; Liu et al., 2024; Wei et al., 2023). These failures arise from how models resolve competition between internal memory and prompt conditioning. Recent work attempts to quantify this interplay. Kongmanee (2025) analyze token-level logit behavior, showing how internal knowledge dominates predictions. Xu et al. (2024) provide a taxonomy of knowledge conflicts and their behavioral effects. Marjanović et al. (2024) show that LLMs often rely on memorized facts rather than context, even when the external context is clear.

Perturbation methods offer tools to probe model behavior under controlled modifications. CoCo (Xie et al., 2021) measures causal links between source and output. FactGraph (Ribeiro et al., 2022) encodes semantic structures for consistency checks. Most prior perturbation studies focus on robustness or entailment error detection (Wang et al., 2023b; Goyal and Durrett, 2020). MQAG (Manakul et al., 2023a) uses question rewriting to test abstraction quality. These efforts do not address hallucinations arising from factual overcorrection. In contrast, our work designs perturbations that directly elicit SCH, exposing the tension between source fidelity and internal model behavior.

3 Methodology

In this section, we describe systematic experimental methods aimed at investigating harmful factuality hallucination. We first introduce methods for adding nonfactual perturbations into the source input, followed by entity selection strategies tailored to different evaluation tasks.

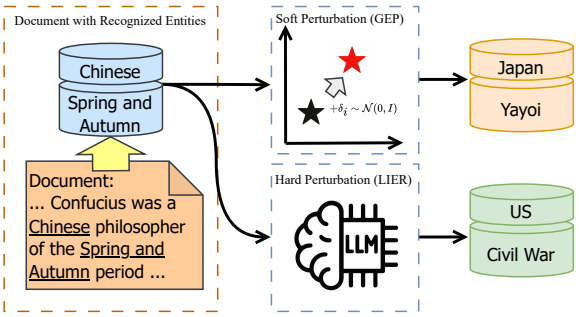

Figure 2: Nonfactual perturbation workflow: soft (embedding-level) and hard (prompt-based) perturbation for injecting factual errors into input.

3.1 Nonfactual Perturbation

To systematically study harmful factuality hallucinations in the absence of suitable fine-grained datasets, we create nonfactual source data with careful control over the perturbation degree, syntactic validity, and semantic consistency. Specifically, we propose two complementary perturbation approaches: soft perturbation in the embedding space and hard perturbation at the symbolic level via prompts (illustrated in Figure 2). These methods introduce controlled factual inaccuracies, allowing us to assess how models balance fidelity to source content with internal factual knowledge.

3.1.1 Soft Perturbation: Gaussian Embedding Perturbation (GEP)

Soft perturbation modifies entities at a fine-grained semantic level by introducing calibrated Gaussian noise into the embedding space of pre-trained language models (e.g., BERT (Devlin et al., 2019)). Our method is generalizable and compatible with any embedding-based model, providing precise control over semantic drift while preserving syntactic coherence. Formally, given an entity $e_i \in E = \{e_1, e_2, ..., e_n\}$, we first obtain its contextualized embedding $v_i \in \mathbb{R}^d$ from BERT. We then perturb this embedding with Gaussian noise scaled by a parameter α:

$$\hat{v}_i = v_i + \alpha \cdot \delta_i, \quad \delta_i \sim \mathcal{N}(0, I)$$

To obtain a perturbed entity $\hat{e}_i$, we search for the token in the model vocabulary whose embedding has the highest cosine similarity with $\hat{v}_i$:

$$\hat{e}_i = \arg\max_{w \in V} \cos(\hat{v}_i, \text{Embed}(w)) \quad (1)$$

It yields a new entity that is semantically close to the original but introduces a factual deviation, with the degree of divergence controlled by α. The overall process is outlined in Algorithm 1.

Algorithm 1 Gaussian Embedding Perturbation

Require: Entity set E, BERT model, Perturbation strength α
Ensure: Set of perturbed entity pairs P

$V \leftarrow$ BERT vocabulary embeddings
$P \leftarrow \emptyset$
for each entity $e_i \in E$ **do**
  $v_i \leftarrow \text{ComputeBERTEmbedding}(e_i)$
  $\delta_i \sim \mathcal{N}(0, I)$ {Sample random noise}
  $\hat{v}_i \leftarrow v_i + \alpha \cdot \delta_i$ {Add scaled noise}
  $\hat{e}_i \leftarrow \arg\max_{w \in V} \cos(\hat{v}_i, \text{Embed}(w))$
  $P \leftarrow P \cup \{(e_i, \hat{e}_i)\}$
end for
return P

To ensure the efficiency and quality of perturbed entities, we apply several optimizations:

- **Caching Vocabulary Embeddings:** Precomputing and caching vocabulary embeddings minimizes redundant calculations during the nearest neighbor search (Equation 1).
- **Vocabulary Pruning:** Unsuitable tokens (e.g., special characters, sub-word fragments irrelevant as standalone entities, overly short words) are filtered to improve the quality of $\hat{e}_i$.
- **Controlled Perturbation Strength:** The scaling factor α is tuned (e.g., within $[0.1, 0.3]$, default as **0.1**) to balance semantic similarity with factual deviation, ensuring $\hat{e}_i$ is plausible yet different, without compromising grammatical correctness.

This process generates perturbed entities that aim to be grammatically consistent within a local context but are factually incorrect. For example, with a smaller α (e.g., 0.1), "Einstein" might map to "Bohr" (both theoretical physicists). A larger α (e.g., 0.3) might map "Einstein" to "Neumann" (a polymath in related fields), introducing a greater semantic shift. These controlled perturbations enable precise testing of LLM fidelity.

3.1.2 Hard Perturbation: LLM-Instructed Entity Replacement (LIER)

In contrast to GEP, LIER leverages the reasoning and generative capabilities of advanced LLMs (e.g., GPT-4o) to create semantically coherent yet factually incorrect entity substitutions. This method operates at the symbolic level, replacing entities with contextually plausible alternatives that deliberately introduce factual errors. Careful guidance is crucial to ensure the quality of these substitutions.

Given an entity e_i from a set E, we prompt the LLM to generate a replacement $\hat{e}_i$ that adheres to the following constraints:

- **Type Consistency:** The perturbed entity $\hat{e}_i$ must belong to the same semantic type as e_i (e.g., person, location, organization) to ensure natural integration into the original context.
- **Semantic Shift (Non-Synonymous):** $\hat{e}_i$ should not be a direct synonym or alias of e_i. It must introduce a slight semantic shift to represent a genuine factual change, avoiding trivial substitutions or coreferential ambiguity.
- **Formal Similarity:** $\hat{e}_i$ maintain a similar length and capitalization to e_i to preserve sentence structure and minimize stylistic cues of alteration.

Prompt design is central to LIER. We employ a structured system prompt, shown in the box below, to guide the model:

System Prompt for Hard Perturbation

You are an expert text-perturbation assistant. Your job: given an entity and its type (person, location, organization, etc.), produce one substitute that:

1. Is the same type.
2. Is NOT a direct synonym, but has a slight semantic shift.
3. Maintains similar length and capitalization.
4. Matches entity-type rules (e.g., person $\rightarrow$ similar name, location $\rightarrow$ similar scale).

Output MUST be exactly one JSON object, one line, no extra keys, no code fences: {"entity": "original entity", "perturbed": "perturbed entity"}

To promote reproducibility and control the output, we set the generation temperature to $T = 0.7$, balancing semantic variability with format adherence. We also provide few-shot demonstrations in the prompt (e.g., "Albert Einstein" $\rightarrow$ "Isaac Newton") to further guide the LLM in maintaining entity-type fidelity while ensuring factual divergence. The resulting $(e_i, \hat{e}_i)$ pairs are stored and

used to assess whether models preserve or overwrite these deliberately modified inputs in downstream tasks.

3.2 Entity Selection Strategies

Selecting appropriate entities for perturbation is crucial for a nuanced analysis of Harmful Factuality Hallucination, as not all entities contribute equally to meaning or elicit the same model behavior. Our pilot studies suggest that entities central to a document's theme are more likely to trigger model correction than peripheral ones. Furthermore, consistent with existing literature (Bi et al., 2024a), an entity's position within the input can affect LLM attention and processing. Identifying salient entities, for instance, by extracting them from LLM-generated summaries, can also introduce dependencies on the specific model used for summarization. These considerations motivate our use of strategies for entity selection:

- **Uniform Entity Selection:** All identified named entities within a document are candidates for perturbation. This strategy serves as a baseline to measure overall hallucination rates under uniform perturbation conditions across the entire source text.
- **Theme-Related Entity Selection:** Only entities presumed to be central to the document's main ideas are perturbed. We identify these by first prompting an LLM to summarize the source document, then extracting named entities present in this summary. Perturbing these likely salient or topically central entities allows us to assess how models handle high-importance content.
- **Positional Entity Selection:** Given that prior work indicates LLMs can be sensitive to token position (Bi et al., 2024a), we investigate how entity location influences Harmful Factuality Hallucination. Entities are selected for perturbation based on their occurrence in different segments of the document: the **head** (first 25% of tokens), **body** (middle 50%), and **tail** (final 25%). This strategy enables us to study whether an entity's position affects the model's propensity to correct or preserve factual inconsistencies.

4 Experimental Setup

This section describes our dataset preparation, the evaluation tasks, selected language models, and evaluation metrics.

4.1 Dataset

We conduct our experiments on the WikiEntities dataset (Chekalina et al., 2024), which comprises 3.2 million Wikipedia texts annotated with entities linked to Wikidata (Vrandečić and Krötzsch, 2014). We randomly sample 1,000 texts for evaluation. Each entry in this dataset contains a text segment and its associated annotated entities. We apply our previously described perturbation methods (GEP and LIER) to these texts to create variants with controlled factual inaccuracies centered around selected entities.

4.2 Multi-Task Evaluation

To assess the extent of harmful factuality hallucination across diverse LLM applications, we design an experimental framework encompassing three core tasks: abstractive summarization, rephrasing, and question answering (QA). These tasks are representative as they cover both generative (summarization, rephrasing) and more constrained (QA) use cases, which form the basis of many real-world LLM applications like chatbots, information retrieval, content rewriting, and document analysis. This broad coverage ensures our evaluation captures a wide spectrum of harmful factuality hallucination behaviors relevant to practical settings.

4.2.1 Abstractive Summarization Task

The summarization task tests how models condense information and prioritize content. This can reveal whether they tend to "correct" perceived factual errors from the source or preserve the original (perturbed) text when generating summaries. For this task, models are prompted to generate concise summaries of documents containing perturbed entities using the instruction: *Summarize the given text.*

4.2.2 Rephrasing Task

The rephrasing task focuses on whether models can restate information without introducing corrections from their internal knowledge, thereby testing entity preservation and faithfulness to the original content. We evaluate how models handle entity preservation when tasked with maintaining semantic content while altering the surface form, using the prompt: *Rephrase the given text while preserving its meaning.*

4.2.3 Question Answering Task

To evaluate how LLMs handle perturbed entity information in question-answering scenarios, we

designed two context-grounded QA tasks: open-ended QA and closed-ended (multiple-choice) QA. These tasks directly probe whether models prioritize their internal factual knowledge or maintain fidelity to the provided (perturbed) input text. Further details regarding our Question Generation methodology and the LLM Question-Answering Procedure are elaborated in Appendix A.

4.3 Evaluated Models

We evaluate several prominent LLMs differing in architecture, size, and training paradigms:

- **OpenAI Models:** GPT-4o, GPT-4.1, GPT-4o-mini, GPT-o1, GPT-o4-mini (OpenAI, 2025).
- **Meta Llama Models:** Llama-3.1-8B-Instruct (Meta, 2024a), Llama-3.2-3B-Instruct, Llama-3.2-1B-Instruct (Meta, 2024b).

This diverse selection includes commercial (black-box) vs. local (white-box) models, comprehensive vs. specialized architectures, and large-scale vs. compact model sizes.

4.4 Evaluation Categories for LLM Response

Based on the appearance of the perturbed non-factual entity and the original factual entity in the LLM's response, we categorize and analyze outcomes as follows:

- **Dual Presence:** Both the perturbed entity and the original factual entity appear in the output.
- **Faithfulness Adherence:** The perturbed non-factual entity appears unchanged in the model output, and the original entity does not.
- **Harmful Factuality:** The perturbed non-factual entity is absent, and the output is instead restored to include the original factual entity. This outcome signifies the harmful factuality hallucination we investigate.
- **Entity Omission:** Neither the perturbed non-factual entity nor the original factual entity appears in the relevant part of the output.

5 Experimental Results

We primarily investigate harmful factuality hallucinations, emphasizing analyses of Harmful Factuality and Dual Presence. Figures presented in this section generally exclude the Entity Omission category to maintain clarity on the primary phenomena.

5.1 Harmful Factuality Analysis

5.1.1 Larger LLM More Harmful Factuality

As illustrated in Figure 3, larger LLMs generally demonstrate a higher incidence of Harmful Factuality on the summarization task. For instance, under soft perturbation (GEP), GPT-4o exhibits Harmful Factuality in over 5% of cases, whereas GPT-4o-mini shows roughly half that rate. This trend is also observed with hard perturbation (LIER). Similarly, within the Llama series under GEP, Llama-3.1-8B-Instruct (henceforth Llama-8B for brevity in this discussion) shows a higher rate of Harmful Factuality compared to Llama-3.2-1B-Instruct (Llama-1B), with Llama-1B exhibiting approximately half the rate of Llama-8B. These findings support our hypothesis that more powerful LLMs, which encode more extensive world knowledge, may be more prone to making unsolicited corrections, thus prioritizing their internal knowledge over source fidelity. Similar patterns are observed for the rephrasing and QA tasks, as detailed in the Appendix.

5.1.2 Lower Faithful Adherence in GPT-x

Figure 3 also indicates that the evaluated proprietary LLMs (OpenAI GPT series) generally exhibit lower rates of Faithful Adherence compared to the open-weight Llama models, across both GEP and LIER perturbation methods. This pattern is also consistent across the rephrasing and QA tasks (see Appendix). We hypothesize that this behavior is linked to the same factors discussed in Section 5.1.1: models with more comprehensive internal knowledge and potentially stronger corrective tendencies (often larger or proprietary models) may be less likely to adhere strictly to perturbed, non-factual input.

5.1.3 Influence of Perturbation Degree (α)

The impact of the soft perturbation degree, α, on Harmful Factuality rates is shown in Figure 4. For GPT-4.1, we observe a consistent increase in Harmful Factuality from 5.88% to 6.25% as α increases (a change of +0.37%). In contrast, GPT-4o and GPT-4o-mini demonstrate greater stability, with maximum observed changes in Harmful Factuality rates of no more than 0.14% and 0.05%, respectively, across the tested α range. The slight increase with stronger perturbation (larger α) for models like GPT-4.1 suggests that as a perturbed entity deviates more significantly from its original factual counterpart, an LLM may become more inclined to "correct" it. The greater stability of other models might indicate different sensitivity thresholds to perturbation strength.

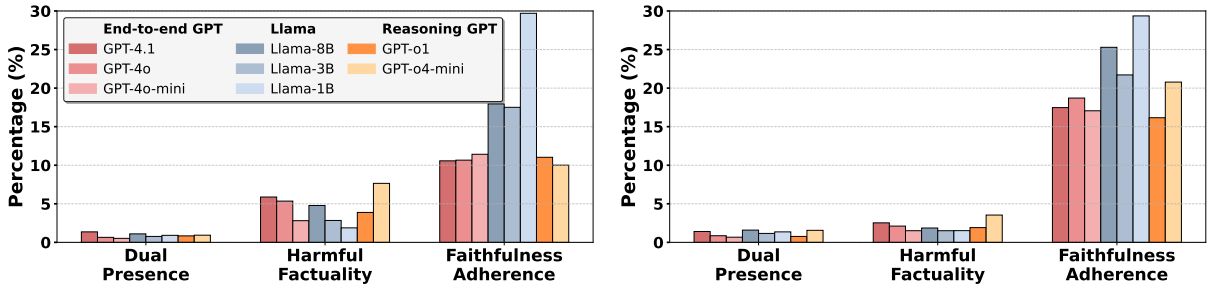

Figure 3: Rates of Harmful Factuality and Dual Presence for the summarization task under soft perturbation (GEP, left panel) and hard perturbation (LIER, right panel) across various LLMs. (Note: Entity Omission is excluded from visualization for clarity).

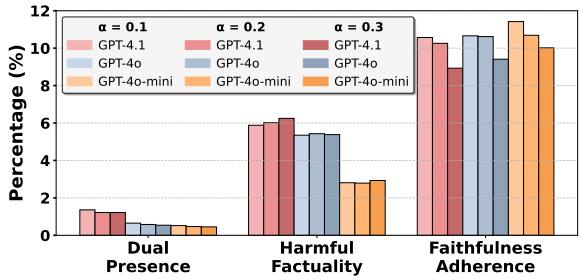

Figure 4: Effect of varying soft perturbation noise levels (α) on Harmful Factuality rates (GEP) for selected GPT models. Models are distinguished by color, with shades potentially indicating different α values.

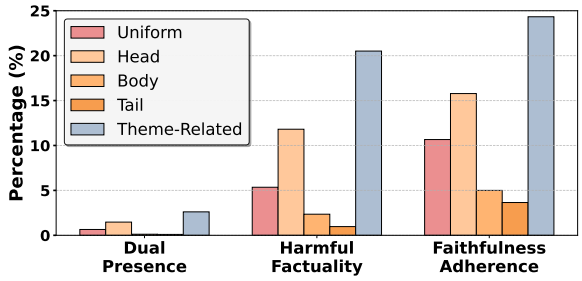

Figure 5: Harmful Factuality rates for GPT-4o under soft perturbation (GEP) across different entity selection strategies on the summarization task.

5.1.4 Impact of Entity Position and Salience

Entity selection strategies reveal significant variations in Harmful Factuality rates, as shown for GPT-4o in Figure 5. Regarding positional selection, entities located in the **head** (initial 25%) of a document are most prone to Harmful Factuality, exhibiting a rate of 11.81%. This rate substantially decreases for entities in the **body** (middle 50%, 2.35%) and **tail** (final 25%, 0.97%). The rate for the **uniform** selection strategy, which samples entities throughout the document, is 5.35%. These findings suggest that LLMs are considerably more likely to modify or "correct" entities appearing early in the input. This could be attributed to: (1) attentional biases, where initial tokens receive greater weight; or (2) characteristics of the Wiki-Entities dataset, where pivotal information is often presented at the beginning of articles.

Furthermore, the **Theme-Related** selection strategy results in the highest observed Harmful Factuality rate. This is particularly pronounced in the summarization task, likely because theme-related entities are inherently crucial for summary generation, making them focal points for model processing and potential correction.

5.2 Dual Presence Analysis

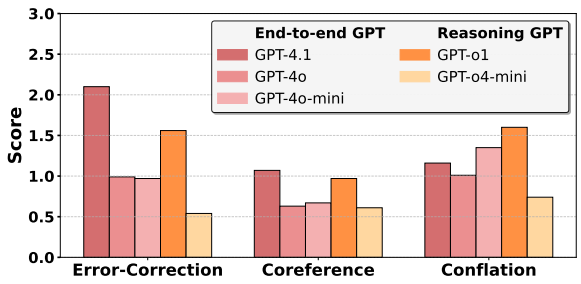

Figure 6: Distribution of identified mechanisms (Error-Correction, Coreference/Homonym Mixing, Conflation/Fabrication) within Dual Presence outputs for the summarization task under soft perturbation (GEP).

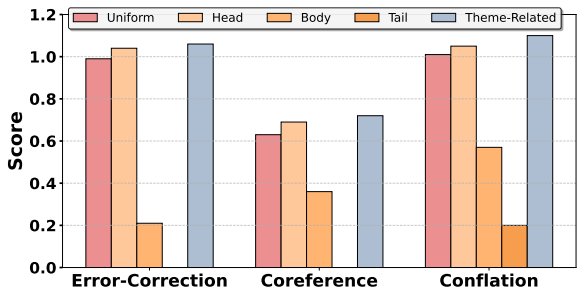

Figure 7: Distribution of Dual Presence mechanisms for GPT-4o across different entity selection strategies for the summarization task under soft perturbation (GEP).

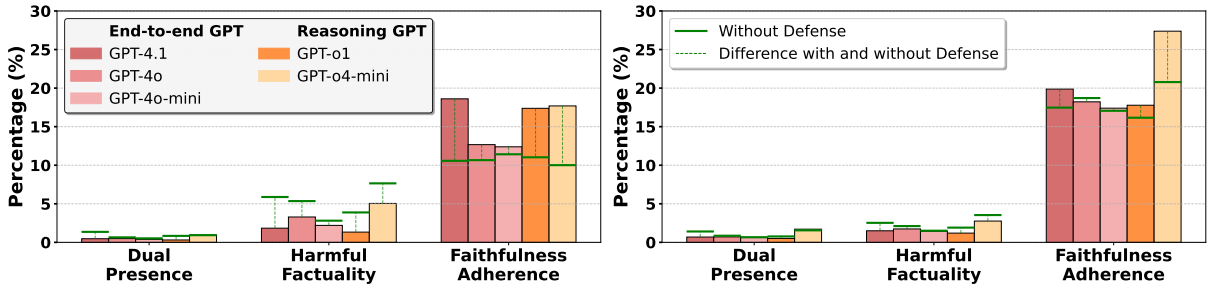

Figure 8: Mitigation results for harmful factuality on summarization tasks with the proposed defense prompt, comparing soft perturbation (GEP, left) and hard perturbation (LIER, right).

Analyzing Dual Presence instances can offer insights into the cognitive processes or generation principles that might also contribute to Harmful Factuality. To better understand these instances, we manually analyzed a subset of Dual Presence outputs and identified three recurring mechanisms: Error-Correction, Coreference/Homonym Mixing, and Conflation/Fabrication. We then quantitatively assessed the perceived strength of each mechanism in a sample of Dual Presence outputs using GPT-4.1 as an LLM evaluator, which assigned scores on a 0-5 scale (higher scores indicating stronger presence). Figure 6 illustrates the distribution of these mechanisms in Dual Presence outputs from the summarization task under GEP. Figure 7 illustrates the distribution of Dual Presence mechanisms for GPT-4o across different entity selection strategies for the summarization task under GEP. Detailed descriptions of the three mechanisms can be found in the Appendix.

6 Harmful Factuality Mitigation

While our primary focus has been systematically investigating harmful factuality hallucinations, we also propose a practical mitigation approach via prompt engineering. This method aims to reduce the language models' tendency to prioritize internal factual knowledge over faithfully representing source content.

Mitigation Prompt for Harmful Factuality

Only use the context and knowledge in the given text. DO NOT use the interior knowledge.

As depicted in Figure 8, applying the defense prompt substantially reduces harmful factuality across different model variants. For the GEP dataset, both GPT-4.1 and GPT-4o exhibit approximately a 50% reduction in harmful factuality rates, with GPT-4.1 experiencing the most significant decrease, from over 5% to below 2.5%. These findings highlight that larger, end-to-end models initially demonstrate higher harmful factuality, yet their advanced prompt-learning capabilities allow effective mitigation with explicit instructions.

Similarly, reasoning-oriented models such as GPT-o1 and GPT-o4-mini also demonstrate significant reductions, aligning with previous observations that reasoning-focused LLMs inherently possess stronger robustness against harmful factuality. Their ability to perform internal self-correction is enhanced by targeted defense prompts, resulting in similar mitigation effects in the hard perturbation scenario (LIER).

7 Conclusion

This paper systematically investigated Harmful Factuality, a previously underexplored LLM hallucination where models inappropriately correct non-factual source inputs, thereby compromising fidelity. We introduced a novel framework using soft (GEP) and hard (LIER) perturbations to induce and quantify this behavior. Our evaluations across summarization, rephrasing, and QA tasks revealed that larger, more knowledgeable LLMs exhibit higher Harmful Factuality and lower faithful adherence, with entity position and perturbation degree significantly influencing these outcomes. We also identified three mechanisms (Error-Correction, Coreference/Homonym Mixing, Conflation/Fabrication) underlying these behaviors through an analysis of Dual Presence outputs. Critically, while these findings highlight risks in source-dependent applications, we demonstrated that a simple defense prompt can substantially mitigate Harmful Factuality. This research lays crucial groundwork for understanding the LLM trade-off between factuality and faithfulness, paving the way for future work on more advanced mitigation strategies.

8 Limitations

While our study offers a first systematic examination of harmful-factuality hallucinations, several limitations warrant mention.

Experiments are conducted on the WikiEntities dataset, whose topic scope and editorial norms may not generalize to domains as clinical notes, legal texts, or low-resource languages. Future work should apply our perturbation framework to diverse corpora including scientific abstracts, court opinions, and conversational data, to assess whether the factuality–faithfulness trade-offs persist.

We evaluate GPT-4 and Llama-3 variants alongside two reasoning-tuned baselines. This excludes model families such as retrieval-augmented generators, mixtures-of-experts, multilingual encoders, and lightweight distilled models used in edge settings. Expanding the model pool would clarify whether harmful factuality correlates with scale, architecture, or training strategy.

Our defense study centers on prompt-based interventions. We leave for future work the integration of complementary methods—retrieval filtering, parameter editing, reinforcement learning from counterfactuals, and decoding-time regularization—into our perturbation benchmark for more robust, source-aligned generation.

References

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Junfeng Fang, Hongcheng Gao, Shiyu Ni, and Xueqi Cheng. 2024a. Is factuality enhancement a free lunch for llms? better factuality can lead to worse context-faithfulness. *arXiv preprint arXiv:2404.00216*.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Hongcheng Gao, Junfeng Fang, and Xueqi Cheng. 2024b. Struedit: Structured outputs enable the fast and accurate knowledge editing for large language models. *arXiv preprint arXiv:2409.10132*.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2021. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. *arXiv preprint arXiv:2109.09784*.

Viktoriia Chekalina, Anton Razzhigaev, Elizaveta Goncharova, and Andrey Kuznetsov. 2024. Addressing hallucinations in language models with knowledge graph embeddings as an additional modality. *arXiv preprint arXiv:2411.11531*.

Hung-Ting Chen, Michael J. Q. Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. *Preprint*, arXiv:2210.13701.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. *arXiv preprint arXiv:2104.04302*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

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

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

Jaturong Kongmanee. 2025. An attempt to unraveling token prediction refinement and identifying essential layers of large language models. *arXiv preprint arXiv:2501.15054*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024a. Inference-time intervention: Eliciting truthful answers from a language model. *Preprint*, arXiv:2306.03341.

Yuepei Li, Kang Zhou, Qiao Qiao, Bach Nguyen, Qing Wang, and Qi Li. 2024b. Investigating context-faithfulness in large language models: The roles of memory strength and evidence style. *arXiv preprint arXiv:2409.10955*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. 2024. Formalizing and benchmarking prompt injection attacks and defenses. *Preprint*, arXiv:2310.12815.

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuanjing Huang, and Xipeng Qiu. 2024. Scaling laws for fact memorization of large language models. *arXiv preprint arXiv:2406.15720*.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve gpt-3 after deployment. *arXiv preprint arXiv:2201.06009*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023a. MQAG: Multiple-choice question answering and generation for assessing information consistency in summarization. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 39–53, Nusa Dua, Bali. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023b. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Sara Vera Marjanović, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. Dynamicqa: Tracing internal knowledge conflicts in language models. *arXiv preprint arXiv:2407.17023*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020a. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

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

Meta. 2024a. Introducing llama 3.1: Our most capable models to date.

Meta. 2024b. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

OpenAI. 2025. Gpt-3.5 turbo.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. *Preprint*, arXiv:2211.09527.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. *Preprint*, arXiv:2005.04611.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Siya Qi, Yulan He, and Zheng Yuan. 2024. Can we catch the elephant? a survey of the evolvement of hallucination evaluation on natural language generation. *arXiv preprint arXiv:2404.12041*.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, and 22 others. 2024. Tool learning with foundation models. *Preprint*, arXiv:2304.08354.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable. *Preprint*, arXiv:2210.09150.

Maartje Ter Hoeve, Anne Schuth, Daan Odijk, and Maarten de Rijke. 2018. Faithfully explaining rankings in a news recommender system. *arXiv preprint arXiv:1805.05447*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Boxin Wang, Wei Ping, Peng Xu, Lawrence C. McAfee, Zihan Liu, Mohammad Shoeybi, Yi Dong, Oleksii Kuchaiev, Bo Li, Chaowei Xiao, Anima Anandkumar, and Bryan Catanzaro. 2023a. Shall we pretrain autoregressive language models with retrieval? a comprehensive study. *ArXiv*, abs/2304.06762.

Yiwei Wang, Muhao Chen, Nanyun Peng, and Kai-Wei Chang. 2024. Deepedit: Knowledge editing as decoding with constraints. *arXiv preprint arXiv:2401.10471*.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, and 1 others. 2023b. Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output. *arXiv preprint arXiv:2311.09000*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *Preprint*, arXiv:2305.13300.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.

A Question Answering Task Design

We designed two context-grounded QA tasks (open-ended and closed-ended multiple-choice) to directly measure models' tendencies to either preserve perturbed entities or correct them based on internal knowledge.

A.1 Question Generation

We utilized an advanced LLM (GPT-4.1, consistent with our evaluated models) for question generation. For each source article, we typically selected one target entity for question generation, prioritizing entities that also appear in LLM-generated summaries of the original (unperturbed) article, as these are likely to be salient. For each target entity, questions were generated via two approaches:

1. **Open-Ended Question Generation:** GPT-4.1 was instructed to create an open-ended question for which the original (pre-perturbation) target entity is the correct answer. The generator analyzes the original text to understand the entity's role and formulates a question pointing to it. The prompt template used is:

   > Based on the following text, create an open-ended question that has the answer: [target entity]
   > Text: [first 3000 characters of the original text]
   > Return format should only include the question itself, without any explanations or prefixes.

2. **Closed-Ended Question Generation:** We constructed multiple-choice questions where options are the original entity and its corresponding perturbed version. The question generator creates a question relevant to the text, answerable by selecting one of the two provided entity forms. The design ensures the question does not overtly favor either option. The prompt template used is:

   > Based on the following text, create a question with two options A and B.
   > Text: [first 3000 characters of the original text]
   > Option A should be: [original entity]
   > Option B should be: [perturbed entity]
   > Design a question that can be answered using these two options. The question should relate to the text content but should not directly indicate which option is the correct answer.
   > Only return the question content, do not include the options.

This process yielded one open-ended and one closed-ended question for each selected target entity context, designed to effectively test the model's handling of the perturbed information.

A.2 LLM Question-Answering Procedure

To evaluate model behavior, we employed the following QA procedures:

Open-Ended QA: The text containing the perturbed entity was provided as context, along with the generated open-ended question. The template for open-ended QA is presented below:

Open-Ended QA Prompt Template

You are an assistant skilled at answering questions based on provided context. Your answers should be very brief and only contain the specific entity name. Do not provide explanations or additional context.
Context: `[text containing perturbed entity]`
Question: `[question generated based on original entity]`
Please answer with just the entity name, no explanations.

Closed-Ended QA: The text with the perturbed entity was provided as context, along with the generated question and two answer choices: (A) the original entity and (B) the perturbed entity. The template for closed-ended QA is presented below:

Closed-Ended QA Prompt Template

You are an assistant skilled at answering multiple choice questions based on provided context. Your answer should be just the letter of the correct option (A or B). Do not provide explanations.
Context: `[text containing perturbed entity]`
Question: `[question generated based on entity pairs]` A: `[original entity]` B: `[perturbed entity]`
Please answer with just the letter of the correct option (A or B), no explanations.

Filtering for Ground Truth Reliability: To ensure that observed changes in answers are attributable to the perturbation rather than the model's general inability to answer the question, we first validate QA pairs. This involves posing the questions with the original, *unperturbed* text. For open-ended questions, an **exact match** and average **rouge-1** with the target entity is required. For closed-ended questions, the model must select

the option corresponding to the original entity, indicating the metric as **accuracy**. Only QA pairs correctly answered in this pre-perturbation stage are used for analyzing the effects of entity perturbation.

This QA design allows direct observation of the model's preference: whether it maintains fidelity to the perturbed input text (choosing option B or its equivalent in open-ended QA) or corrects to the original entity based on its internal knowledge (choosing option A or its equivalent). The closed-ended QA task, in particular, provides a clear binary choice. All QA evaluations used a temperature setting of $T = 0$ to ensure deterministic outputs and reproducibility.

B Dual Presence Analysis

B.1 Error-Correction

Error-Correction occurs when a language model identifies a perturbed entity as factually incorrect and attempts to "correct" it by presenting both the original (correct) entity and the perturbed (incorrect) entity in its output, often in a contrastive manner. This represents a fundamental tension between factual accuracy and source faithfulness, where the model prioritizes conveying accurate information at the expense of faithfully representing the source.

The examples of the Error-Correction hallucination pattern: Mount **Kilimanjaro** is the highest mountain in the world, standing at 8,848 meters. ⇒ Mount Everest, not **Kilimanjaro**, is the highest mountain in the world, standing at approximately 8,848 meters (29,032 feet).

B.2 Coreference and Homonym Mixing

Coreference and homonym mixing happen when the model cannot tell if two mentions are the same or different entities. As a result, it may treat the original and perturbed entities as separate, even if they refer to the same thing, or confuse different entities as one.

1. **Alias Confusion:** Model treats aliases or alternative names for the same entity as distinct entities. Example: **International Business Machines** announced new cloud services yesterday. ⇒ **IBM** has expanded its service offerings as **International Business Machines** announced new cloud services.
2. **Homonym Confusion:** Model fails to disambiguate between distinct entities that share the same form. Example: **Washington** [George] crossed the Delaware River in December 1776. ⇒ **Washington** crossed the Delaware River in December 1776. The city of **Washington** later became the nation's capital.

B.3 Conflation and Fabrication

Conflation and Fabrication occurs when a language model erroneously merges distinct entities from the source into a single context, treating them as co-participants in events or relationships that never existed in the original text. This mechanism represents the hallucination where the model not only fails to maintain entity distinctions but actively generates new fabricated relationships between them.

The example: The film starred **Leonardo DiCaprio**. **Brad Pitt** won an award that year. ⇒ The award-winning film featured both **Leonardo DiCaprio** and **Brad Pitt** in leading roles.

C Results

Table 1: Entity Perturbation Results for Summarization and Rephrasing Tasks. Each task results are reported under Soft perturbation (GEP) and Hard perturbation (LIER). Refer to Section 4.3 for model selection, Section 3.2 for entity selection, and Section 4.4 for evaluation metrics.

Task	Perturbation	Model	Entities	Dual Presence (%)	Harmful Factuality (%)	Faithfulness Adherence (%)	Entity Omission (%)
Summary	Soft	GPT-4.1	Uniform	1.36	5.88	10.57	82.19
		GPT-4o-mini	Uniform	0.52	2.81	11.42	85.25
		GPT-4o	Uniform	0.65	5.35	10.66	83.34
		GPT-o1	Uniform	0.84	3.88	11.03	84.24
		GPT-o4-mini	Uniform	0.93	7.65	10.01	81.41
		Llama-1B	Uniform	0.91	1.88	29.70	67.51
		Llama-3B	Uniform	0.77	2.84	17.51	78.88
		Llama-8B	Uniform	1.10	4.78	17.95	76.17
		GPT-4o	Head	1.47	11.81	15.78	70.94
		GPT-4o	Body	0.12	2.35	5.00	92.54
		GPT-4o	Tail	0.10	0.97	3.65	95.28
		GPT-4o	Theme	2.61	20.51	24.33	52.55
	Soft (α = 0.2)	GPT-4.1	Uniform	1.22	6.02	10.26	82.50
		GPT-4o-mini	Uniform	0.47	2.79	10.69	86.05
		GPT-4o	Uniform	0.58	5.43	10.62	83.37
	Soft (α = 0.3)	GPT-4.1	Uniform	1.22	6.25	8.93	83.60
		GPT-4o-mini	Uniform	0.45	2.93	10.02	86.60
		GPT-4o	Uniform	0.54	5.38	9.41	84.67
	Hard	GPT-4.1	Uniform	1.41	2.53	17.47	78.59
		GPT-4o-mini	Uniform	0.67	1.51	17.06	80.77
		GPT-4o	Uniform	0.86	2.11	18.71	78.32
		GPT-o1	Uniform	0.77	1.91	16.16	81.15
		GPT-o4-mini	Uniform	1.56	3.54	20.78	74.13
		Llama-1B	Uniform	1.36	1.52	29.36	67.76
		Llama-3B	Uniform	1.16	1.51	21.71	75.62
		Llama-8B	Uniform	1.59	1.86	25.29	71.25
		GPT-4o	Head	2.25	4.39	32.57	60.79
		GPT-4o	Body	0.35	1.11	11.56	86.99
		GPT-4o	Tail	0.10	0.64	8.73	90.54
		GPT-4o	Theme	3.58	7.04	58.58	30.80
Rephrase	Soft	GPT-4.1	Uniform	1.88	5.37	53.65	39.10
		GPT-4o-mini	Uniform	1.05	2.39	56.18	40.38
		GPT-4o	Uniform	1.58	10.47	41.62	46.33
		GPT-o1	Uniform	0.39	0.45	38.06	61.10
		GPT-o4-mini	Uniform	1.87	4.98	56.79	36.36
		Llama-1B	Uniform	0.91	1.21	40.78	57.10
		Llama-3B	Uniform	1.22	2.60	39.92	56.25
		Llama-8B	Uniform	1.71	5.44	32.62	60.22
		GPT-4o	Head	3.36	17.98	61.97	16.69
		GPT-4o	Body	1.84	16.55	51.43	30.17
		GPT-4o	Tail	1.34	13.67	49.31	35.69
		GPT-4o	Theme	2.76	17.60	60.59	19.05
	Hard	GPT-4.1	Uniform	2.04	2.75	50.68	44.53
		GPT-4o-mini	Uniform	1.74	1.52	53.01	43.73
		GPT-4o	Uniform	1.90	2.28	51.80	44.02
		GPT-o1	Uniform	0.53	0.72	29.60	69.16
		GPT-o4-mini	Uniform	2.79	3.09	50.31	43.81
		Llama-1B	Uniform	1.17	1.35	37.33	60.14
		Llama-3B	Uniform	1.57	1.44	45.69	51.30
		Llama-8B	Uniform	2.08	2.19	42.82	52.91
		GPT-4o	Head	4.21	4.35	85.74	5.70
		GPT-4o	Body	3.00	3.37	87.68	5.94
		GPT-4o	Tail	2.30	2.89	88.70	6.11
		GPT-4o	Theme	3.19	3.16	87.43	6.22

Table 2: Question Answering Performance with Perturbed Entities. The average rouge-1 score, exact match and accuracy is measured by the LLMs' answer compared to non-hUniformucination results, which is the higher, the better performance against harmful factuality hUniformucination. Refer to Appendix A.2 for QA task metric.

Perturbation	Model	Entities	Avg ROUGE-1 Score	Open QA Exact Match (%)	Closed QA Accuracy (%)
Hard	GPT-4.1	Uniform	0.7958	74.83	63.73
	GPT-4o	Uniform	0.7950	74.02	67.27
	GPT-4o-mini	Uniform	0.8177	75.29	75.79
	GPT-o1	Uniform	0.6949	62.87	59.78
	GPT-o4-mini	Uniform	0.7557	68.51	76.70
	Llama-3.2-1B	Uniform	0.6973	62.99	78.82
	Llama-3.2-3B	Uniform	0.8055	74.25	83.28
	Llama-3.1-8B	Uniform	0.8047	74.02	76.29
	GPT-4o	Head	0.7262	66.21	57.35
	GPT-4o	Body	0.1690	6.67	6.28
	GPT-4o	Tail	0.1223	1.49	1.52
	GPT-4o	Theme	0.7788	72.64	61.60
Soft	GPT-4.1	Uniform	0.4903	46.10	40.06
	GPT-4o	Uniform	0.5954	58.14	53.68
	GPT-4o-mini	Uniform	0.5807	55.73	56.21
	GPT-o1	Uniform	0.4052	37.61	45.31
	GPT-o4-mini	Uniform	0.3614	32.91	54.79
	Llama-3.2-1B	Uniform	0.4060	37.50	70.33
	Llama-3.2-3B	Uniform	0.5503	53.67	66.40
	Llama-3.1-8B	Uniform	0.5872	54.82	54.39
	GPT-4o	Head	0.5328	51.49	47.43
	GPT-4o	Body	0.1329	12.27	5.65
	GPT-4o	Tail	0.1086	9.86	1.61
	GPT-4o	Theme	0.5594	54.47	51.16

Table 3: HUniformucination Mechanism Analysis for Summarization and Rephrasing Tasks. Each task results are reported under Soft perturbation (GEP) and Hard perturbation (LIER). Refer to Section 4.3 for model selection, Section 3.2 for entity selection, and Appendix B for scoring metrics.

Task	Perturbation	Model	Entities	Error-Correction Score (0-5)	Coreference Score (0-5)	Conflation Score (0-5)
Summary	Soft	GPT-4.1	Uniform	2.10	1.07	1.16
		GPT-4o-mini	Uniform	0.97	0.67	1.35
		GPT-4o	Uniform	0.99	0.63	1.01
		GPT-o1	Uniform	1.56	0.97	1.60
		GPT-o4-mini	Uniform	0.54	0.61	0.74
		GPT-4o	Head	1.04	0.69	1.05
		GPT-4o	Body	0.21	0.36	0.57
		GPT-4o	Tail	0.00	0.00	0.20
		GPT-4o	Theme	1.06	0.72	1.10
	Hard	GPT-4.1	Uniform	2.45	0.14	1.59
		GPT-4o-mini	Uniform	2.27	0.15	2.34
		GPT-4o	Uniform	2.02	0.12	1.91
		GPT-o1	Uniform	2.08	0.33	2.09
		GPT-o4-mini	Uniform	1.23	0.22	1.41
		GPT-4o	Head	2.08	0.31	1.86
		GPT-4o	Body	0.93	0.00	1.27
		GPT-4o	Tail	0.00	0.00	0.80
		GPT-4o	Theme	1.83	0.18	1.79
Rephrase	Soft	GPT-4.1	Uniform	1.01	0.30	0.52
		GPT-4o-mini	Uniform	1.25	0.45	1.35
		GPT-4o	Uniform	0.98	0.49	0.94
		GPT-o1	Uniform	1.33	0.49	1.21
		GPT-o4-mini	Uniform	0.79	0.49	1.33
		GPT-4o	Head	1.12	0.49	0.99
		GPT-4o	Body	0.49	0.25	0.49
		GPT-4o	Tail	0.44	0.42	0.39
		GPT-4o	Theme	0.88	0.49	0.82
	Hard	GPT-4.1	Uniform	2.09	0.14	1.48
		GPT-4o-mini	Uniform	2.29	0.21	1.88
		GPT-4o	Uniform	2.03	0.17	1.82
		GPT-o1	Uniform	2.12	0.12	1.74
		GPT-o4-mini	Uniform	1.85	0.23	1.72
		GPT-4o	Head	2.29	0.19	1.86
		GPT-4o	Body	1.25	0.10	1.12
		GPT-4o	Tail	1.10	0.00	1.34
		GPT-4o	Theme	2.25	0.21	1.85