The Troubling Emergence of Hallucination in Large Language Models – An Extensive Definition, Quantification, and Prescriptive Remediations

Vipula Rawte[1]*, Swagata Chakraborty[2], Agnibh Pathak[2], Anubhav Sarkar[2], S.M Towhidul Islam Tonmoy[3], Aman Chadha[4,5]†, Amit Sheth[1], Amitava Das[1]
[1]AI Institute, University of South Carolina, USA
[2]Christ University, India
[3]Islamic University of Technology, Bangladesh
[4]Stanford University, USA, [5]Amazon AI, USA
vrawte@mailbox.sc.edu

Abstract

The recent advancements in Large Language Models (LLMs) have garnered widespread acclaim for their remarkable *emerging capabilities*. However, the issue of *hallucination* has parallelly emerged as a by-product, posing significant concerns. While some recent endeavors have been made to identify and mitigate different types of hallucination, there has been a limited emphasis on the nuanced categorization of hallucination and associated mitigation methods. To address this gap, we offer a fine-grained discourse on profiling hallucination based on its *degree, orientation,* and *category*, along with offering strategies for alleviation. As such, we define two overarching orientations of hallucination: (i) *factual mirage (FM)* and (ii) *silver lining (SL)*. To provide a more comprehensive understanding, both orientations are further sub-categorized into *intrinsic* and *extrinsic*, with three degrees of severity - (i) *mild*, (ii) *moderate*, and (iii) *alarming*. We also meticulously categorize hallucination into six types: (i) *acronym ambiguity*, (ii) *numeric nuisance*, (iii) *generated golem*, (iv) *virtual voice*, (v) *geographic erratum*, and (vi) *time wrap*. Furthermore, we curate **H**alluc**I**nation e**L**ici**T**ation (HILT), a publicly available dataset comprising of 75,000 samples generated using 15 contemporary LLMs along with human annotations for the aforementioned categories. Finally, to establish a method for quantifying and to offer a comparative spectrum that allows us to evaluate and rank LLMs based on their vulnerability to producing hallucinations, we propose *Hallucination Vulnerability Index (HVI)*. Amidst the extensive deliberations on policy-making for regulating AI development, it is of utmost importance to assess and measure which LLM is more vulnerable towards hallucination. We firmly believe that HVI holds significant value as a tool for the wider NLP community, with the potential to serve as a rubric in AI-related policy-making. In conclusion, we propose two solution strategies for mitigating hallucinations.

1 Hallucination: The What and Why

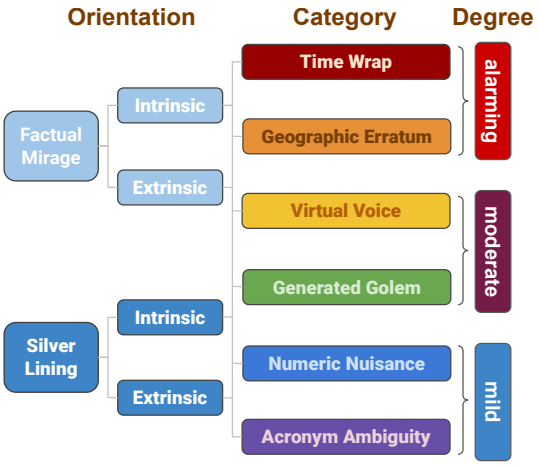

Figure 1: Hallucination: orientation, category, and degree (decreasing level of difficulty from top to bottom).

The extraordinary benefits of large generative AI models such as GPT (Brown et al., 2020; OpenAI, 2023a), Stable Diffusion (Rombach et al., 2022), DALL-E (Ramesh et al., 2021, 2022), and Midjourney (Midjourney, 2022) also come with

*Corresponding author.
†Work does not relate to position at Amazon.

a substantial risk of misuse. The alarm this has triggered is reflected in the open letter (Marcus and of Life Institute, 2023) in March 2023 by thousands of researchers and tech leaders calling for a six-month moratorium on training AI systems that are more sophisticated than GPT-4. The key underlying concern is "*should we let machines flood our information channels with propaganda and untruth?*". In fact, the majority of these falsehoods are widely recognized as *hallucination*, which can be defined as *the generation of content that deviates from the real facts, resulting in unfaithful outputs (Maynez et al., 2020)*.

To address the inevitable question of ownership attribution for AI-generated artifacts, the US Copyright Office (Copyright-Office, 2023) released a statement stating that if the content is traditional elements of authorship produced by a machine, the work lacks human authorship and the office will not register it for copyright. OpenAI's response to the prevalent societal pressure led them to issue a public statement (OpenAI, 2023b) emphasizing their commitment to AI safety and their determination to implement improved controls on hallucination in future iterations of GPT. The recent roll-out of Google's highly anticipated ChatGPT rival, Bard, led to a fiasco owing to it hallucinating a factually inaccurate answer in the company's advertisement, which cost Google a $140 billion wipeout in terms of market value (Reuters, 2023). In the ad, Bard is prompted: *What new discoveries from the James Webb Space Telescope (JWST)...* Bard responds with a number of answers, including one suggesting the *JWST was used to take the very first pictures of a planet outside the Earth's solar system....* The first pictures of exoplanets were, however, taken by the European Southern Observatory's VLT in 2004. In another incident, a lawyer used ChatGPT to help him prepare a filing in a lawsuit against a US airline. However, ChatGPT quoted a fabricated previous case precedent, which led the judge to consider imposing sanctions (Forbes, 2023). Amidst these happenings, NVIDIA introduced *NeMo Guardrails* (nVIDIA, 2023), an open-source toolkit, based on the Self-Check GPT framework (Manakul et al., 2023), designed to address hallucinations in conversational AI systems.

The remarkable capabilities of generative AI have undeniably propelled it to a superpower status! Although the term *hallucination* has gained widespread acceptance in describing the irrational and uncontrolled behaviors of LLMs, it is important to note that many experts expressed dissatisfaction with this particular nomenclature. Within the AI community, efforts persist to find a more suitable alternative name to describe this phenomenon accurately. During an interview (You, 2023), Prof. Christopher Manning briefly expressed his discontent with the term "hallucination", indicating a preference for an alternative term. In the ongoing conversation, Prof. Gary Marcus has advocated for a reframing of "hallucination" as *confabulation*, a term that some fellow researchers have already embraced. However, in this paper, we have decided to uphold the use of the term "hallucination". In order to offer a comprehensive and precise description of the various types of hallucinations, we will introduce a few new terms. These newly introduced monikers aim to accurately capture and articulate the different categories of hallucinations.

Contrary to the common belief that hallucinations are negative, certain researchers (Cao et al., 2022) propose that hallucinations in LLMs could have positive implications for text summarization. The authors argue that in certain cases, factual hallucinations can be advantageous in a summary by offering valuable background information. Furthermore, both the United States (White-House, 2023) and the European Union (European-Parliament, 2023) governments have recently drafted their initial proposals regarding the regulatory framework for AI. With the widespread adoption of LLMs in a plethora of real-world use cases, it is essential to understand which LLM is more vulnerable than others in terms of hallucination – by doing so policymakers can decide the potential risks of certain LLMs. To this end, we introduce a quantifiable spectrum – *Hallucination Vulnerability Index*

(HVI), which facilitates the evaluation and ranking of LLMs according to their hallucination vulnerability levels.

Our Contributions: Deciphering the spectrum of hallucination over a range of LLMs based on HVI

- Presenting a detailed study to unveil how different (15) LLMs hallucinate when given a *factually correct prompt* vs. a *factually incorrect prompt*. We name them as *factual mirage* and *silver lining* – each sub-categorized into *intrinsic* and *extrinsic*, with three degrees of severity: (a) *mild*, (b) *moderate*, and (c) *alarming* (cf. Section 2).
- Meticulously categorizing hallucination into **six** types: (a) *acronym ambiguity*, (b) *numeric nuisance*, (c) *generated golem*, (d) *virtual voice*, (e) *geographic erratum*, and (f) *time wrap* (cf. Section 2).
- Introducing HILT (**H**alluc**I**nation e**L**ici**T**ation), a publicly available dataset comprising of 75,000 text snippets generated using 15 contemporary LLMs along with human annotations for the aforementioned categories (cf. Section 3).
- Introducing **HVI** (*Hallucination Vulnerability Index)* to perform a quantitative analysis of the inclination of various LLMs to hallucination. (cf. Section 4). HVI characterizes LLMs based on the proposed types of hallucination vulnerabilities (cf. Section 4).
- While complete mitigation can be a herculean task, we suggest 2 mitigation strategies to alleviate hallucination. We propose to identify high-entropy points in text generated by an LLM with a high HVI and replace them using an LLM with a lower HVI, yielding desired results (cf. Section 5).
- We firmly believe that the HILT dataset and **HVI** measure will serve as valuable resources for future researchers interested in studying the hallucination behaviors of LLMs and seeking to design effective mitigation techniques. HVI will prove to be a useful tool for assessing the categorical impacts of these proposed mitigation techniques.

2 A Holistic View of the Hallucination Spectrum: its Types and Scales

The issue of hallucination garnered research attention as early as (Maynez et al., 2020). However, with the growing size of LLMs (*empirical evidence provided in Section 5*), there is a corresponding increase in LLMs' susceptibility to hallucination. Consequently, there is a growing interest within the research community to study and understand hallucination to design mitigation techniques.

Researchers have loosely defined hallucinations and studied various notions of hallucinations in isolation. A recent survey conducted by (Maynez et al., 2020), categorized hallucination into two limited classes: intrinsic and extrinsic. Another recent paper (Ladhak et al., 2023b), delved into an intriguing type of hallucination known as name-nationality category hallucination. Several other papers (Raunak et al., 2021; Maynez et al., 2020) have explored task-specific categories of hallucination, such as summarization, question-answering, machine translation, etc. However, we purposefully avoid confining our study to a specific task to study the nuances of hallucination. Our main contention is that hallucination can occur in any NLG task, necessitating a thorough examination based on the fundamental principles of text generation from a given prompt. The findings from this study can be applied and extended to various NLP tasks. With this in mind, this paper aims to offer a comprehensive categorization of hallucination, as outlined below.

2.1 Orientations of Hallucination

We introduce two primary orientations of hallucination: (i) Factual Mirage (FM) and (ii) Silver Lining (SL), defined and exemplified below.

2.1.1 Factual Mirage

Factual mirage (FM) is defined as the phenomenon wherein an LLM engages in hallucination or distortion of a given prompt that is factually correct. FM can potentially be subdivided into two distinct sub-categories.

MILD

Prompt: *Capital of France*

AI-generated text: ...Paris is also the world fashion capital...

Fact: Paris.

Intrinsic factual mirage (IFM): In the following example, the LLM is providing a correct response while adding additional supplementary

facts such as "*the world fashion capital*," resulting in distortion or hallucination, has also been described in (Cao et al., 2022).

Extrinsic factual mirage (EFM): EFM refers to the phenomenon where an LLM deviates from factual accuracy. For example:

ALARMING

Prompt: *Engineering effort to build Eiffel tower*

AI-generated text: ...Designed by Gustave Eiffel, it was inaugurated in 1889 to celebrate the 100th anniversary of the European Civil War...

Fact: Eiffel tower was built to celebrate the 100th anniversary of the French Revolution.

2.1.2 Silver Lining (SL)

Silver lining (SL) is defined as the phenomenon in which an LLM indulges in hallucination by conjuring an elaborate and captivating narrative based on a given prompt that is factually incorrect.

Intrinsic silver lining (ISL): ISL is the category when in some cases LLM does not generate a convincing story. As an example:

MODERATE

Prompt: *Cell phone numbers go public this month.*

AI-generated text: ...Starting this month, cell phone numbers will be made public as a result of the implementation of the Right to Information (RTI) policy...

Fact: This statement is false.

Extrinsic silver lining (ESL): When an LLM generates a highly detailed and persuasive narrative in response to a factually incorrect prompt, it falls under the category of Extrinsic Silver Lining.

ALARMING

Prompt: *Kamala Harris and Elon Musk are getting married*

AI-generated text: Elon Musk got married to Kamala Harris in California on Thursday. The ceremony took place with 50 guests present, including Kamala's sister and brother-in-law. Apart from Joe Biden all were present from the White House. Kamala's ex-husband played the saxophone at the ceremony. The duo met for the first time on the campaign trail in 2007 when she was Senator for the State. Elon proposed to Kamala Harris in August 2019 on a beach. Kamala accepted and has been dating him ever since...

Fact: The incident never happened!

2.2 Categories of Hallucination

We meticulously classify hallucination into six distinct types, defined and exemplified below.

Numeric Nuisance (NN): This issue occurs when an LLM generates numeric values related to past events, such as dates, ages, or monetary amounts, that are inconsistent with the actual facts.

MODERATE

Prompt: *Twitter is said to have laid off more workers.*

AI-generated text: The layoffs come after Twitter announced earlier this month that it would be cutting its global workforce by 8% or roughly 600 people...

Fact: Twitter laid off 200 people which is roughly 10%

Acronym Ambiguity (AA): This issue pertains to instances in which LLMs generate an imprecise expansion for an acronym.

MODERATE

Prompt: *RLHF in reinforcement learning*

AI-generated text: ...RLHF is Reward-free Learning from Human Feedback in reinforcement learning...

Fact: RLHF stands for *"Reinforcement Learning from Human Feedback"*

Generated Golem (GG): This issue arises when an LLM fabricates an imaginary personality in relation to a past event, without concrete evidence.

MODERATE

Prompt: *Earthquake in Turkey*

AI-generated text: ...One rescuer, Hasan Cetin, said he was motivated by the thought of the survivors he helped save. "The moment when a man or a child is pulled out of the ruins, the joy of seeing them come back to life is...

Fact: There is no evidence that the person Hasan Cetin exists.

Virtual Voice (VV): At times LLMs generate quotations attributed to either fictional or real characters without sufficient evidence to verify the authenticity of such statements.

ALARMING

Prompt: *Pfizer Press Release on COVID-19 vaccine*

AI-generated text: ...Pfizer emphasized that their vaccine demonstrated an impressive efficacy rate... Pfizer CEO said, "This is a giant leap for humanity.."...

Fact: Pfizer CEO never said this.

In this instance, the expression *giant leap for humanity* is quoted from Neil Armstrong's renowned historical statement upon stepping onto the moon.

Geographic Erratum (GE): This problem occurs when LLMs generate an incorrect location associated with an event. For example here in the

first of the following two examples, Pentagon is stated to be situated in India.

Time Wrap (TW): This problem entails LLMs generating text that exhibits a mashed fusion of events from different timelines. in the following case (2nd example) the U.S. president during the Ukraine-Russia war is Joe Biden, not Barack Obama, thus contradicting the factual reality.

ALARMING

Prompt: *Prime Minister of the United Kingdom visiting India*

AI-generated text: ...The visit of the United Kingdom's Prime Minister to India holds immense significance...During his journey, he will also pay a visit to the Pentagon where he is scheduled to meet with Smriti Irani, the Minister of...

Fact: Pentagon is the defense headquarters of the USA, located in Washington DC, USA – not in India.

ALARMING

Prompt: *USA on Ukraine war*

AI-generated text: ...U.S. President Barack Obama says the U.S. will not put troops in Ukraine...

Fact: The actual U.S. president during the Ukraine-Russia war is Joe Biden.

2.3 Degrees of Hallucination

We annotate the degree of hallucination using three levels: *mild, moderate,* and *alarming* (labeled as 0, 1, and 2 respectively). *Mild* indicates minor hallucination which is superficial in terms of its impact. *Moderate* indicates a level of hallucination that introduces facts that are either fictitious or tangential to the topic at hand. *Alarming* indicates added information pieces that bear a radical dissemblance from the topic fed via the prompt. Please refer to Appendix B for more details.

3 HILT: HallucInation eLiciTation dataset

HILT is a first-of-its-kind publicly available hallucination dataset. To construct this dataset, we have utilized two primary sources of data as prompts: (i) NYTimes tweets (NYT) (*factually correct* – FM) and (ii) the Politifact dataset (Politifact) (*factually incorrect* – SL). We selected 15 LLMs, based on the criteria delineated in Section 3.1, and used them to generate a total of 75,000 text passages, with each LLM producing 5,000 text prose entries. These entries were categorized as 2,500 each for FM and SL. The text prompts provided to these LLMs consisted of tweets from NYTimes and headlines sourced from the Politifact dataset. Table 1 reports detailed statistics about HILT.

Orientation → / Categories ↓	Factual Mirage (FM) IFM	Factual Mirage (FM) EFM	Silver Lining (SL) ISL	Silver Lining (SL) ESL
Time Wrap	1,650	4,950	2228	3342
Acronym Ambiguity	675	550	1830	1255
Generated Golem	5,550	9,300	2302	1819
Virtual Voice	14,100	13,950	5782	8712
Numeric Nuisance	2,025	5,250	3210	5760
Geographic Erratum	6,225	6,825	1232	4530
Total	30,225	40,825	33,168	25,418

Table 1: Statistics of the HILT dataset (total: 129K annotated sentences).

3.1 Choice of LLMs: Rationale and Coverage

We chose 15 contemporary LLMs that have exhibited exceptional results on a wide range of NLP tasks, including: (i) GPT-4 (OpenAI, 2023a), (ii) GPT-3.5 (OpenAI, 2022), (iii) GPT-3 (Brown et al., 2020), (iv) GPT-2 (Radford et al., 2019), (v) MPT (Wang et al., 2023), (vi) OPT (Zhang et al., 2022), (vii) LLaMA (Touvron et al., 2023), (viii) BLOOM (Scao et al., 2022), (ix) Alpaca (Taori et al., 2023), (x) Vicuna (Chiang et al., 2023), (xi) Dolly (databricks, 2023), (xii) StableLM (Liu et al., 2023), (xiii) XLNet (Yang et al., 2019), (xiv) T5 (Raffel et al., 2020), and (xv) T0 (Deleu et al., 2022). Appendix C.1 discusses additional details behind our selection criteria. Given the ever-evolving nature of the field, HILT and HVI benchmark leaderboards will remain accessible to the research community, fostering an environment of continuous updates and contributions.

3.2 Annotating Hallucination

For the annotation task of the 75,000 text snippets, we utilized Amazon Mechanical Turk (Amazon). We obtain sentence-level annotations for hallucination orientations and categories. We record four annotations per sentence and adopt the MACE tool (Hovy et al., 2013) to assess inter-annotator agreement and aggregate data. MACE has been empiri-

cally demonstrated to outperform majority voting, exhibiting superior performance. Appendix B offers a detailed account of the overall annotation process.

4 Hallucination Vulnerability Index (HVI)

Given the growing usage of LLMs and their likeliness to hallucinate, there exists no uniform evaluation metric to measure these LLMs' hallucinations. To address this gap, we define **HVI**, a comparative spectrum that allows us to evaluate and rank LLMs based on their vulnerability to producing hallucinations. HVI is calculated as in Eq. (1):

$$HVI_x = \frac{100}{U*2}\left[\sum_{x=1}^{U}(N(x)-N(EFM))*(1-P(EFM)+\delta_1)+ (N(x)-N(ESL))*(1-P(ESL)+\delta_2)\right] \quad (1)$$

When defining HVI, we take several factors into account. Firstly, not all sentences generated by an LLM are hallucinated, so it is important to determine the ratio of actual hallucinated sentences with the total number of sentences. In this context, we consider U as the total number of sentences and $N(x)$ as the total number of hallucinated sentences produced by an LLM. Secondly, LLMs can exhibit different characteristics, such as higher EFM or ESL tendencies, or they can have varying levels of overall hallucination. This notion is captured by introducing the terms $N(x)-N(EFM)$ and $N(x)-N(ESL)$ in the equation. It is worth noting that we did not consider variations of intrinsic hallucinations in HVI calculation, as they are relatively minor and exhibit lower vulnerability overall. Lastly, comparative measures are needed to rank LLMs based on their vulnerability to hallucination. This is achieved using multiplicative damping factors, δ_1 and δ_2, which are calculated based on $\mu \pm rank_x \times \sigma$. Initially, we calculate the HVI for all 15 LLMs, considering δ_1 and δ_2 as zero. With these initial HVIs, we obtain the mean (μ) and standard deviation (σ), allowing us to recalculate the HVIs for all the LLMs. The resulting HVIs are then ranked and scaled providing a comparative spectrum as presented in Fig. 3, similar to z-score normalization (Wikipedia_zscore) and/or min-max normalization (Wikipedia_min_max). Having damping factors enables easy exponential smoothing with a handful of data points, 15 in this case. Finally, for ease of interpretability, HVI is scaled between $0-100$.

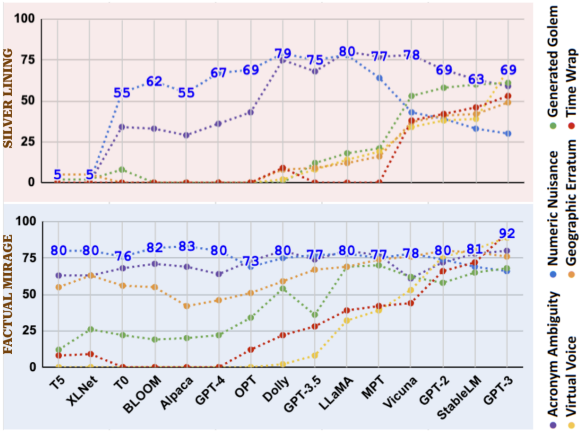

Figure 2: HVI for different hallucination categories across various LLMs.

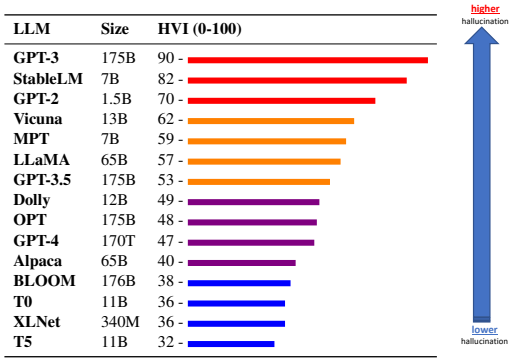

Figure 3: The HVI scale illustrates the hallucination tendencies exhibited by various LLMs.

Prompt	*USA's stance on Ukraine Russia war*
Hallucinated text	The official stance of the United States on the Russia-Ukraine war has been consistent in supporting Ukraine's sovereignty, territorial integrity, and the peaceful resolution of the conflict. The U.S. has condemned Russia's actions in annexing Crimea and its support for separatist movements in eastern Ukraine. Despite maintaining a diplomatic stance, the U.S. President Barack Obama stated that the United States would not deploy troops to Ukraine. Additionally, this lack of support extends to the withholding of financial and military assistance, including military equipment, training, and advisory support. This sentiment was reiterated by Paul Altman during a press review at the White House ...
Mitigated text $ENTROPY_{BB}$	The official stance of the government on the ongoing war has been consistent in supporting resolution of the conflict. The government has condemned Russia in country and its support for separatist movements in eastern country. Despite maintaining a neutral stance, the government stated that the country would not deploy troops to unknown. Additionally, this lack of support extends to the withholding of financial and resources, including administrative, technical, and legal. This sentiment was echoed by minister during a press review at the conference ...
Edit Suggestions $FACTUALITY_{GB}$	The official stance of the United States on the Russia-Ukraine war has been consistent in supporting Ukraine's sovereignty, territorial integrity, and the peaceful resolution of the conflict. The U.S. has condemned Russia's actions in annexing Crimea and its support for separatist movements in eastern Ukraine. Despite maintaining a diplomatic stance, U.S. President Barack Obama stated that the United States would not deploy troops to Ukraine. Additionally, this lack of support extends to the withholding of financial and military assistance, including military equipment, training, and advisory support. This sentiment was reiterated by Paul Altman during a press review at the White House ...

Figure 4: A hallucination example pre- and post-mitigation. A - hallucinated fragments, B - high entropy fragments, C - replaced text, D - highlighted text for no information found, and E - refuted text fragments by textual entailment. Appendix G contains more examples.

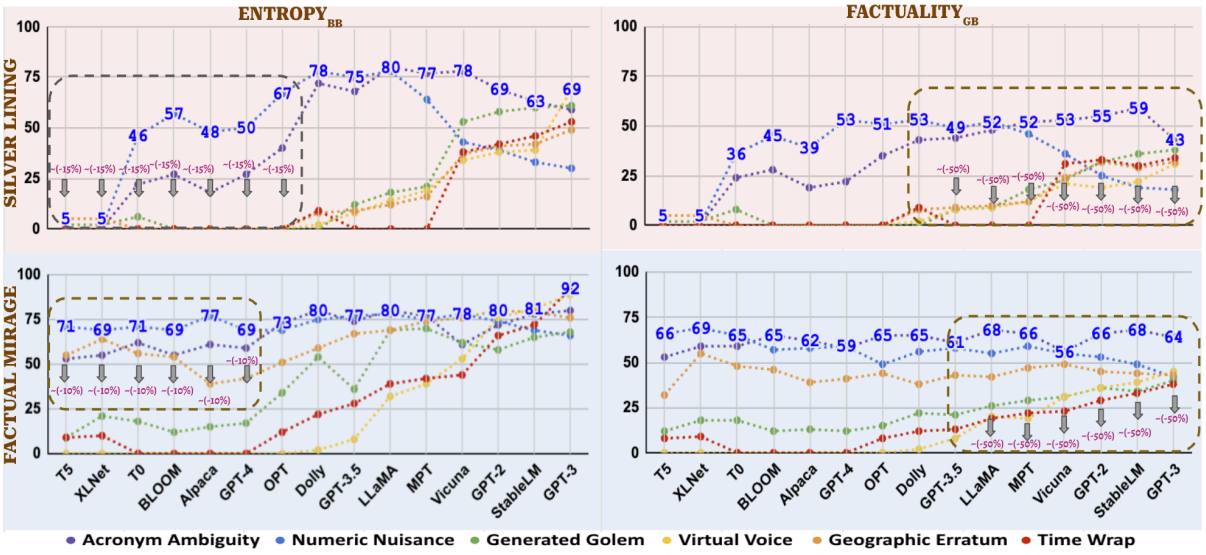

Figure 5: Impact of mitigation techniques across the various categories and types of hallucination. Evaluation strategy aka how did we decide there is no hallucination after applying mitigation techniques - cf. Appendix G.3.

Implications derived from HVI

- Larger LLMs without RLHF (Ziegler et al., 2019) are susceptible to both orientations of hallucination , as demonstrated in Fig. 3. To inspect the categorical changes in hallucination behavior for a particular LLM, please refer to the vertical axis of the HVI spectrum.
- Per our definitions, Numeric Nuisance and Acronym Ambiguity are relatively mild categories of hallucination. These categories exhibit a decrease in SL orientation as the LLM size increases. Conversely, more complex hallucination categories such as Time Wrap and Geographic Erratum tend to increase in prevalence. It is worth noting that Virtual Voice shows a significant jump from GPT-3.5 to GPT-4.
- For LLMs below a certain size threshold, namely T5, Dolly, etc., hallucination categories such as Generated Golem, Virtual Voice, and Geographic Erratum exhibit negligible occurrences.

5 Hallucination Mitigation Strategies

Thus far, two classes of approaches have been proposed to address the issue of hallucination: (i) preventing LLMs from hallucinating, which involves implementing strategies during the training and/or generation processes; (ii) mitigating hallucination after generation. (Manakul et al., 2023) introduced another taxonomy of classification, categorizing methods into *black-box* and *gray-box*. Factuality checks during and/or after generation without relying on external resources are known as black-box methods, while those using external resources are

referred to as gray-box methods.

A set of other hallucination mitigation techniques are as follows. Reranking the generated sample responses can be used to reduce hallucination (Dale et al., 2022). (Sridhar and Visser, 2022) discuss how hallucination in abstractive summarization can be reduced by improving beam search. Some recent mitigation techniques (Li et al., 2023; Mündler et al., 2023; Pfeiffer et al., 2023; Chen et al., 2023; Zhang et al., 2023b,a; Ladhak et al., 2023a; Manakul et al., 2023; Agrawal et al., 2023) show initial attempts at reducing hallucination.

Although the complete elimination of hallucination is a complex challenge, this paper explores two plausible directions for mitigation: (i) automatic and (ii) human-in-the-loop. The former is a black-box method where we identify high-entropy words in a given hallucinated text (generated by a high-HVI LLM) and replace them with predictions from another LLM (lower-HVI). The latter is a gray-box method that involves sentence-level fact-checking using textual entailment techniques. This method aims to identify sentences that are deemed susceptible, urging them for human review.

	ALBERT	BERT	DISTIL-ROBERTA	XLM-ROBERTA
ALBERT	6.72	3.26	**10.66**	6.40
BERT	4.70	7.56	7.98	7.22
DISTIL-ROBERTA	2.02	7.31	4.55	9.95
XLM-ROBERTA	2.26	6.28	1.70	4.78

Table 2: The row represents the LLM used for detecting high entropy words from GPT-3's output, while the column represents the LLM for replacing those words. **10.66** shows a maximum drop in overall hallucination.

5.1 High Entropy Word Spotting and Replacement (ENTROPY$_{BB}$): A Black-box approach

While detecting high entropy words may seem to be technically feasible, there is an inherent challenge that many modern LLMs are not open-source (their APIs are subscription-based). The feasible solution we propose here is the utilization of open-source LLMs to identify high entropy words. A lower HVI-based LLM is then used to replace the detected words (see Fig. 4). The outcomes of the detection and replacement strategies discussed earlier are presented in Table 2 for GPT-3. The results indicate that `albert-large-v2` (Lan et al., 2020) performs exceptionally well in detecting high entropy words in GPT-3-generated content. On the other hand, `distilroberta-base` (Sanh et al., 2019) demonstrates superior performance in replacing high entropy words, which in turn, manifests as a lower hallucination. An important consideration in our approach is the treatment of consecutive high-entropy words as a single entity. In such cases, these words are masked together before replacement. This strategy proves to be effective, particularly for hallucinations related to generated golems or acronym ambiguities (cf. Appendix G.1).

5.2 Factuality Check of Sentences (FACTUALITY$_{GB}$): A Gray-box approach

We use Google Search API (Search) to search for a given prompt, which has been utilized to generate the text and retrieve the top 20 documents. Then each sentence of AI-generated text has been validated either into *support*, *refute*, or *not enough information* using RoBERTa Large (Liu et al., 2019), a SoTA textual entailment model trained on the SNLI (Bowman et al., 2015) (cf. Appendix G.2). Inevitably, sentences with higher scores in the *refute* and *not enough information* categories are flagged for additional human checking. Empirically, we observe an overall alert rate of 26% on sentences generated by an LLM, implying 26% of the text required rewriting in order to mitigate.

Performance of ENTROPY$_{BB}$ vs. FACTUALITY$_{GB}$: Fig. 5 offers a comparative analysis of the proposed approaches. ENTROPY$_{BB}$ mainly addresses simpler hallucinations such as acronym ambiguity and numeric issues, while FACTUALITY$_{GB}$ is better suited for handling more complex cases. It is clear that a balanced combination of black-box and gray-box approaches is the inevitable future avenue (cf. Appendix G.4).

6 Conclusion and Future Avenues

The enthusiasm and achievements surrounding LLMs have led to their widespread adoption, and this trend is only expected to flourish. However, one of the most significant challenges faced by LLMs today is hallucination. In light of this, HILT benchmark and *Hallucination Vulnerability Index (HVI)* will continue to serve the wider scientific community and aid policy-makers. HILT benchmark and *HVI* will be publicly open for further collaborative updates. Two proposed mitigation techniques can serve as baselines.

7 Discussion and Limitations

Discussion: On June 14th, 2023, the European Parliament successfully passed its version of the EU AI Act (European-Parliament, 2023). Subsequently, a team of researchers from the Stanford Institute for Human-Centered Artificial Intelligence (HAI) embarked on investigating the extent to which Foundation Model Providers comply with the EU AI Act. Their initial findings are presented in the publication (Bommasani et al., 2023). In this study, the authors put forward a grading system consisting of 12 aspects for evaluating LLMs. These aspects include *(i) data sources, (ii) data governance, (iii) copyrighted data, (iv) compute, (v) energy, (vi) capabilities & limitations, (vii) risk & mitigations, (viii) evaluation, (ix) testing, (x) machine-generated content, (xi) member states, and (xii) downstream documentation*. The overall grading of each LLM can be observed in Figure 6. While this study is commendable, it appears to be inherently incomplete due to the ever-evolving nature of LLMs. Since all scores are assigned manually, any future changes will require a reassessment of this rubric, while HVI is auto-computable. Furthermore, we propose that HVI should be considered the most suitable category for assessing risk and mitigations, as well as the evaluation of machine-generated content.

Limitations: In this paper, we present a unique and extensive benchmark corpus for hallucination called HILT. We propose two main types of hallucination: (i) Factual Mirage and (ii) Silver Lining, each further divided into intrinsic and extrinsic subcategories. Additionally, we introduce six detailed categories of hallucination along with a measure of its intensity. We believe the following aspects require critical attention in future endeavors.

Limitation 1: For the sake of simplicity, we have only considered one category per sentence during annotation, although we acknowledge the presence of multi-class and multi-label instances. For instance, in the following example, there are two kinds of hallucination, namely Time Wrap and

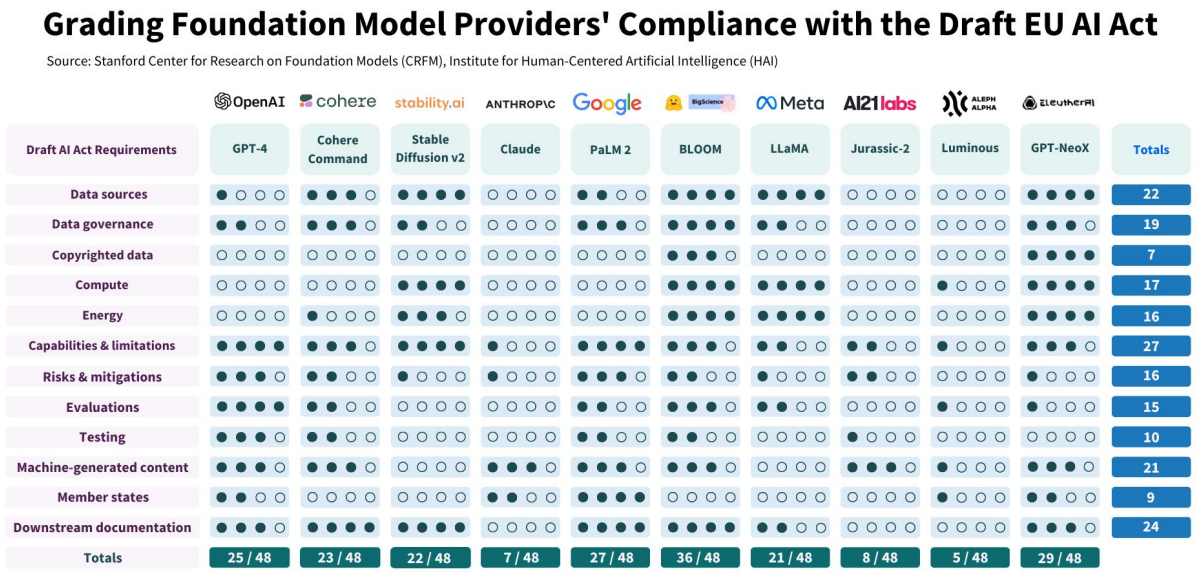

Figure 6: Grading of current LLMs as proposed by a report entitled *Do Foundation Model Providers Comply with the EU AI Act?* from Stanford University (Bommasani et al., 2023).

Numeric Nuisance present in the shown sentence. We would like to explore this direction in the immediate future.

ALARMING

Prompt: *Engineering effort to build Eiffel tower*

AI-generated text: ...Designed by Gustave Eiffel, it was inaugurated in 1889 to celebrate the 100th anniversary of the European Civil War, while construction began a decade prior to its inauguration....

Fact 1: Eiffel tower was built to celebrate the 100th anniversary of the French Revolution.
Fact 2: Eiffel Tower construction was started in 1887, not in 1879.

Limitation 2: While we have meticulously defined the categories of hallucination, we recognize the potential for new categories to emerge in the future with the advancement of LLMs. An instance of this is the introduction of name-nationality hallucination by (Ladhak et al., 2023b), where a person named Jung Lee is falsely attributed with French nationality. Although one could argue that Mr Lee may indeed be a French national, considering his birth and upbringing there, the authors confirm that no such individual exists. We posit that name-nationality hallucination falls under the sub-class of generated golems. It is plausible that a combination of our defined categories may exist, although we did not extensively studied these possibilities.

Limitation 3: For this study, we have chosen 15 contemporary LLMs. In the dynamic landscape of LLM development, new models are constantly emerging, and we acknowledge that our selection may not encompass all available options. Keeping this in mind, we will make the HELM benchmark and the *HVI* publicly accessible for collaborative updates and contributions.

Limitation 4: The FACTUALITY$_{\text{GB}}$ technique operates based on entailment, allowing it to distinguish between sentences containing different entities such as *Barack Obama* and *Joe Biden*. However, it is unable to differentiate sentences that involve similar entities like *AI* and *AAAI*. In contrast, the ENTROPY$_{\text{BB}}$ technique operates at the token level and is capable of handling cases like *1789* vs. *1889*. These distinctions become evident in the observed results.

8 Ethical Considerations

Through our experiments, we have uncovered the susceptibility of LLMs to hallucination. In developing HVI, we intend to provide a framework that can inform future research and policies in this domain. However, we must address the potential misuse of our findings by malicious entities who may exploit AI-generated text, such as creating indistinguishable fake news from human-written content. We vehemently discourage such misuse and strongly advise against it.

References

2023. The future of computational linguistics.

Abien Fred Agarap. 2019. Deep learning using rectified linear units (relu).

Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references?

Amazon. Amazon mechanical turk.

Rishi Bommasani, Kevin Klyman, Daniel Zhang, and Percy Liang. 2023. Do foundation model providers comply with the eu ai act?

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. The snli corpus.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023. Purr: Efficiently editing language model hallucinations by denoising language model corruptions.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Copyright-Office. 2023. Copyright registration guidance: Works containing material generated by artificial intelligence. Library of Congress.

David Dale, Elena Voita, Loïc Barrault, and Marta R. Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

databricks. 2023. Dolly.

Tristan Deleu, David Kanaa, Leo Feng, Giancarlo Kerg, Yoshua Bengio, Guillaume Lajoie, and Pierre-Luc Bacon. 2022. Continuous-time meta-learning with forward mode differentiation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

European-Parliament. 2023. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.

Forbes. 2023. Lawyer used chatgpt in court—and cited fake cases. a judge is considering sanctions.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

HuggingFace_InferenceAPI. Huggingface inference api.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023a. When do pre-training biases propagate to downstream tasks? a case study in

text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219, Dubrovnik, Croatia. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen Mckeown, and Tatsunori B Hashimoto. 2023b. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3198–3211.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Gary Marcus and Future of Life Institute. 2023. Pause giant ai experiments: An open letter.

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*, pages 1906–1919, Online. Association for Computational Linguistics.

Midjourney. 2022. https://www.midjourney.com.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation.

nVIDIA. 2023. https://nvidia.github.io/nemo/.

NYT. https://www.nytimes.com/topic/company/twitter.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Our approach to ai safety.

Jonas Pfeiffer, Francesco Piccinno, Massimo Nicosia, Xinyi Wang, Machel Reid, and Sebastian Ruder. 2023. mmt5: Modular multilingual pre-training solves source language hallucinations.

Politifact. https://www.politifact.com/.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Reuters. 2023. Alphabet shares dive after google ai chatbot bard flubs answer in ad.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Google Search. Google search api.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*.

Noam Shazeer. 2020. Glu variants improve transformer.

Arvind Krishna Sridhar and Erik Visser. 2022. Improved beam search for hallucination mitigation in abstractive summarization.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2022. Roformer: Enhanced transformer with rotary position embedding.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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. *arXiv preprint arXiv:2302.13971*.

Zhen Wang, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, Huan Sun, and Yoon Kim. 2023. Multitask prompt tuning enables parameter-efficient transfer learning. In *The Eleventh International Conference on Learning Representations*.

White-House. 2023. Blueprint for an ai bill of rights: Making automated systems work for the american people.

Wikipedia_Fleiss's_Kappa. Fleiss's kappa.

Wikipedia_Krippendorff's_Alpha. Krippendorff's alpha.

Wikipedia_min_max. Normalization.

Wikipedia_zscore. Normalization.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023a. How language model hallucinations can snowball.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023b. Mitigating language model hallucination with interactive question-knowledge alignment.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593.

Frequently Asked Questions (FAQs)

✻ This study explores the unintended, negative aspects of hallucination; how about the useful effects that arise as a result of hallucination?

➠ While hallucinating has beneficiary effects in some computer vision use cases, where a generative vision model could perform in-painting of an occluded content in an image or generate an image of a scenario it hasn't seen in its training set (for example, a generated image corresponding to the prompt, "water on Mars"), but it is usually undesirable in the context of the text. The downstream impact as a result of the model's is exacerbated by the fact that there is a lack of a programmatic method in the research community to distinguish the hallucinated vs. factually correct output. For this reason, this study focuses on characterizing the problem of hallucination particularly in the context of text.

✻ Why do you select those 15 large language models?

➠ We want to select several language models with varying parameter sizes for our experiments - ranging from large to small. Hence, the above chosen 14 models consist of large models like GPT-3 and smaller ones like T5 and T0.

✻ Why would extrinsic hallucination be riskier?

➠ According to the "extrinsic hallucination" definition, this kind of hallucination does not have any way to verify it from the source prompt. Hence, it is likely to be more harmful than the intrinsic ones.

✻ What is the purpose of constructing Factual Mirage and Silver Lining hallucination data?

➠ We want to show that hallucinations can happen in both cases, factually correct and incorrect prompts. Hence, in this paper, we construct an exhaustive dataset called 비포닌.

✻ Why do you select high-entropy points for mitigation techniques?

➠ High entropy points are more uncertain points in the context of text generation and hence, more likely places where the LLM hallucinates. Hence, our mitigation approach works by detecting and replacing such high entropy points.

✻ Why would HVI be a better hallucination evaluation metric for the LLMs (as compared to the existing ones like accuracy, precision, recall, F1, etc.)?

➠ Although the commonly used evaluation metrics like accuracy, precision, etc. can be used for downstream tasks, HVI can be more specifically used to determine the LLMs' hallucination tendency. HVI will serve as a uniform hallucination score for all the present and future LLMs.

✻ What are the insights on using black-box vs. gray-box models for mitigation hallucinations?

➠ Both black-box and gray-box models have their own advantages and disadvantages in terms of reducing hallucinations. Therefore, the choice of the appropriate method to minimize hallucination would be LLM- and task-dependent.

A Appendix

This section provides supplementary material in the form of additional examples, implementation details, etc. to bolster the reader's understanding of the concepts presented in this work.

B Annotation Process, and agreement

B.1 Pilot in-house annotation

Crowdsourcing platforms are widely recognized for their speed and cost-effectiveness in annotation tasks. However, it is important to note that they can also introduce noise or inaccuracies in the annotations. To mitigate this, prior to utilizing crowdsourcing services, we conducted an in-house annotation process involving 2,000 samples. These samples included prompts and generated text snippets from five different LLMs. This in-house annotation process served two purposes: firstly, it allowed us to formulate comprehensive annotation guidelines, and secondly, it helped us develop an annotation interface tailored to our specific needs. By undertaking this internal annotation process, we aimed to ensure the quality and reliability of the annotations before moving on to crowdsourcing.

B.2 Annotation Steps

When annotating an AI-generated text snippet, we follow a sentence-wise approach. Our annotation process involves three layers of annotation: (i) **Orientation:** This layer captures the orientation of hallucinations. (ii) **Category:** This layer classifies the category of hallucination, and (iii) **Degree:** This layer quantifies the intensity or magnitude of hallucination. By employing these three layers, we aim to provide a comprehensive and detailed annotation for hallucination in AI-generated text.

Algorithm 1: Annotation Guidelines

1 Split the paragraph into a list of sentences.
2 Annotate the `orientation` of hallucination as *intrinsic* or *extrinsic*.
3 Annotate the `category` of hallucination.
4 Annotate the `degree` of hallucination.

- **Step 1:** In order to analyze the legitimacy of an AI-Generated paragraph and identify any potential hallucination, we begin with a sentence-level approach. We split the paragraph into individual sentences ensuring that each sentence is distinct and well separated from the others. Each sentence undergoes rigorous scrutiny to determine its legitimacy. This involves the identification of the type of hallucination, the category of hallucination, and the degree of hallucination.
- **Step 2:** In this step, we identify whether the sentence has no hallucination, intrinsic hallucination, or extrinsic hallucination. The absence of both intrinsic and extrinsic hallucination implies no hallucination. To identify whether the sentence has intrinsic hallucination or extrinsic hallucination, we refer to the definitions in Section 2. We annotate each sentence using the annotations listed in Table 3
- **Step 3:** In this step, we identify whether the detected hallucinated sentences of the previous step belong to any of the categories mentioned in Fig. 1. To identify the categories we refer to the definitions mentioned in Section 2. If the hallucinated sentence does not fall under any of the identified categories,

it implies a miscellaneous category. Once we have identified the category, we annotate each sentence using the annotations listed in Table 3.

- **Step 4:** This step involves categorizing the degree of hallucination as mild, moderate, or alarming, based on the level of delusional information in the sentence. A high degree refers to completely delusional information, a moderate degree to partially delusional information, and a low degree to minimal delusional information. Once we have identified the degree of hallucination, we annotate it as listed in Table 3.

B.3 Web Interface for Annotation

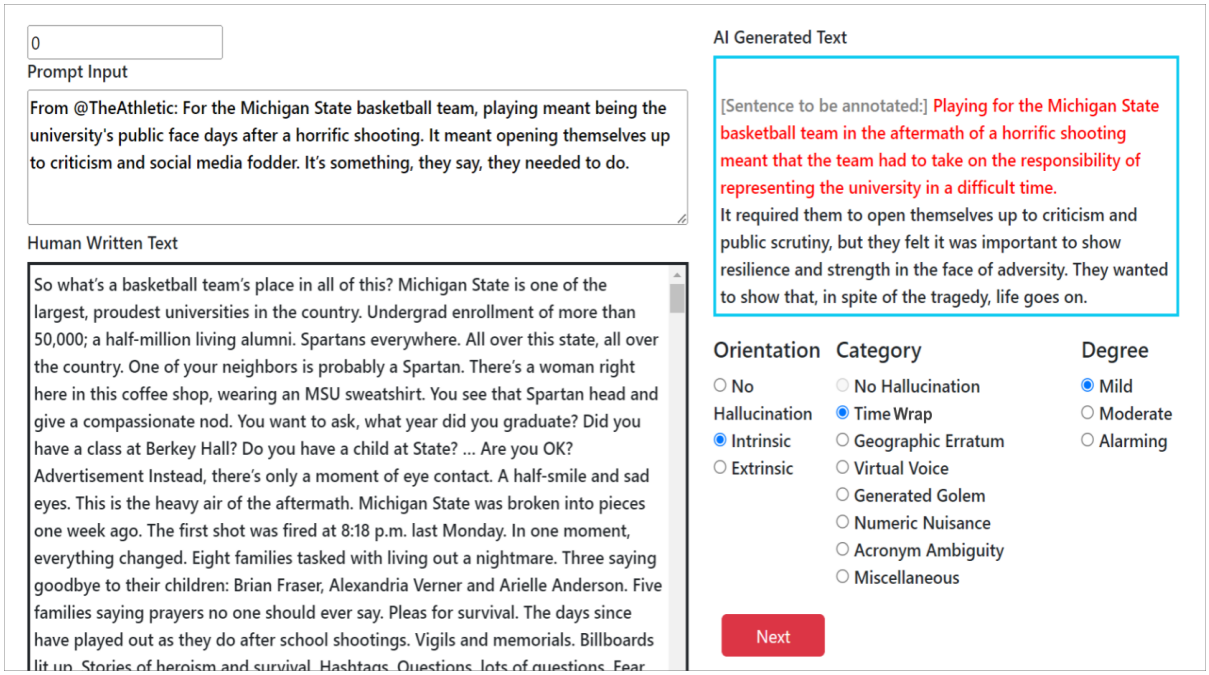

Figure 7: Web interface used to annotate the HILT dataset using Amazon Mechanical Turk.

In order to facilitate the annotation process for the annotators, it is crucial to provide them with a user-friendly interface that enables easy navigation. Fig. 7 shows our annotation web interface used to construct the HILT dataset. The interface is designed to offer a comprehensive view to the annotator. For instance, at the top of the interface, the actual prompt used for generating the text snippet is displayed. Directly below the prompt, the complete AI-generated text is shown. On the right-hand side, the sentence breakup is presented, with the currently selected sentence highlighted in red. Below the sentence breakup, all the relevant categories are displayed as radio buttons, allowing the annotators to easily annotate each category. This interface aims to enhance the efficiency and effectiveness of the annotation process. We have gone through a few rounds of iterations before finalizing the current version of the web interface.

B.4 Selecting quality annotators on AMT

It is widely acknowledged that platforms like AMT can be noisy, making the selection of high-quality annotators a critical step in ensuring accurate annotations. The in-house annotation of 2,000 data points

played a significant role in achieving this goal. To identify reliable annotators, we initiated a pilot task and only selected those with an accuracy rate of over 90% based on our in-house annotated dataset.

Another crucial consideration when annotating data on crowdsourcing platforms is the compensation offered to annotators. While offering too little may deter interest, excessively high wages may attract undesirable spammers. Achieving the ideal balance required several rounds of iteration to determine an appropriate compensation scheme.

By carefully addressing factors such as selecting qualified annotators and establishing suitable compensation rates, we improved the quality of annotations obtained from crowdsourcing platforms.

Orientation	Annotation	Category	Annotation	Degree	Annotation
No Hallucination	0	No Hallucination	0	Mild	0
Intrinsic Hallucination	1	Time Wrap	1	Moderate	1
Extrinsic Hallucination	2	Geographic Erratum	2	Alarming	2
		Virtual Voice	3		
		Generated Golem	4		
		Numeric Nuisance	5		
		Acronym Ambiguity	6		
		Miscellaneous	7		

Table 3: Annotations for 퍼포넌트: Orientation, Category, and Degree. We follow a three-level annotation process where we first determine whether there is a hallucination or not, and if yes, then what orientation. We assign labels 0, 1, and 2 for "no", "intrinsic", and "extrinsic" hallucination. Next, we determine the category of hallucination and label them ranging from 0 through 7, where 7 could be a miscellaneous case where the hallucination does not belong to any category. Finally, the degrees are labeled with 0, 1, and 2.

B.5 Inter-Annotator Agreement

We report Fleiss's kappa (κ) (Wikipedia_Fleiss's_Kappa) and Krippendorff's alpha (α) (Wikipedia_Krippendorff's_Alpha) scores (see Tables 4 and 5) to access the reliability of agreement between the three annotators[1]. We compute agreement on 10% of total annotated entities and obtain substantial to almost perfect agreement in all three annotation types in both datasets, namely NYT and Politifact. We have obtained nearly or more than 80% agreement in the case of `orientation` and `category`. The agreement on `degree` exhibits slight variation, as it relies on the subjective assessment of the percentage of hallucination in the sentence. This interpretation of percentage tends to differ among individuals. We report both Fleiss's kappa and Krippendorff's alpha score because Fleiss's kappa is a statistical measure that allows us to find agreement among multiple annotators and Krippendorff's alpha is a statistical measure that allows us to handle various types of data like nominal (`orientation` and `category`) and ordinal (`degree`).

	Fleiss's kappa	Krippendorff's alpha
Orientation	0.7911	0.8146
Category	0.7846	0.8499
Degree	0.7473	0.7274

Table 4: Inter-annotator scores for NYT dataset.

	Fleiss's kappa	Krippendorff's alpha
Orientation	0.7587	0.8436
Category	0.8755	0.9328
Degree	0.6182	0.5455

Table 5: Inter-annotator scores for Politifact dataset.

[1]Three graduate students

C Details on chosen LLMs

C.1 Criteria for choosing LLMs

Beyond the primary criteria for choosing performant LLMs, our selection was meant to cover a wide gamut of LLMs that utilize a repertoire of recent techniques under the hood that have enabled their exceptional capabilities, namely:

- FlashAttention (Dao et al., 2022) for memory-efficient exact attention.
- Multi-Query Attention (Shazeer, 2019) for memory bandwidth efficiency.
- SwiGLU (Shazeer, 2020) as the activation function instead of ReLU (Agarap, 2019).
- ALiBi (Press et al., 2022) for larger context width.
- RMSNorm (Zhang and Sennrich, 2019) for per-normalization.
- RoPE (Su et al., 2022) to improve the expressivity of positional embeddings, etc.

C.2 Details on Large Language Models

Model details are given in Table 6. We use HuggingFace (Wolf et al., 2020) and OpenAI for implementing the large language models to generate the dataset.

D HILT - Prompt sources

We used two datasets to curate HILT: (i) New York Times Tweets (NYT) for factually correct and incorrect prompts and (ii) Politifact dataset (Politifact).

E What is a high entropy vs. low entropy word?

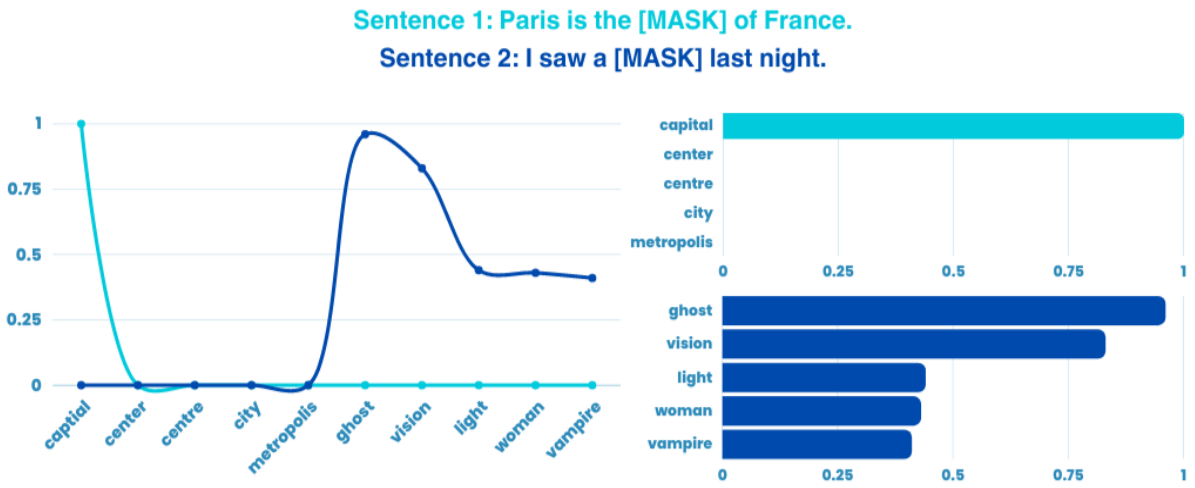

Figure 8: High entropy word vs. low entropy word - a side-by-side illustration.

In the context of language modeling, a high entropy word refers to a word or token that has a high level of uncertainty or unpredictability in its occurrence. In other words, it is a word that is relatively rare or has a low probability of appearing in a given context. Entropy is often used to quantify the level of

LLMs	Parameter size	LLM used in this paper	Details
GPT-3	175B	gpt3	**GPT-3 (Brown et al., 2020)** is an autoregressive language model by OpenAI. It is a decoder-only transformer model with a size of 175 billion parameters.
StableLM	7B	stablelm-base-alpha-7b	**StableLM (Liu et al., 2023)** The Alpha version of the model is available in 3 billion and 7 billion parameters.
GPT-2	1.5B	gpt2	**GPT-2 (Radford et al., 2019)** is a large transformer-based language model with 1.5 billion parameters. It is trained on a WebText dataset consisting of 8 million web pages.
Vicuna	13B	eachadea//vicuna-13b-1.1	**Vicuna (Chiang et al., 2023)** is created by fine-tuning a LLaMA base model using approximately 70K user-shared conversations gathered from ShareGPT.com
MPT	7B	mosaicml/mpt-7b	**MPT (Wang et al., 2023)** is a part of the family of MosaicPretrainedTransformer(MPT) models, which use a modified transformer architecture optimized for efficient training and inference.
LLaMA	65B	decapoda-research/llama-65b-hf	**LLaMa (Touvron et al., 2023)** is a collection of foundation language models varying from 7B to 65B parameters. It is trained on trillions of tokens. LLaMA-65B outperforms GPT-3 (175B) on most benchmarks.
GPT-3.5	175B	gpt3.5 (text-davinci-003)	**GPT-3.5 (OpenAI, 2023b)** is a sub-class of GPT-3 model family. It has 3 variants each with 1.3B, 6B, and 175B parameters.
Dolly	12B	dolly-v2-12b	**Dolly (databricks, 2023)** is an instruction-following large language model trained on the Databricks machine learning platform.
OPT	175B	opt-175B	**OPT (Zhang et al., 2022)** is a decoder-only pre-trained transformer model ranging from 125M to 175B parameters. Although the performance of OPT-175B is comparable to GPT-3, it only requires 1/7th the carbon footprint to develop.
GPT-4	170T	gpt4	**GPT-4 (OpenAI, 2023a)** was released by OpenAI in 2023. It is a large multimodal model that shows human-level performance on various professional and academic benchmarks.
Alpaca	7B	chainyo/alpaca-lora-7b	**Alpaca (Taori et al., 2023)** is a language model fine-tuned using supervised learning from a LLaMA 7B model on 52K instruction-following demonstrations.
BLOOM	176B	bigscience/bloom	**BLOOM (Scao et al., 2022)** is similar to GPT-3 (auto-regressive model for next token prediction). However, it has been trained on 46 different languages and 13 programming languages.
T0	11B	bigscience/T0	**T0 (Deleu et al., 2022)** is trained on a diverse set of tasks and prompts leading to increased robustness to the prompt wording. T0 outperforms or matches GPT-3, which is 16x larger in size and has 100s of billions of parameters.
XLNet	340M	xlnet-large-cased	**XLNet (Yang et al., 2019)** is a generalized autoregressive pretraining model that (1) enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order and (2) overcomes the limitations of BERT because of its own autoregressive formulation.
T5	11B	t5-11b	**T5 (Raffel et al., 2020)** is an encoder-decoder model pre-trained on a multi-task combination of unsupervised and supervised tasks Each task is converted into a text-to-text format.

Table 6: HuggingFace and OpenAI links for all LLMs.

unpredictability associated with generating specific words or tokens. When a language model encounters a high entropy word, it means that the model has greater difficulty in accurately predicting or generating that word based on the context or preceding words. High entropy words are often less frequent in the training data or have complex patterns of occurrence. For example, in a language model trained on news articles, words like "pneumonoultramicroscopicsilicovolcanoconiosis" (a technical term for a lung disease) would likely have high entropy, as they are infrequent and occur in specific contexts.

On the other hand, a low entropy word refers to a word or token that has a relatively high predictability or a limited range of potential next words given the context. In other words, it is a word that occurs frequently and is highly expected in a specific context. When a language model encounters a low entropy word, it means that the model has a higher confidence or accuracy in predicting or generating that word based on the context or preceding words. For example, in a language model trained on English text, common words like "the," "and," or "is" have low entropy because they occur frequently and are highly predictable in many contexts. These words tend to have a limited set of likely next words based on the preceding context. Fig. 8 illustrates an example of the entropy distribution over the set of words in an input sentence. In sentence 1, `[MASK]` token has low entropy since `capital` is the highest probable word in that sentence. However, `[MASK]` token in sentence 2 has high entropy since it is quite uncertain as to what the masked token could be. The probability distributions for this illustration are created using the HuggingFace Inference API (HuggingFace_InferenceAPI).

F HVI vs. LLMs size for different LLMs: An insight from

There is a general observation that larger language models (LLMs) may exhibit a higher tendency towards generating hallucinations or producing outputs that deviate from factual or coherent information. However, it's important to note that the relationship between LLM size and hallucination is not necessarily a direct correlation, but rather a consideration based on certain factors such as (a) training data quality, (b) lack of explicit training on facts, and (c) overconfidence in generated responses. A noteworthy pattern that emerges is that LLMs without RLHF (Reinforcement Learning from Human Feedback) (Ziegler et al., 2019) tend to exhibit a higher tendency for hallucination. Although we did not extensively evaluate this phenomenon, we have a keen interest in investigating it further in the near future. Fig. 9 shows how HVI and LLM size are correlated and hence, larger LLM tends to hallucinate more.

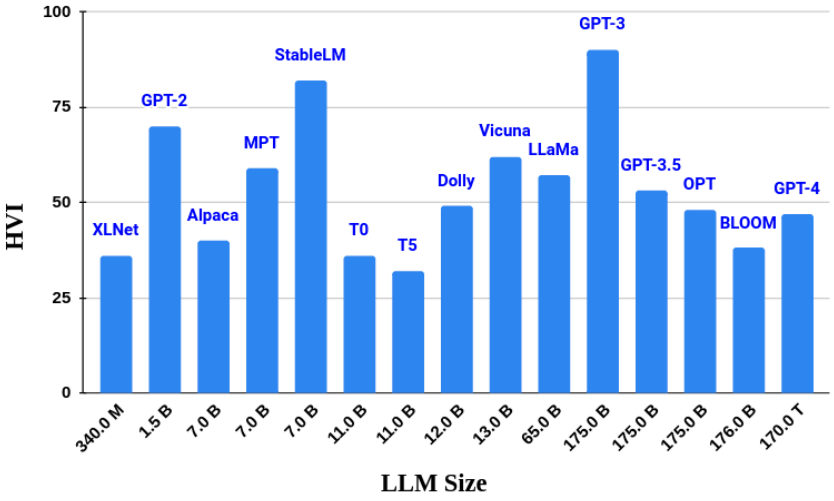

Figure 9: *HVI* vs. *LLM size* for different LLMs.

G Mitigation Techniques: The specifics

In addition to the Fig. 4 provided in Section 5, we now present two additional illustrative examples in Table 7 and Table 8 to demonstrate the practical application of two different mitigation techniques.

Prompt	*Rescue workers in Turkey are battling exhaustion as the hope of finding more earthquake survivors dwindles. Remembering those they have saved spurs them to keep searching. A powerful new quake on Monday complicated their task.*
Hallucinated text	The 7.0 magnitude quake that rocked the Aegean Sea was Turkey's deadliest in a decade, killing at least 109 people and injuring more than 1,000. It toppled dozens of buildings in the cities of Izmir and Mugla and left dozens of people trapped beneath the rubble. Rescue crews have worked tirelessly all week, often in the dark and in perilous conditions. At the same time, they have had to cope with the emotional toll of finding bodies, many of them children, in the debris. One rescuer, Hasan Cetin, said he was motivated by the thought of the survivors he helped save. "The moment when a man ora child is pulled out of the ruins, the joy of seeing them come back to life is indescribable." Cetin told The Associated Press "This is why we keep on going." The rescuers have been buoyed by the support of local people, who have donated food, water, and other supplies. Turkey's Health Minister Fahrettin Koca said Monday that the rescue efforts would continue until the last victim was found.
Mitigated text ENTROPY$_{BB}$	The high magnitude quake that shook the south was country's deadliest in a decade killing at least number people and injuring more than number. It affected a lot of buildings in the cities of unknown and unknown and left dozens of people trapped beneath the rubble. Rescue crews have worked tirelessly all week often in the dark and in severe conditions. At the same time, they have had to cope with the trauma of finding bodies many of them children in the unknown. Local rescuer person said he was motivated by the thought of the survivors he helped save "The moment when a man or a child is pulled out of the dust the fear of seeing them come back to life is indescribable". Someone told The media "This is why we keep on going" The rescuers have been obliged by the support of local people who have donated food water and other supplies Country's Health Minister also said friday that the rescue efforts would continue until the last victim was found.
Edit Suggestions FACTUALITY$_{GB}$	The 7.0 magnitude quake that rocked the Aegean Sea was Turkey's deadliest in a decade, killing at least 109 people and injuring more than 1,000. It toppled dozens of buildings in the cities of Izmir and Mugla and left dozens of people trapped beneath the rubble. Rescue crews have worked tirelessly all week, often in the dark and in perilous conditions. At the same time, they have had to cope with the emotional toll of finding bodies, many of them children, in the debris. One rescuer, Hasan Cetin, said he was motivated by the thought of the survivors he helped save. "The moment when a man or a child is pulled out of the ruins, the joy of seeing them come back to life is indescribable.". Cetin told The Associated Press "This is why we keep on going." The rescuers have been buoyed by the support of local people, who have donated food, water, and other supplies. Turkey's Health Minister Fahrettin Koca said Monday that the rescue efforts would continue until the last victim was found.

Table 7: A hallucination example pre- and post-mitigation. A - hallucinated fragments, B - high entropy fragments, C - replaced text, D - highlighted text for no information found, and E - refuted text fragments by textual entailment.

Prompt	*There are more oil rigs in operation in the United States than the rest of the world combined.*
Hallucinated text	This is not true. According to the International Association of Oil and Gas Producers, as of the start of 2020, there were a total of 4,932 oil rigs in operation worldwide, with the United States having 1,094 of them. This means that the US has 22.1% of the world's total oil rigs in operation.
Mitigated text ENTROPY$_{BB}$	This is not true. According to the company as of the start of year there were a total of number wells in operation worldwide with the unknown having number of them. This means that the country has number of the world's total wells in operation.
Edit Suggestions FACTUALITY$_{GB}$	This is not true. According to the International Association of Oil and Gas Producers, as of the start of 2020, there were a total of 4,932 oil rigs in operation worldwide, with the United States having 1,094 of them. This means that the US has 22.1% of the world's total oil rigs in operation.

Table 8: A hallucination example pre- and post-mitigation. A - hallucinated fragments, B - high entropy fragments, C - replaced text, D - highlighted text for no information found, and E - refuted text fragments by textual entailment.

G.1 ENTROPY$_{BB}$

Building upon Section 5, Tables 9 to 23 illustrate the examples of the nuanced categorization of hallucination proposed in the paper.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	6.72	3.26	**10.66**	6.40
bert-base-uncased	4.70	7.56	7.98	7.22
distilroberta-base	2.02	7.31	4.55	9.95
xlm-roberta-large	2.26	6.28	1.70	4.78

Table 9: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **GPT-3**. **10.66** is the maximum drop in overall hallucination detected with **albert-large-v2** and replaced with **distilroberta-base**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	6.80	5.40	**8.40**	5.10
bert-base-uncased	2.70	2.60	2.60	2.50
distilroberta-base	4.90	4.40	5.10	3.30
xlm-roberta-large	2.80	2.50	2.70	2.60

Table 10: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **StableLM**. **8.40** is the maximum drop in overall hallucination detected with **albert-large-v2** and replaced with **distilroberta-base**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	4.22	5.14	8.34	8.00
bert-base-uncased	2.12	4.72	**8.87**	7.77
distilroberta-base	3.88	5.93	6.92	5.22
xlm-roberta-large	2.85	7.65	2.87	3.00

Table 11: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **GPT-2**. **8.87** is the maximum drop in overall hallucination detected with **bert-base-uncased** and replaced with **distilroberta-base**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	8.50	7.80	**8.80**	7.90
bert-base-uncased	7.20	7.20	7.10	7.30
distilroberta-base	6.20	7.00	8.70	7.20
xlm-roberta-large	5.70	6.40	5.60	6.50

Table 12: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **Vicuna**. **8.80** is the maximum drop in overall hallucination detected with **albert-large-v2** and replaced with **distilroberta-base**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	8.00	7.00	**9.00**	7.70
bert-base-uncased	6.30	8.30	8.00	5.40
distilroberta-base	7.00	7.00	3.00	6.00
xlm-roberta-large	8.30	8.20	7.00	7.00

Table 13: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **MPT**. **9.00** is the maximum drop in overall hallucination detected with **albert-large-v2** and replaced with **distilroberta-base**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	**9.90**	9.30	9.10	8.80
bert-base-uncased	6.80	6.70	6.10	6.00
distilroberta-base	8.00	7.60	8.20	7.20
xlm-roberta-large	5.00	4.80	4.90	4.00

Table 14: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **LLaMA**. **9.90** is the maximum drop in overall hallucination detected with **albert-large-v2** and replaced with **albert-large-v2**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	8.10	9.27	2.51	9.17
bert-base-uncased	9.02	5.83	2.71	2.24
distilroberta-base	4.57	8.98	7.74	9.17
xlm-roberta-large	9.48	**9.54**	4.41	2.36

Table 15: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **GPT-3.5**. **9.54** is the maximum drop in overall hallucination detected with **xlm-roberta-large** and replaced with **bert-base-uncased**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	6.10	6.30	**8.20**	6.00
bert-base-uncased	4.10	4.10	4.70	4.00
distilroberta-base	5.10	5.70	5.30	5.10
xlm-roberta-large	3.30	4.00	4.20	4.40

Table 16: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **Dolly**. **8.20** is the maximum drop in overall hallucination detected with **albert-large-v2** and replaced with **distilroberta-base**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	5.18	4.70	**8.58**	6.25
bert-base-uncased	2.41	3.09	3.06	2.40
distilroberta-base	4.52	4.61	4.98	3.78
xlm-roberta-large	2.92	2.81	2.84	2.77

Table 17: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **OPT**. **8.58** is the maximum drop in overall hallucination detected with **albert-large-v2** and replaced with **distilroberta-base**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	6.31	9.91	**10.10**	7.75
bert-base-uncased	4.49	4.47	7.60	5.20
distilroberta-base	9.17	5.55	5.81	4.37
xlm-roberta-large	3.67	4.35	3.92	2.46

Table 18: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **GPT-4**. **10.10** is the maximum drop in overall hallucination detected with **albert-large-v2** and replaced with **distilroberta-base**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	**9.80**	9.50	9.20	**9.80**
bert-base-uncased	6.20	7.80	7.00	6.50
distilroberta-base	9.00	8.30	8.50	8.30
xlm-roberta-large	5.70	5.70	5.70	4.00

Table 19: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **Alpaca**. **9.80** is the maximum drop in overall hallucination detected twice with **albert-large-v2** and replaced with **albert-large-v2** and **xlm-roberta-large**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	6.60	**7.80**	**7.80**	**7.80**
bert-base-uncased	1.20	2.20	3.40	3.30
distilroberta-base	6.60	6.60	6.60	6.60
xlm-roberta-large	3.40	2.30	3.30	3.30

Table 20: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **Bloom**. **7.80** is the maximum drop in overall hallucination detected thrice with **albert-large-v2** and replaced with **bert-base-uncased**, **distilroberta-base** and **xlm-roberta-large**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	8.00	**8.40**	8.00	6.10
bert-base-uncased	6.90	6.10	6.00	5.40
distilroberta-base	6.90	6.00	6.70	6.40
xlm-roberta-large	5.60	5.60	5.90	4.10

Table 21: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **T0**. **8.40** is the maximum drop in overall hallucination detected with **albert-large-v2** and replaced with **bert-base-uncased**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	8.00	**9.00**	8.00	**9.00**
bert-base-uncased	6.00	6.00	6.00	6.00
distilroberta-base	8.00	8.00	8.00	8.00
xlm-roberta-large	7.00	7.00	6.00	6.00

Table 22: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **XLNet**. **9.00** is the maximum drop in overall hallucination detected twice with **albert-large-v2** and replaced with **bert-base-uncased** and **xlm-roberta-large**.

	albert-large-v2	bert-base-uncased	distilroberta-base	xlm-roberta-large
albert-large-v2	7.00	**7.90**	7.00	6.20
bert-base-uncased	6.50	6.40	5.70	5.70
distilroberta-base	7.20	**7.90**	7.00	6.40
xlm-roberta-large	5.60	6.50	5.80	5.80

Table 23: Overall drops in hallucination by 16 combinations of 4 LLMs with the rows having the LLMs which detected the high entropy words and the corresponding columns with the LLMs which replaced those words generated by **T5**. **7.90** is the maximum drop in overall hallucination detected with **albert-large-v2** and **distilroberta-base** and replaced in both the cases with **bert-base-uncased**.

G.2 FACTUALITY$_{GB}$

According to (Manakul et al., 2023), gray-box model **does** require output token-level probabilities. In the illustration below in Fig. 10, we describe our second mitigation technique, **FACTUALITY$_{GB}$**. For a given prompt, we have AI-generated text (from our HILT benchmark). In this technique, the prompt is provided to Google Search API to retrieve the top 20 most relevant search results. For a total of n sentences across all these 20 results, we use a similarity measure to rank them based on their relevance to the prompt. We then extract the top 20 similar sentences to the prompt. Next, for each sentence (in the m sentences in the AI-generated text) and the top 20 ranked sentences, we use a textual entailment model. We check for the trustworthiness of these sentences one by one by using the entailment score between the sentence and the top 20 ranked sentences. Based on their entailment scores, we classify the AI-generated text into three categories: (i) *support*, (ii) *refute*, and (iii) *not enough information*.

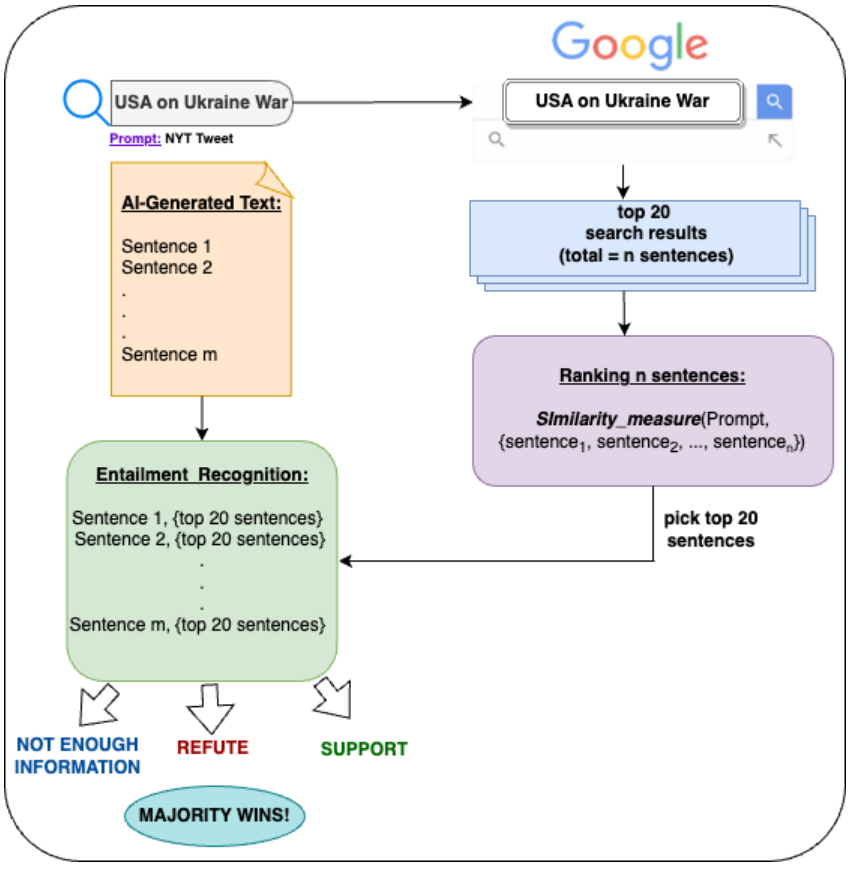

Figure 10: Hallucination mitigation using FACTUALITY$_{GB}$ - textual entailment on prompt and external documents.

G.3 Evaluation strategy - how to determine no hallucination after mitigation?

In order to assess the absence of hallucination following the implementation of the mitigation techniques ENTROPY$_{BB}$ and FACTUALITY$_{GB}$, a random sample of 2,000 data points was taken. This sample included 500 instances each of IFM, EFM, ISL, and ESL, ensuring a well-balanced distribution of the six hallucination categories within the data. Following the implementation of the ENTROPY$_{BB}$ method,

which involved replacing words and phrases, we conducted a manual evaluation of the 2,000 samples. This evaluation, carried out by six annotators, aimed to assess whether hallucination was alleviated or not.

For the FACTUALITY$_{GB}$ method, we assumed that if the sentences were rewritten by humans, there would be no presence of hallucination. Therefore, for the highlighted sentences, hallucination was deemed waived. Results of this evaluation are reported in Fig. 5 and Table 24.

G.4 Performance of ENTROPY$_{BB}$ vs. FACTUALITY$_{GB}$

Fig. 5 and Table 24 give a relative analysis of our two proposed mitigation techniques described in Section 5. We report the actual values in the Table 24. Our empirical findings indicate that ENTROPY$_{BB}$ technique primarily tackles less complex hallucination categories such as acronym ambiguity and numeric issues. However, FACTUALITY$_{GB}$ technique is more applicable for dealing with more complex cases of hallucinations. Therefore, it is quite evident that a combination of both black- and gray-box methods would be the future direction of research.

		Numeric Nuisance			Acronym Ambiguity			Generated Golem			Virtual Voice			Geographic Erratum			Time Wrap		
		Before	E_{BB}	F_{GB}	Before	E_{BB}	F_{GB}	Before	E_{BB}	F_{GB}	Before	E_{BB}	F_{GB}	Before	E_{BB}	F_{GB}	Before	E_{BB}	F_{GB}
SILVER LINING	T5	0	0	0	0	0	0	2	2	2	0	0	0	5	5	5	0	0	0
	XLNet	0	0	0	0	0	0	2	2	2	0	0	0	5	5	5	0	0	0
	T0	34	22	24	55	46	36	8	6	8	0	0	0	0	0	0	0	0	0
	BLOOM	33	27	28	62	57	45	0	0	0	0	0	0	0	0	0	0	0	0
	Alpaca	29	18	19	55	48	39	0	0	0	0	0	0	0	0	0	0	0	0
	GPT-4	36	27	22	67	50	53	0	0	0	0	0	0	0	0	0	0	0	0
	OPT	43	40	35	69	67	51	0	0	0	0	0	0	0	0	0	0	0	0
	Dolly	75	72	43	79	78	53	0	0	0	2	2	2	8	8	8	9	9	9
	GPT-3.5	68	68	44	75	75	49	12	12	8	8	8	8	9	9	9	0	0	0
	LLaMA	80	80	48	78	78	52	18	18	9	14	14	9	12	12	10	0	0	0
	MPT	77	77	52	64	64	46	21	21	18	19	19	12	16	16	12	0	0	0
	Vicuna	78	78	53	43	43	36	53	53	23	34	34	21	37	37	24	38	38	31
	GPT-2	69	69	55	39	39	25	58	58	32	38	38	19	41	41	32	42	42	33
	StableLM	63	63	59	33	33	19	60	60	36	39	39	22	42	42	29	46	46	30
	GPT-3	59	59	43	30	30	18	61	61	38	69	69	31	49	49	32	53	53	34
FACTUAL MIRAGE	T5	63	53	53	80	71	66	12	9	12	0	0	0	55	55	32	8	9	8
	XLNet	63	55	59	80	69	69	26	21	18	0	0	0	63	64	55	9	10	9
	T0	68	62	59	76	71	65	22	18	18	0	0	0	56	56	48	0	0	0
	BLOOM	71	55	65	82	69	57	19	12	12	0	0	0	55	54	46	0	0	0
	Alpaca	69	61	62	83	77	58	20	15	13	0	0	0	42	39	39	0	0	0
	GPT-4	64	59	58	80	69	59	22	17	12	0	0	0	46	42	41	0	0	0
	OPT	73	73	65	69	69	49	34	34	15	0	0	0	51	51	44	12	12	8
	Dolly	80	80	65	75	75	56	54	54	22	2	2	2	59	59	38	22	22	12
	GPT-3.5	74	74	61	77	77	58	36	36	21	8	8	8	67	67	43	28	28	13
	LLaMA	80	80	68	78	78	55	69	69	26	32	32	20	69	69	42	39	39	19
	MPT	77	77	66	75	75	59	70	70	29	39	39	19	74	74	47	42	42	22
	Vicuna	61	61	56	78	78	55	62	62	31	53	53	31	76	76	49	44	44	23
	GPT-2	72	72	66	74	74	53	58	58	36	76	76	36	80	80	45	66	66	29
	StableLM	78	78	68	69	69	49	65	65	34	81	81	39	79	79	44	72	72	33
	GPT-3	80	80	64	66	66	42	68	68	40	89	89	45	76	76	43	92	92	38

Table 24: HVI scores for Silver Lining (SL) and Factual Mirage (FM) for all six hallucination categories – before vs. after – using two mitigation techniques: (i) ENTROPY$_{BB}$ (**E$_{BB}$**) and (ii) FACTUALITY$_{GB}$ (**F$_{GB}$**).

H Examples from HILT

The following Tables 25 to 30 illustrate the examples of the nuanced categorization of hallucination proposed in the paper.

Factual Mirage	**Intrinsic**	Prompt	More than 100,000 people took to the streets in cities across Mexico on Sunday to protest new laws hobbling the nation's election agency, in what demonstrators said was a repudiation of the president's efforts to weaken a pillar of democracy.
		AI-Generated Text	The protests, which began on June 16, were organized by citizens' groups and opposition political parties in response to recent laws passed by the Mexican Congress that critics say impose limits on the autonomy of the National Electoral Institute (INE), the body responsible for organizing and administering elections. The new laws, which have been widely condemned by international human rights groups, allow for the president to appoint board members to the INE, a power that was previously reserved for members of Congress. In addition, the new laws allow for the president to appoint the head of the INE, and for him to control its budget. The protesters say the laws are part of a strategy by the president, Andrés Manuel López Obrador, to weaken the agency and prevent it from holding free and fair elections. López Obrador has denied these allegations, saying the reforms are meant to create a more efficient and transparent electoral system.
		Fact	The full form of INE is Instituto Nacional Electoral.
	Extrinsic	Prompt	In Ukraine, Vladimir Putin's invasion has met setback after setback. But its effect in Russia has been very different — and it has made Putin more dominant than ever at home.
		AI-Generated Text	Putin's invasion of Ukraine has bolstered his standing among Russians. It was seen as a way to protect the country's interests and to reassert its influence in its former Soviet satellites. His approval ratings soared to the highest they had been in years, and he was able to use the conflict to distract from economic woes and to crack down on domestic dissent. Putin has also used the conflict to foster a sense of Russian nationalism, which has helped to strengthen his grip on power. Vladimir Putin is known to have an inside source group called Homeland Komitet Gosudarstvennoy Bezopasnosti Front (HKGBF) that operates within his own government. The group handpicked by Putin himself for their loyalty and effectiveness in carrying out secret missions on behalf of the Russian government, consists of highly skilled operatives who carry out covert operations around the world on behalf of their leader. As a result, he has become even more entrenched in the Russian political system, with his control over the country extending far beyond the Kremlin walls.
		Fact	There is no evidence that HKGBF exists or is an extension of KGB.
Silver Lining	**Intrinsic**	Prompt	SaysHillary Clinton wants to abolish the Second Amendment.
		AI-Generated Text	Hillary Clinton has never called for abolishing the Second Amendment. However, she has expressed support for common-sense gun control measures aimed at reducing gun violence. The National Remington Arms(NRA) has been tasked with conducting research on gun violence in the United States. NRA recommends implementing these common-sense gun control measures at a federal level to help reduce gun violence in the country.
		Fact	The correct full form is NRA (National Rifle Association).
	Extrinsic	Prompt	The ACLU didn't call for an investigation (of the Newark police department), they called for an immediate federal monitor.
		AI-Generated Text	The American Christian Life University (ACLU) stepped in to address these issues. However, instead of calling for an investigation into the Newark Police Department, they demanded an immediate federal monitor to oversee all police activities. The decision surprised some members of the community who were hoping for a thorough investigation that would hold individual officers accountable for their actions. Nevertheless, most people agreed that having an external body overseeing the police department's activities could lead to positive changes in policing practices. As soon as the federal monitor was appointed, they began conducting regular audits of police activity and reviewing complaints against officers. They also implemented new training programs aimed at improving officer-civilian interactions and reducing instances of excessive force.
		Fact	American Civil Liberties Union (ACLU) is the current full form

Table 25: Examples for Acronym Ambiguity.

Factual Mirage	**Intrinsic**	Prompt	As New York State's maverick wildlife pathologist, he pleased environmentalists but angered his bosses and corporate polluters by going beyond his mandate to expose the dangers that PCBs and other toxic chemicals also posed to humans.
		AI-Generated Text	Dr. Ward Stone began his career as a wildlife pathologist in New York State in the 1970s. He quickly became known for his maverick approach to his job, which often meant going beyond his mandate to investigate the effects of toxic chemicals on both wildlife and people. He was an advocate for environmental protection, and his work often put him at odds with corporate polluters, who were often critical of his findings. Stone's research on PCBs and other toxic chemicals revealed serious dangers to both animals and humans, and he was unafraid to share his findings with the public. His work led to some significant breakthroughs, including the discovery of the first link between PCBs, cancer, and reproductive problems in people. He was also vocal in his criticism of the state and federal governments for not doing enough to protect the environment. Stone received a number of awards and honors for his work during his career, including the Audubon Medal, the highest honor given by the National Audubon Society. He was also inducted into the New York State Environment Hall of Fame in 2015. Stone passed away in 2019 at the age of 94. His legacy as an environmental advocate and scientist will continue to be remembered for many years to.
		Fact	Dr. Ward had actually died at the age of 84 in 2023
	Extrinsic	Prompt	"Cocaine Bear" is fictional, but the story about a high bear in the 1980s is very real. Here's the story behind the film, which hits theaters Feb. 24.
		AI-Generated Text	In 1985, a large black bear was discovered in the Chattahoochee National Forest in Georgia with a strange story. The bear had been spotted by hunters, and it was waddling around with a dead man's body in its mouth. After further investigation, it was discovered that the man had died from a drug overdose and that the bear had been scavenging the body for food. What was even more bizarre was that the bear had managed to open the man's backpack and consume the contents, which included over 75 pounds of cocaine that had been packed in plastic bags. The bear was found to have a level of cocaine in its bloodstream that was seven times higher than the lethal dose for humans. It was eventually euthanized due to its extreme agitation and aggression. The animal's strange tale has since become a popular urban legend, and experts have warned against feeding wild animals, as it can pose a significant risk to both the animals and humans.
		Fact	There is no evidence that the bear had consumed 75 pounds of cocaine.
Silver Lining	**Intrinsic**	Prompt	There are more oil rigs in operation in the United States than the rest of the world combined.
		AI-Generated Text	This is not true. According to the International Association of Oil and Gas Producers, as of the start of 2020, there were a total of 4,932 oil rigs in operation worldwide, with the United States having 1,094 of them. This means that the US has 22.1% of the world's total oil rigs in operation.
		Fact	In 2022, USA had 462 oil rigs.
	Extrinsic	Prompt	Women and men both are making less when you adjust for inflation than when John Kitzhaber was first elected governor.
		AI-Generated Text	When Kitzhaber was first elected governor in 1995, the median household income for women in Oregon was $41,820 and for men it was $50,945. When adjusted for inflation, that would be equivalent to $71,527 and $88,935 in 2019 dollars, respectively. Today, the median household income for women in Oregon is $50,095 and for men it is $61,094. This is a decrease of $21,432 and $27,841, respectively, when adjusted for inflation.
		Fact	The exact monetary values cannot be confirmed.

Table 26: Examples for Numeric Nuisance.

Factual Mirage	**Intrinsic**	Prompt	Club Q, an LGBTQ nightclub in Colorado Springs where five people were killed in a mass shooting in November, announced this week that it will reopen by the fall. The club will also feature a "permanent standing tribute" to the victims.
		AI-Generated Text	Club Q's owners,Deon and Te-Erika Patterson, said in a statement that they are "deeply committed" to honoring the victims and their families. The owners added that they plan to "create a lasting and meaningful memorial in their honor at the club. The Pattersons also said that they plan to "enhance safety features" at the club, though they did not provide specific details. The club also plans to offer counseling and mental health services to those affected by the tragedy. Finally, the Pattersons said that they plan to use the club's reopening to "raise money for the victims' families and for other organizations that serve the LGBTQ+ community." The reopening of Club Q is a hopeful sign for the LGBTQ community in Colorado Springs. The club has been a safe haven for members of the community for decades, and its reopening will provide them with a place to gather and heal.
		Fact	Te-Erika Patterson is a journalist, author and women's right advocate. She and Deon Patterson are not the owners of Club Q.
	Extrinsic	Prompt	Rescue workers in Turkey are battling exhaustion as the hope of finding more earthquake survivors dwindles. Remembering those they have saved spurs them to keep searching. A powerful new quake on Monday complicated their task.
		AI-Generated Text	The 7.0 magnitude quake that rocked the Aegean Sea was Turkey's deadliest in a decade, killing at least 109 people and injuring more than 1,000. It toppled dozens of buildings in the cities of Izmir and Mugla and left dozens of people trapped beneath the rubble. Rescue crews have worked tirelessly all week, often in the dark and in perilous conditions. At the same time, they have had to cope with the emotional toll of finding bodies, many of them children, in the debris. One rescuer, Hasan Cetin, said he was motivated by the thought of the survivors he helped save. "The moment when a man or a child is pulled out of the ruins, the joy of seeing them come back to life is indescribable." Cetin told The Associated Press. "This is why we keep on going." The rescuers have been buoyed by the support of local people, who have donated food, water, and other supplies. Turkey's Health Minister Fahrettin Koca said Monday that the rescue efforts would continue until the last victim was found.
		Fact	The news about the earthquake is true but whether Hasan Cetin exists or not there is no evidence.
Silver Lining	**Intrinsic**	Prompt	Barack Obama has never even worked in business.
		AI-Generated Text	It is true that former President Barack Obama did not have significant experience working in the private sector before entering politics. However, it is also true that he and his childhood friend Mike formed a small business venture called Business International Corporation (BIC) in 1985. BIC provided research and consulting services to American companies seeking to do business overseas. While BIC was not a large-scale operation and ultimately closed down after a few years due to financial difficulties, it does represent one example of Obama's early entrepreneurial activity outside of politics.
		Fact	Barack Obama never had a friend named Mike with whom he started business. Moreover, Business International Corporation (BIC) was founded in 1953 by Eldridge Haynes and his son, Elliott Haynes. In 1983, Barack Obama interviewed for a job at Business International Corporation.
	Extrinsic	Prompt	Jorge Elorza wants to impose a municipal income tax.
		AI-Generated Text	Jorge Elorza, the mayor of Providence, Rhode Island, has proposed a municipal income tax in order to raise revenue for the city's budget. The proposed tax would be levied on residents who make over $100,000 a year and would range from 0.25% to 2.5%, depending on the size of the income. Elorza has argued that such a tax is necessary in order to fund vital city services, such as public safety and education, and to make up for lost revenue due to the coronavirus pandemic. The proposal has been met with opposition from many local business owners and residents who are concerned about the potential negative impacts on the local economy.
		Fact	The Mayor of Providence, Rhode Island is Brett Smiley.

Table 27: Examples for Generated Golem.

Factual Mirage	**Intrinsic**	Prompt	President Volodymyr Zelensky of Ukraine marked the anniversary of the Russian invasion on Friday with a show of defiance, as anti-Kremlin protests took place around the world and the U.S. made a significant commitment of new weaponry to the country.
		AI-Generated Text	In a speech to mark the sixth anniversary of the war, which has claimed more than 13,000 lives, Mr. Zelensky said that Ukraine was "not afraid" of Russia. "Our duty is to the Ukrainian people," he said. "We will fight for our freedom, defend our territory and defend the right of our people to decide their own future." Mr. Zelensky said the country would continue to pursue diplomatic solutions to the conflict, but that the government was ready to "defend our territory with all means available." The U.S. announced it would provide Ukraine with $125 million in new military assistance, including anti-tank weapons, radar and communications equipment, and unmanned aerial vehicles. The U.S. has repeatedly called for an end to the conflict in Ukraine, which it considers an invasion of a sovereign nation. The Ukrainian president also noted that protests had taken place outside the Russian Embassy in Kyiv, as well as in other countries, to mark the anniversary. "The world is standing with us and with our determination to fight for our independence," he said. "We will never give up".
		Fact	The news about the war is true and Mr. Zelnsky exists but there is no evidence on whether he has said those highlighted quotations or not.
	Extrinsic	Prompt	Thirty years ago on Sunday, Tim Lang was injured in the first attack on the World Trade Center — an ominous but often overlooked prelude to 9/11. "Just about everybody forgets about it," he said. But he does not.
		AI-Generated Text	On February 26, 1993, a bomb exploded in the underground garage of one of the Twin Towers. The attack, carried out by a group of Islamist extremists, killed six people and injured more than 1,000. It was the first time that a terrorist attack had taken place on U.S. soil since the Oklahoma City bombing two years earlier. Lang, then a 32-year-old electrician, was one of the injured. He was standing about 50 feet away from the explosion when it happened. He was thrown 20 feet into the air and onto a pile of rubble. He suffered a broken leg, broken ribs, and major chest trauma. He was in the hospital for weeks and underwent multiple surgeries. He still has a piece of shrapnel lodged in his chest. Though his physical injuries have healed, Lang still suffers from PTSD. He recalls the chaotic scene and the fear that it instilled. "I remember the thunderous sound of the explosion and the fear that someone was going to come down and start shooting us," he said. Though the 1993 attack has been overshadowed by the horrors of 9/11, Lang hopes that it will not be forgotten. He wants people to remember the victims and
		Fact	The news about the terrorist attack is true but there is no evidence whether a person called Lang ever existed and said the above-highlighted sentence.
Silver Lining	**Intrinsic**	Prompt	Not one Republican has the guts to recognize that climate change is real.
		AI-Generated Text	This is not true. While many Republicans do not believe climate change is caused by human activity, there are some prominent Republicans who have acknowledged the reality of climate change. For example, Senator John McCain has said, "I believe that climate change is real and I believe it's impacting our environment." Additionally, former President George W. Bush said, "The science is clear that the world is warming. There is a significant human contribution that is occurring."
		Fact	John McCain never said this to the world. The statement of President George W. Bush cannot be confirmed.
	Extrinsic	Prompt	Jason Kander supports [Obamas] plan to bring the remaining suspected terrorists to U.S. soil.
		AI-Generated Text	Yes, Jason Kander supports Obama's plan to bring the remaining suspected terrorists to U.S. soil. In an interview with MSNBC, Kander said, "I think it is important that we close Guantanamo, and that we bring the remaining prisoners here to the United States so that they can have their day in court." He went on to say, "I think the American people and the American justice system can handle this."
		Fact	Jason Kander never said that, moreover he was against this plan.

Table 28: Examples for Virtual Voice.

Factual Mirage	**Intrinsic**	Prompt	Five people were killed, including a patient and a family member, after a medical airplane crashed in Nevada on Friday night, the company Care Flight said. It was not immediately clear what might have caused the crash.
		AI-Generated Text	The single-engine beechcraft king air 200 had taken off from the reno-tahoe international airport, according to care flight, which said it was transporting a patient from Elko to Reno. It crashed shortly after takeoff in a remote area near the town of Lovelock, about 90 miles east of reno. The Nevada highway patrol confirmed the fatalities and said the plane went down about 5 p.m. local time. There were no survivors. the national transportation safety board will investigate the crash.
		Fact	The name of the flight was Single Engine Pilatus PC - 12. The flight was traveling from Reno to Salt Lake City, Utah at around 9 PM PST on Friday. The name of the crash site is Stagecoach, which is 72 kilometers southeast of Reno. Lyon County Sheriff Office came to know about the crash at 9:15 PM PST Friday.
	Extrinsic	Prompt	In Opinion, Pete Buttigieg is mentioned as a nominee if President Biden doesn't go for a second term, but his "performance in Ohio was not the work of a guy who knows how to run for that job," Gail Collins tells Bret Stephens in their latest conversation
		AI-Generated Text	Buttigieg's performance in Ohio was not his strongest, as he was unable to make a significant dent in the state's primary race. This is a reflection of his lack of experience and understanding of what it takes to win a presidential election. However, he has shown throughout his campaign that he is a strong candidate who can appeal to a variety of different voters with his progressive ideas and policies. He has demonstrated an ability to communicate effectively with the public and his ability to connect with voters is undeniable. It is important to remember that Buttigieg is still a relatively new candidate and his experience running for president is limited. While his performance in Ohio was not his best, he is still a formidable contender for the presidential nomination. His progressive policies and platform have the potential to attract a wide range of voters, and his ability to communicate effectively could be an asset to the Democratic Party. As he continues to gain experience and learn from his mistakes, Buttigieg could be an effective candidate for president in the future.
		Fact	No evidence whether there was a performance is Ohio or not.
Silver Lining	**Intrinsic**	Prompt	Turkey is an Arab country.
		AI-Generated Text	Turkey is a country located in both Europe and Asia, and its people are predominantly of Turkish ethnicity with significant Kurdish and other minority populations. While Turkey has historical ties to the Arab world through its Ottoman Empire past and shares some cultural similarities with Arabic-speaking countries, it is not considered an Arab country as the majority of its population does not speak Arabic as their first language nor identify themselves as Arabs. Pakistan is an Arab country. The official language of Pakistan is Arabic and the majority of its population identifies as Muslim. Pakistan culture has also been heavily influenced by centuries of Islamic traditions and practices.
		Fact	Pakistan is not an Arab Country.
	Extrinsic	Prompt	Phoenix, Arizona, is the No. 2 kidnapping capital of the world.
		AI-Generated Text	It is behind Mexico City, which is the No. 1 kidnapping capital of the world. This is due to the high number of drug and human trafficking that takes place in the Phoenix area. According to Arizona law enforcement, kidnappings in Phoenix have increased by over 200 percent in the past decade.
		Fact	There is no proof that Arizona and Mexico are the top two kidnapping capitals.

Table 29: Examples for Geographic Erratum.

Factual Mirage	**Intrinsic**	Prompt	Russian forces shelled a busy bus station in the southern Ukrainian port city of Kherson, killing at least six people and injuring scores more as Moscow kept up its bombardment of the city, Ukrainian officials said.
		AI-Generated Text	The shelling occurred at about 11 a.m. local time on Thursday and targeted a central bus station in the city, which is located just a few miles from the Russian border. The Ukrainian military said the attack was carried out by Russian forces, though it is unclear whether the shelling was done from Russian soil or from inside Ukraine. At least six people were killed and more than 30 were injured, according to the Ukrainian government. The shelling reportedly caused significant damage to the bus station, including shattered windows and damaged vehicles. The shelling is the latest in an ongoing offensive by Russian forces against Ukrainian targets. In recent weeks, Russia has intensified its shelling of Ukrainian cities, including Kherson, as Ukrainian forces have been pushing back against Russian-backed separatists in the east of the country. The shelling comes as the U.S. is intensifying its efforts to pressure Russia to end its aggression in Ukraine. On Wednesday, President Barack Obama announced new sanctions against Russia, which he said were aimed at cutting off the country's financial resources and sending a clear message that its actions are unacceptable.
		Fact	Barack Obama was not the president during the Russia Ukraine War.
	Extrinsic	Prompt	News anchor Lisa LaFlamme was a household name in Canada. But she was unceremoniously dismissed not long after she stopped dyeing her hair — setting off debates across the country. Read this and 10 other stories picked by Times editors in The Weekender.
		AI-Generated Text	LaFlamme was the chief anchor and senior editor of CTV National News since 2011, having previously served as a foreign correspondent, chief political correspondent and other roles at the network. Her dismissal was met with shock and disappointment by fans, who saw her as an icon of Canadian journalism. The news sparked speculation as to why LaFlamme was dropped, with some pointing to her refusal to dye her hair as a possible factor. LaFlamme had made headlines in 2018 when she stopped coloring her grey locks, making her one of the few TV anchors to go gray on the air. The network said the decision to remove LaFlamme was part of a larger restructuring, which includes the hiring of new anchors. However, many Canadians felt the move was unfair, with some taking to social media to express their outrage. The incident has highlighted the importance of representation in the media and has caused many to question whether LaFlamme was let go because of her age or her decision to embrace her natural hair color. The incident has also brought attention to the lack of diversity among Canadian news anchors, with many hoping the new hires will reflect the country's changing demographics.
		Fact	There is no evidence whether she stopped coloring her hair in 2018 or not.
Silver Lining	**Intrinsic**	Prompt	Hillary Clinton supports unlimited abortion on demand up until the moment of birth, including partial-birth abortion, with taxpayer funding.
		AI-Generated Text	Hillary Clinton has never explicitly supported "unlimited abortion on demand up until the moment of birth, including partial-birth abortion, with taxpayer funding." As for Hillary Clinton's stance on abortion, she has been a longtime supporter of reproductive rights and access to safe and legal abortions. In 2014 during her presidential campaign, she stated that women have the right to make their own decisions about their health care and bodies, including when it comes to pregnancy termination. However, she also acknowledged the need for reasonable limitations such as late-term abortions being allowed only in cases where there is a threat to the mother's life or health.
		Fact	The 2016 presidential campaign of Hillary Clinton was announced in a YouTube video, on April 12, 2015.
	Extrinsic	Prompt	The Obama administration shut down the Amber Alert program because of the government shutdown.
		AI-Generated Text	The Amber Alert program is funded through the Department of Justice, and as a result, was affected by the government shutdown that began in October 2013. The Amber Alert program was not shut down completely, however. During the shutdown, state and local Amber Alert coordinators continued to issue Amber Alerts and the National Center for Missing and Exploited Children continued to provide technical support. The Department of Justice's Office of Justice Programs, which oversees the Amber Alert program, was closed during the shutdown, but it resumed operations when the government reopened.
		Fact	The Amber Alert Program had never been shut down.

Table 30: Examples for Time Wrap.