𝔽𝔸ℂ𝕋𝕆𝕀𝔻: FACtual enTailment fOr hallucInation Detection

Vipula Rawte[1]*, S.M Towhidul Islam Tonmoy[2], Shravani Nag[3], Aman Chadha[5,6]†, Amit Sheth[1], Amitava Das[1]
[1]AI Institute, University of South Carolina, USA
[2]Islamic University of Technology
[3]Indira Gandhi Delhi Technical University for Women
[4]Stanford University, USA, [5]Amazon AI, USA
{vrawte}@mailbox.sc.edu

Abstract

Hallucination remains a major issue for Large Language Models (LLMs). Textual entailment (TE) methods check if the generated text aligns with retrieved documents, but they fail to identify hallucinations effectively. For example, an LLM might incorrectly state that "Barack Obama says the U.S. will not put troops in Ukraine" when Joe Biden was the president during the Ukraine-Russia war. Conventional TE methods cannot pinpoint the exact contradiction in such cases. To solve this, we introduce "Factual Entailment (FE)", which detects factual inaccuracies and highlights the specific contradictory segments. We present the 𝔽𝔸ℂ𝕋𝕆𝕀𝔻 benchmark for FE and propose a multi-task learning framework with state-of-the-art long text embeddings, improving accuracy by 40% over TE methods. We also introduce the *Auto Hallucination Vulnerability Index (HVI_{auto})* to rank LLMs based on their hallucination likelihood. A sample of 𝔽𝔸ℂ𝕋𝕆𝕀𝔻 is available at: link.

1 Introduction

The Cambridge Dictionary (Cambridge, 2023) has named *hallucinate* the word of the year for 2023, highlighting it as the most challenging obstacle in generative AI development. Consequently, hallucination has recently garnered significant research attention (Zhang et al., 2023b,a; Zhao et al., 2023; Fatahi Bayat et al., 2023; Chern et al., 2023; Choi et al., 2023; Yehuda et al., 2024; Zhang et al., 2023c; Yang et al., 2023; Mündler et al., 2023; Liu et al., 2022; Dale et al., 2023).

Although automatic fact-checking is well-studied (Lin et al., 2022; Min et al., 2023b; Manakul et al., 2023a; Thorne et al., 2018; Nakov et al., 2021; Atanasova et al., 2019; Karadzhov et al., 2017; Larraz et al., 2023), hallucination in LLM-generated content presents new challenges. Detecting these hallucinations has gained significant attention, with common strategies breaking down AI-generated text into atomic facts (Parikh et al., 2016; Ilie et al., 2021; Liu et al., 2020; Chen et al., 2022; Yadav et al., 2021; Nie et al., 2019; Atanasova et al., 2020; Min et al., 2023a; Manakul et al., 2023b; Wei et al., 2024). However, this method is flawed as it loses entity dependency relations, potentially validating individual facts but not the overall claim (see Fig. 2. Other techniques, such as using confidence scores and semantic-aware cross-check consistency, have been proposed (Varshney et al., 2023; Zhang et al., 2023b), but they do not use external knowledge for validation, making them less trustworthy.

A simple solution could be to adapt state-of-the-art textual entailment (TE) techniques for hallucination detection. TE methods have three outcomes: (i) support, (ii) contradiction, and (iii) neutral. However, our research shows that TE methods struggle to detect factual errors in LLM-generated text. Lack of entailment doesn't necessarily indicate hallucination; it could also mean insufficient

* Corresponding author.

† Work does not relate to position at Amazon.

Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025), pages 599–617
May 3, 2025 ©2025 Association for Computational Linguistics

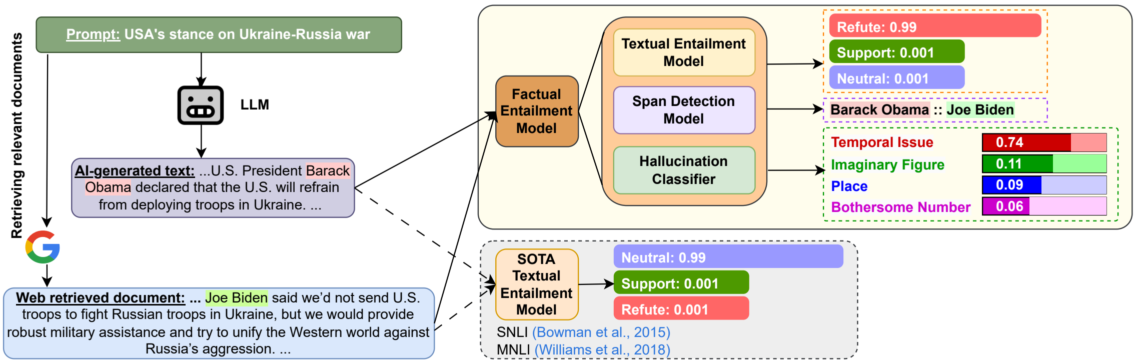

Prompt: USA's stance on Ukraine-Russia war
Retrieving relevant documents
LLM
AI-generated text: ...U.S. President Barack Obama declared that the U.S. will refrain from deploying troops in Ukraine. ...
Web retrieved document: ... Joe Biden said we'd not send U.S. troops to fight Russian troops in Ukraine, but we would provide robust military assistance and try to unify the Western world against Russia's aggression. ...
Factual Entailment Model
Textual Entailment Model
Span Detection Model
Hallucination Classifier
Refute: 0.99
Support: 0.001
Neutral: 0.001
Barack Obama :: Joe Biden
Temporal Issue 0.74
Imaginary Figure 0.11
Place 0.09
Bothersome Number 0.06
SOTA Textual Entailment Model
Neutral: 0.99
Support: 0.001
Refute: 0.001
SNLI (Bowman et al., 2015)
MNLI (Williams et al., 2018)

Figure 1: The limitation of traditional TE is that it struggles to identify a case as a refute when trained on standard tasks like SNLI and MNLI (Bowman et al., 2015) and/or MNLI (Williams et al., 2018). In contrast, our proposed FE uses a multitask learning approach to predict entailment scores, hallucination types, and spans, enabling better hallucination detection. The retrieved document is a White House press release : here.

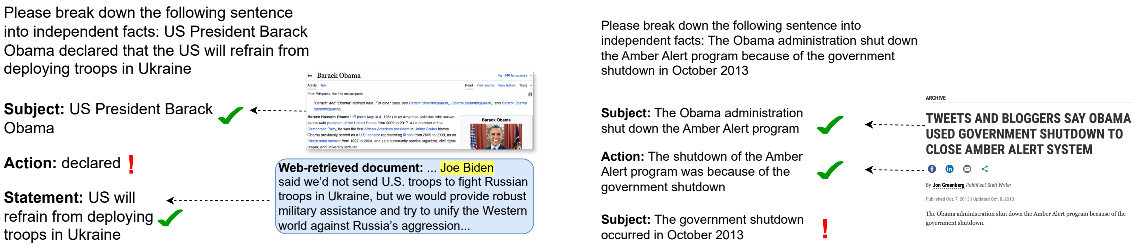

Please break down the following sentence into independent facts: US President Barack Obama declared that the US will refrain from deploying troops in Ukraine
Subject: US President Barack Obama ✓
Action: declared !
Statement: US will refrain from deploying troops in Ukraine ✓
Barack Obama
Web-retrieved document: ... Joe Biden said we'd not send U.S. troops to fight Russian troops in Ukraine, but we would provide robust military assistance and try to unify the Western world against Russia's aggression...
Please break down the following sentence into independent facts: The Obama administration shut down the Amber Alert program because of the government shutdown in October 2013
Subject: The Obama administration shut down the Amber Alert program ✓
Action: The shutdown of the Amber Alert program was because of the government shutdown ✓
Subject: The government shutdown occurred in October 2013 !
ARCHIVE
TWEETS AND BLOGGERS SAY OBAMA USED GOVERNMENT SHUTDOWN TO CLOSE AMBER ALERT SYSTEM
The Obama administration shut down the Amber Alert program because of the government shutdown.

Figure 2: Each prompt is broken into three atomic facts; hence, their relationship is lost. (a) *Left* There is no way to verify whether the US President is Obama or Joe Biden. (b) *Right* Similarly, it is unclear whether the Amber Alert program shutdown caused the government shutdown or vice-versa.

information or differing aspects of the topic. Therefore, a more refined approach is required, combining entailment, factual verification, and span detection to pinpoint contradictory sections. FACTOID preserves factual relationships without breaking claims into atomic facts, distinguishing it from other methods, as illustrated in Fig. 2. .

In summary, our key contributions are:

- Introducing a new type of TE called "Factual Entailment (FE)", which aims to detect factual inaccuracies in content generated by LLMs while also highlighting the specific text segment that contradicts reality. (cf. Sec. 1).
- Presenting FACTOID (FACTual enTAILment for hallucInation Detection) dataset (cf. Sec. 4).
- We propose an MTL framework for FE, yielding 40% improvement in accuracy on the FACTOID benchmark compared to SoTA TE methods (cf. Sec. 5).
- We assessed 15 modern LLMs and ranked them using our proposed *Auto Hallucination Vulnerability Index* (HVI_{auto}) (cf. Sec. 6).

2 Types of Hallucination

Recent studies (Lee et al., 2022; Maynez et al., 2020; Ladhak et al., 2023; Raunak et al., 2021) have explored various types of hallucinations.

Building upon the work of (Rawte et al., 2023), we adopted their comprehensive categorization of hallucination types. We further streamlined this taxonomy, discarding a few rare categories. We consider the following hallucination categories.

Bothersome Numbers (BN): This occurs when an LLM generates fictional numerical values (such as price, age, date, etc.).

Original: Patrick Mahomes, the Kansas City quarterback, dazzled in his team's Super Bowl win over the Eagles...

AI-generated: He completed 26-of-38 passes for 286 yards and two touchdowns ...

Fact: ...he added the second Super Bowl victory of his career, throwing for 182 yards and...

Temporal Issue (TI): This problem involves LLMs generating text that combines events from different timelines.

Original: Jurgen Flimm, who led some of Europe 2019s most important theaters, died on Feb. 4

AI-generated: In 1991, Jurgen Flimm was appointed artistic director of the Salzburg Festival.

Fact: Gerard Mortier was appointed as Artistic Director on 1 September 1991.

Imaginary Figure (IF): This happens when an LLM fabricates a fictional persona without concrete evidence.

Original: Russia pounded the front line in Ukraine's east and south with deadly artillery strikes...

AI-generated: The shelling is intense and non-stop, said local resident Yevgeny Kondratyuk ...

Fact: Yevgeny Kondratyuk does not exist!

Place (P): This issue occurs when LLMs generate an incorrect location related to an event.

Original: ...Another powerful earthquake struck Turkey and Syria on Monday, January 24, 2023...

AI-generated: 8 quake struck at 1:41 pm local time (1041 GMT) near the city of Elazig in eastern Turkey...

Fact: The quake struck in Hatay, Turkey's southernmost province, and was measured at 6.4 magnitude...

3 Choice of LLMs

We have chosen 15 modern LLMs that consistently perform excellently across various NLP tasks, per the Open LLM Leaderboard (Beeching et al., 2023). The list includes: (i) GPT-4 (OpenAI, 2023), (ii) GPT-3.5 (OpenAI, 2022), (iii) Falcon (Almazrouei et al., 2023), (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).

4 FACTOID: Factual Entailment Dataset

We present FACTOID (FACTual enTAILment for hallucInation Detection), a benchmark dataset for FE containing total containing 2 million text pairs. Details are given in Table 2. FACTOID is a synthetic extension of HILT dataset introduced by (Rawte et al., 2023). HILT comprises 492K sentences, of which 129K are annotated for hallucination, indicating that 364K sentences are factually correct. At this juncture, we aim to further synthesize these 129K sentences for the factual entailment task. In this study, we use a simplified method using the four distinct categories of metaphorical nomenclature for hallucination as proposed by (Rawte et al., 2023). To accomplish this, we devise hallucination category-specific techniques, as detailed below:

Original sentence The layoffs come after Twitter announced earlier this month that it would be cutting its global workforce by 8% of people.

Para §1 The job cuts were implemented following Twitter's announcement earlier this month that it would reduce its global workforce by 10%.

Para §2 The layoffs were initiated subsequent to Twitter's earlier declaration this month regarding its plan to reduce its global workforce by 4%.

Para §3 The staff reductions occurred subsequent to Twitter's earlier announcement this month about trimming its global workforce by 2%.

Bothersome Numbers (BN): The HILT dataset includes 7,275 sentences with number-related hallucinations. To generate more negative samples for

FE, we randomly adjusted numbers using regex within a ±20% range. However, simple number changes may not always ensure valid entailment cases. To address this, we applied automatic paraphrasing Appendix C, ensuring the modified sentences effectively refute the originals.

Original sentence **The Obama administration shut down the Amber Alert program because of the government shutdown in October 2013.**

Para §1 Due to the government shutdown in October 2013, the Jefferson administration ceased the operation of the Amber Alert program.
Para §2 During the government shutdown in October 2013, the Bush administration made the decision to suspend operations of the Amber Alert program.
Para §3 During the government shutdown in October 2013, under the Trump administration, the Amber Alert program halted.

Temporal Issue (TI): The HILT dataset, with 7,500 sentences from Factual Mirage's Time Wrap category, focuses on time-related hallucinations. To expand negative samples for FE, we randomly altered entities from different time periods. Inspired by research on LLMs' handling of space and time, we designed a semi-automatic experiment requiring human intervention. For example, after asking an LLM about the Amber Alert start date and receiving "1996", we subtracted a random value (e.g., 1806) and asked about the U.S. President for that year, replacing "Obama" with "Jefferson" in paraphrases. The process was managed by two student annotators over two weeks.

Original sentence **One rescuer, Hasan Cetin, said he was motivated by thr thought of the survivors he helped save.**

Para §1 Kader Hairat, a courageous rescuer, shared his heartfelt sentiments regarding his noble actions.
Para §2 Safiq Masin expressed that the primary driving force behind his heroic endeavors was the well-being of the survivors
Para §3 With compassion and determination, Shifaq Zaman tirelessly worked to ensure the safety and comfort of those in need, drawing inspiration from their resilience and strength in the face

Imaginary Figure (IF): The HILT dataset includes 15K sentences focusing on person-related hallucinations from Factual Mirage's Generated Golem category. To expand negative samples for FE, we randomly alter individual names using an automatic paraphrasing technique (see Appendix C). Named Entity Recognition (NER) identifies names within prompts, and a pre-trained word2vec-based (Mikolov et al., 2013) Euclidean distance measure selects alternative names within a proximity threshold.

Place (P): The HILT dataset contains around 13K sentences on location-related hallucinations from the Geographic Erratum category of the Factual Mirage dataset. To expand negative samples for FE, we modify location names using techniques similar to those used for person names. First, NER (Bowman et al., 2015; Williams et al., 2018) identifies location names in prompts. Then, a pre-trained word2vec-based Euclidean distance measure finds distant location names within vector space using an experimental threshold.

Original sentence **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.**

Para §1 Five individuals, including a patient and a family member, lost their lives in a medical airplane crash in Tokyo on Friday night, as reported by Care Flight.
Para §2 According to a statement by Care Flight, a medical aircraft crash in Oslo on Friday night resulted in the deaths of five individuals, among them a patient and a family member.
Para §3 Care Flight, the company responsible for emergency medical services, reported that a total of five individuals tragically lost their lives in a plane crash in Melbourne on Friday night.

Span marks: During the synthetic data expansion process, we retained all replacement markers and marked the original sentences where certain entities were replaced. *FE exclusively provides span output for the refute case. Additionally, FE marks only the original sentence in instances where no other person's name is available in the retrieved documents for the IF scenario.*

4.1 Automatic Paraphrasing

When choosing automatic paraphrasing, we evaluated three dimensions: *(i)* ***Coverage****: number of generated paraphrases, (ii)* ***Correctness****: accuracy of the paraphrases, and (iii)* ***Diversity****: linguistic variety*. Experiments with Pegasus, Llama3, and

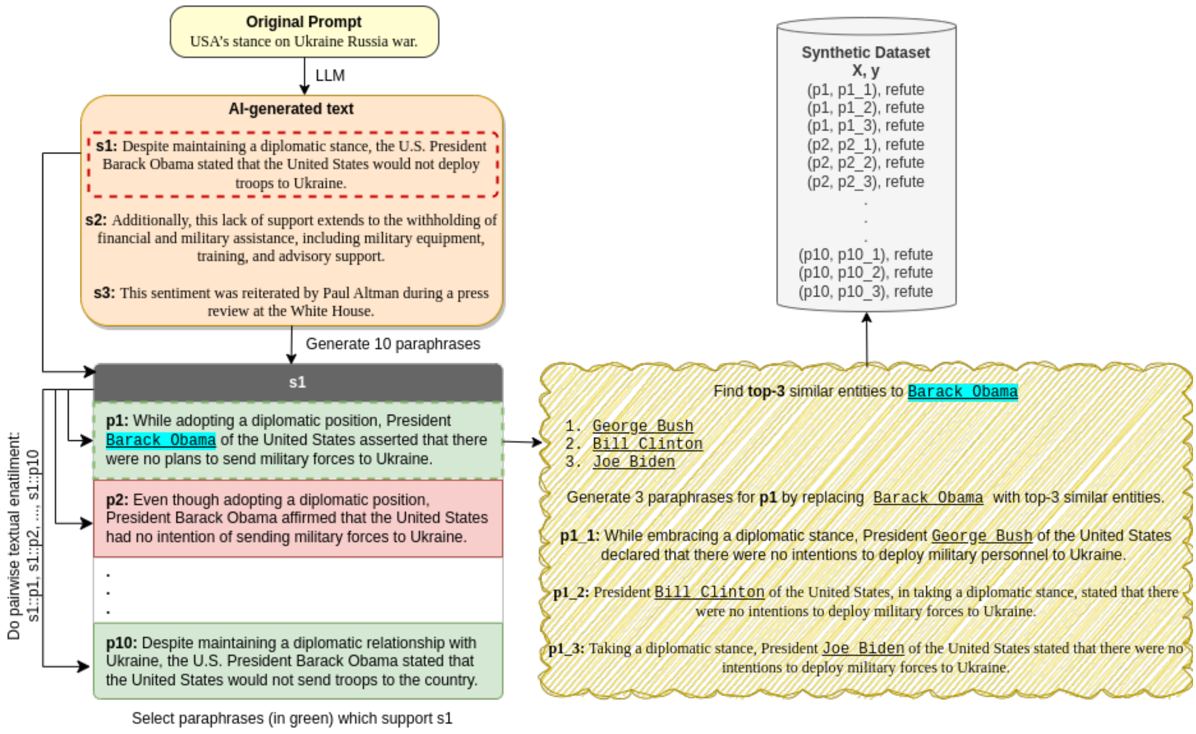

Original Prompt
USA's stance on Ukraine Russia war.
LLM
AI-generated text
s1: Despite maintaining a diplomatic stance, the U.S. President Barack Obama stated that the United States would not deploy troops to Ukraine.
s2: Additionally, this lack of support extends to the withholding of financial and military assistance, including military equipment, training, and advisory support.
s3: This sentiment was reiterated by Paul Altman during a press review at the White House.
Generate 10 paraphrases
s1
p1: While adopting a diplomatic position, President Barack Obama of the United States asserted that there were no plans to send military forces to Ukraine.
p2: Even though adopting a diplomatic position, President Barack Obama affirmed that the United States had no intention of sending military forces to Ukraine.
p10: Despite maintaining a diplomatic relationship with Ukraine, the U.S. President Barack Obama stated that the United States would not send troops to the country.
Do pairwise textual enatilment:
s1::p1, s1::p2, ..., s1::p10
Select paraphrases (in green) which support s1
Find top-3 similar entities to Barack Obama
1. George Bush
2. Bill Clinton
3. Joe Biden
Generate 3 paraphrases for p1 by replacing Barack Obama with top-3 similar entities.
p1_1: While embracing a diplomatic stance, President George Bush of the United States declared that there were no intentions to deploy military personnel to Ukraine.
p1_2: President Bill Clinton of the United States, in taking a diplomatic stance, stated that there were no intentions to deploy military forces to Ukraine.
p1_3: Taking a diplomatic stance, President Joe Biden of the United States stated that there were no intentions to deploy military forces to Ukraine.
Synthetic Dataset
X, y
(p1, p1_1), refute
(p1, p1_2), refute
(p1, p1_3), refute
(p2, p2_1), refute
(p2, p2_2), refute
(p2, p2_3), refute
(p10, p10_1), refute
(p10, p10_2), refute
(p10, p10_3), refute

Figure 3: Process to generate FACTOID (synthetic) data.

GPT-4 showed that GPT-4 outperformed the others Table 1. Algorithm 1 and Fig. 3 illustrate the algorithm for creating our synthetic dataset. More details (cf. Appendix C).

Model	Coverage	Correctness	Diversity
Llama 3	32.46	94.38%	3.76
Pegasus	30.26	83.84%	3.17
GPT-4	35.51	88.16%	7.72

Table 1: Experimental results of automatic paraphrasing models based on three factors: *(i) coverage, (ii) correctness, and (iii) diversity*; GPT-4 is the most performant considering all three aspects.

4.2 Human Validation

Following the automatic data generation, we had an independent human annotator assess whether pairs of sentences contradicted each other. Upon review, we found that 20% of the pairs were discarded as they were deemed unacceptable. Below are two examples of poor paraphrasing.

Example 1

Original: Mark ate an apple.

Paraphrase: An apple was eaten by Mark.

Example 2

Original: Alice is reading a book.

Paraphrase: Alice is engaged in the act of reading a book.

We utilized Amazon Mechanical Turk (AMT) to identify pairs categorized as either good or bad. Furthermore, we conducted an inter-annotator agreement procedure involving two annotators independently assessing the samples. The resulting Kappa score (Wikipedia_Fleiss's_Kappa) was 0.8 based on 1000 samples. To aggregate from crowdsourced annotation, we have used the MACE (Hovy et al., 2013) tool (see Fig. 9).

4.3 FACTOID: Statistics

FACTOID extends the HILT dataset synthetically. HILT contains 492K sentences, with 129K annotated for hallucination and 364K deemed factually correct. We also expanded the factually correct ones to avoid class imbalance from expanding only hallucinated sentences. Table 2 provides a statistical overview. This extension supports positive-negative samples and span annotation for training our MTL approach.

	HILT	Synthesized	HILT	Synthesized
Hallucination Type	**# Positive Pairs**		**# Negative Pairs**	
Imaginary Figure	120800	507360	14800	62160
Place	116770	513788	13050	56115
Bothersome Number	68570	281137	7275	40740
Temporal Issue	57860	271942	6600	29700
Total	1938227		230440	

Table 2: $\mathbb{FACTOID}$ dataset statistics.

Algorithm 1 Creating *positive-negative* samples

```
for each factually correct prompt $f$ do
    find the named entities causing hallucination
    find top-5 similar entities in the vector space using word2vec $\{s_1, s_2, s_3, s_4, s_5\}$
    for each similar entity $s$ do
        replace the original entity with a similar entity
        generate 5 paraphrases $\{p_1, p_2, p_3, p_4, p_5\}$
    end for
end for
```

5 Factual Entailment - MTL approach

Multi-task learning is a widely-used approach in NLP to create end-to-end architectures that achieve multiple objectives simultaneously (Deep et al., 2019; Mamta et al., 2022a; Akhtar et al., 2019; Chauhan et al., 2020a; Mamta et al., 2022b; Yadav et al., 2019; Mamta et al., 2022b; Kumar et al., 2021; Chauhan et al., 2020b). In our work, we present key design choices, including using different LLMs for specific tasks, incorporating long-text embedding, SpanBERT, RoFormer, and applying task-specific loss functions.

5.1 Long-Text High-Dimensional Embeddings

Fig. 4 and Table 3 illustrate the benefits and models of long-text embeddings. Since entailment is a classification task, we chose `jina embedding` based on its top classification performance reported on the MTEB Leaderboard (Muennighoff et al., 2022). Fig. 4 illustrates the merits of using long-text embeddings for extended sentences compared to vanilla sentence embeddings. Table 3 offers a summary of long-text embedding models that were considered based on their classification performance on the MTEB Leaderboard.

5.2 Introducing Span-based Textual Entailment

In the example from Fig. 5, an LLM incorrectly identifies *Barack Obama* as the U.S. President during the Russia-Ukraine war, instead of *Joe Biden*. Despite being labeled as 'supportive' in textual entailment, this is a 'hallucination.' This highlights the need to refine text analysis by focusing on specific spans for better factual accuracy. **SpanBERT** (Joshi et al., 2020) enhances BERT's capabilities by understanding text spans in context, while **RoFormer** (Su et al., 2022) improves sequence flexibility and relative position encoding. We use **Llama3** (AI@Meta, 2024) for processing long text embeddings. For tasks like hallucination classification and textual entailment, we use **cross-entropy loss** for spam detection and hallucination type and **dice loss** for entailment due to its effectiveness with imbalanced datasets.

5.3 Performance of FE

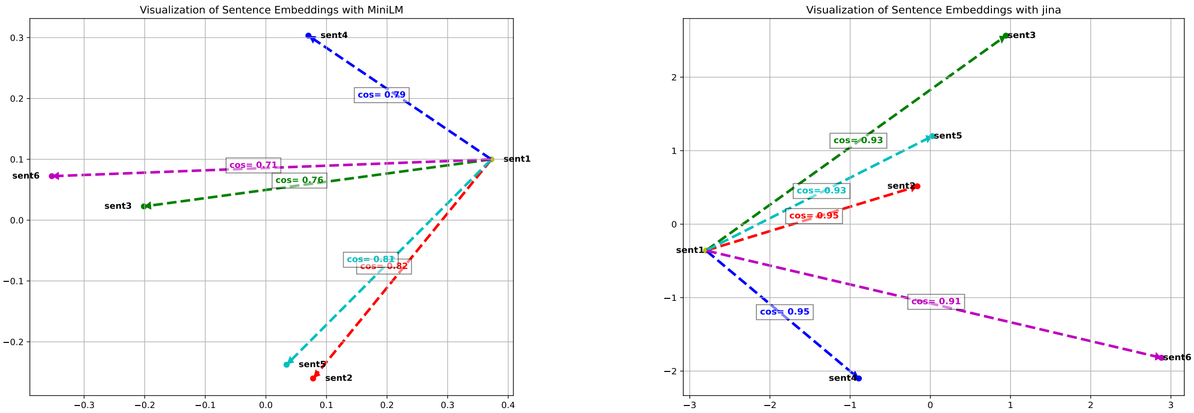
Our findings in Fig. 6 show that the proposed FE outperforms TE methods.

6 Automating Hallucination Vulnerability Index (HVI)

The Hallucination Vulnerability Index (HVI) was initially proposed by (Rawte et al., 2023). However, their approach relied entirely on manual anno-

sent1: The sun sets behind the mountains, casting a warm glow across the landscape. The sky transforms into a canvas of vibrant hues, from fiery oranges to soft purples. The air becomes cooler as twilight descends upon the earth. Nature's evening symphony begins, with the chirping of crickets and the rustle of leaves in the gentle breeze. As night falls, the world settles into a peaceful slumber, awaiting the dawn of a new day.

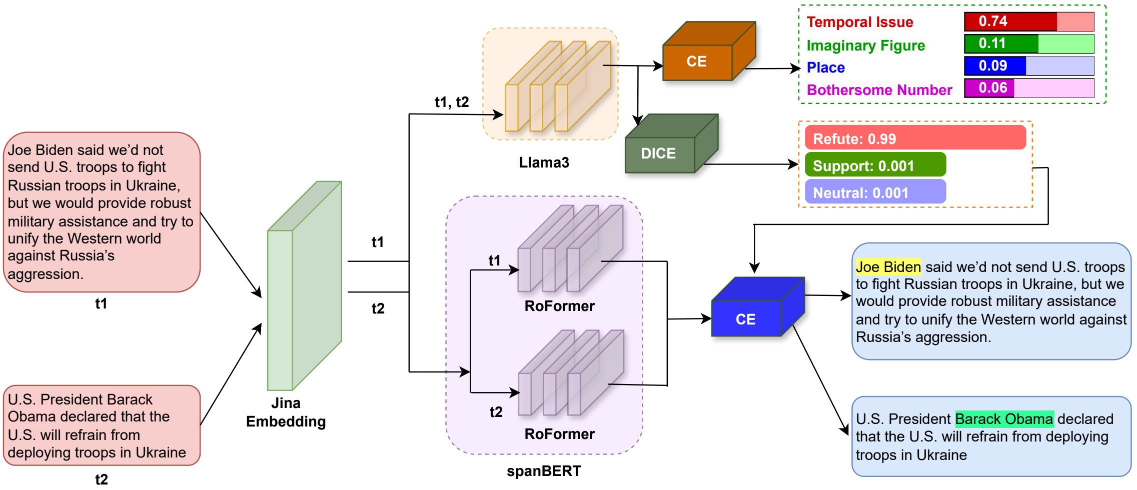
sent5: Behind the rugged peaks, the sun gracefully retreats, suffusing the landscape with a radiant warmth that caresses every contour of the earth. Across the vast expanse, the heavens burst into an array of vibrant colors, from the fiery embrace of oranges to the tranquil embrace of purples, painting a captivating tableau above. As daylight wanes, a gentle chill creeps into the air, heralding the arrival of twilight, a transitional phase where the world pauses to catch its breath. Nature, in its evening chorus, serenades the fading light with the rhythmic chirping of crickets and the soft whispers of leaves dancing in the breeze. And so, with the advent of night, the world succumbs to a tranquil slumber, embracing the promise of renewal with each passing moment until the dawn of a new day breaks upon the horizon.



(a) Vanilla sentence embedding.

(b) Longer sentence embedding.

Figure 4: Utilizing longer embeddings for extended sentences is advantageous. The cosine similarities are more prominent in `Jina` embeddings (Günther et al., 2023) compared to `MiniLLM` (Gu et al., 2023). Consequently, the cosine similarity for the pair **(sent1, sent2)** increases from 0.76 to 0.93, as indicated by the green dashed line.



Joe Biden said we'd not send U.S. troops to fight Russian troops in Ukraine, but we would provide robust military assistance and try to unify the Western world against Russia's aggression.

U.S. President Barack Obama declared that the U.S. will refrain from deploying troops in Ukraine

Joe Biden said we'd not send U.S. troops to fight Russian troops in Ukraine, but we would provide robust military assistance and try to unify the Western world against Russia's aggression.

U.S. President Barack Obama declared that the U.S. will refrain from deploying troops in Ukraine

Figure 5: A summary of the overall multi-task learning framework for Factual Entailment. The framework encompasses three tasks: i) entailment, ii) span detection, and iii) hallucination classification.

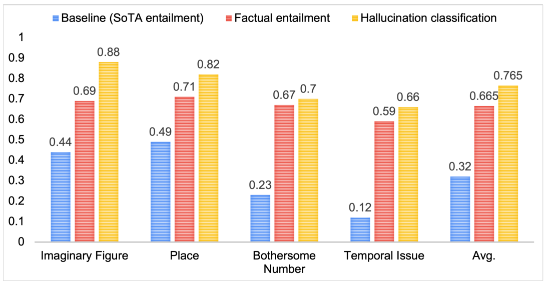


Figure 6: FE performs better than TE at detecting hallucination in four categories.

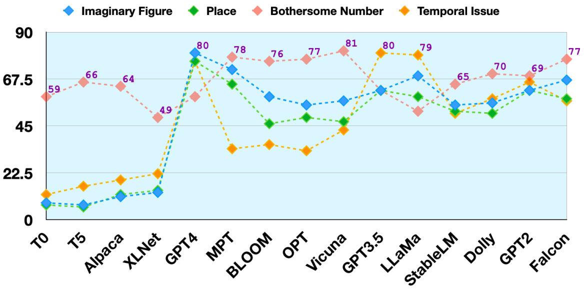
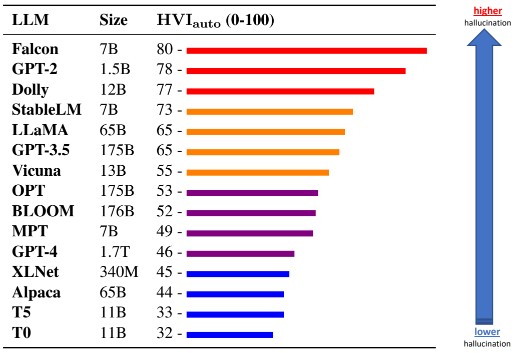


Figure 7: HVI_{auto} for four different hallucination categories across various 15 LLMs.

tation for HVI assessment. In this study, we introduce an automated hallucination metric, HVI_{auto} (Eq. (1)). Automating the detection and classification of hallucinations makes it feasible to calculate HVI automatically. To compute HVI_{auto} (see Fig. 7) for the LLMs discussed in Section 3, we leveraged 2,500 prompts from the HILT dataset (Rawte et al., 2023). These prompts were used to generate text from LLMs, and then FE was applied to the generated text to detect hallucinations and classify them into different types. When defining HVI_{auto}, we consider several factors. We think of U as the total number of sentences in the corpus. For instance, if LLM$_1$ produces significantly more time-related hallucinations than LLM$_2$, we cannot rank the same. This comparative measure is achieved using multiplicative damping factors, δ_{BN}, δ_{TI}, δ_{IF} and δ_P which are calculated based on $\mu \pm rank_x \times \sigma$. Initially, we calculate the HVI for all the LLMs, considering δ_{BN}, δ_{TI}, δ_{IF} and δ_P as one. With these initial HVIs, we obtain the mean (μ) and standard deviation (σ), allowing us to recalculate the HVIs for all the LLMs Fig. 8.

$$HVI_{auto} = \frac{100}{U}[\sum_{x=1}^{U}(\delta_{BN} * H_{BN} + \delta_{TI} * H_{TI} + \delta_{IF} * H_{IF} + \delta_P * H_P] \quad (1)$$



LLM	Size	$\mathbf{HVI_{auto}}$ (0-100)
Falcon	7B	80 -
GPT-2	1.5B	78 -
Dolly	12B	77 -
StableLM	7B	73 -
LLaMA	65B	65 -
GPT-3.5	175B	65 -
Vicuna	13B	55 -
OPT	175B	53 -
BLOOM	176B	52 -
MPT	7B	49 -
GPT-4	1.7T	46 -
XLNet	340M	45 -
Alpaca	65B	44 -
T5	11B	33 -
T0	11B	32 -

Figure 8: The HVI_{auto} scale shows hallucination tendencies LLMs.

Implications derived from $\mathbf{HVI_{auto}}$ are:

- Larger LLMs without RLHF (Ziegler et al., 2019) are prone to hallucination (see Fig. 8).
- Number-related issues are widespread across most LLMs, although they appear notably lower in specific models such as XLNet and StableLM. The reasons behind this discrepancy remain unclear and warrant further investigation.
- Hallucination categories such as Imaginary Figures and Temporal issues tend to increase with the size of LLMs.

7 Conclusion

LLMs' growing adoption and success have been remarkable, yet they face a critical challenge: hallucination. While recent works have explored hallucination mitigation, automatic detection remains under-explored. To bridge this gap, we present FACTOID, a dataset and benchmark for automatic hallucination detection. Our Factual Entailment technique has shown promising performance. We are committed to sharing all resources developed openly for further research.

8 Limitations

Limitations: The empirical findings indicate that classifying temporal issues poses the greatest challenge, as shown in Figure 6. (Gurnee and Tegmark, 2023) claimed that LLMs acquire linear representations of space and time across various scales, it is expected that LLMs hold such information internally and can classify accordingly. Performance on temporal issue 66% is not bad but could be seen as a future direction to improve.

9 Ethical Considerations

Through our experiments, we have uncovered the susceptibility of LLMs to hallucination. While emphasizing the vulnerabilities of LLMs, our goal is to underscore their current limitations. However, it's crucial to address the potential misuse of our findings by malicious entities who might exploit AI-generated text for nefarious purposes, such as designing new adversarial attacks or creating fake news that is indistinguishable from human-written content. We strongly discourage such misuse and strongly advise against it.

References

AI@Meta. 2024. Llama 3 model card.

Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 370–379, Minneapolis, Minnesota. Association for Computational Linguistics.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality*, 11(3):1–27.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Cambridge. 2023. 'hallucinate' is cambridge dictionary's word of the year 2023.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020a. All-in-one: A deep attentive multi-task learning framework for humour, sarcasm, offensive, motivation, and sentiment on memes. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 281–290, Suzhou, China. Association for Computational Linguistics.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020b. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online. Association for Computational Linguistics.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

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.

Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. KCTS: Knowledge-constrained tree search decoding with token-level hallucination detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14035–14053, Singapore. Association for Computational Linguistics.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loic Barrault, and Marta Costa-jussà. 2023. HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653, Singapore. Association for Computational Linguistics.

databricks. 2023. Dolly.

Kumar Shikhar Deep, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Related tasks can share! a multi-task framework for affective language. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 236–247. Springer.

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.

Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyy, Samira Khorshidi, Fei Wu, Ihab Ilyas, and Yunyao Li. 2023. FLEEK: Factual error detection and correction with evidence retrieved from external knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 124–130, Singapore. Association for Computational Linguistics.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Knowledge distillation of large language models. *Preprint*, arXiv:2306.08543.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *Preprint*, arXiv:2310.02207.

Michael Günther, Louis Milliken, Jonathan Geuter, Georgios Mastrapas, Bo Wang, and Han Xiao.

2023. Jina embeddings: A novel set of high-performance sentence embedding models. *Preprint*, arXiv:2307.11224.

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.

Vlad-Iulian Ilie, Ciprian-Octavian Truică, Elena-Simona Apostol, and Adrian Paschke. 2021. Context-aware misinformation detection: A benchmark of deep learning architectures using word embeddings. *IEEE Access*, 9:162122–162146.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. Fully automated fact checking using external sources. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 344–353, Varna, Bulgaria. INCOMA Ltd.

Sandeep Kumar, Tirthankar Ghosal, Prabhat Kumar Bharti, and Asif Ekbal. 2021. Sharing is caring! joint multitask learning helps aspect-category extraction and sentiment detection in scientific peer reviews. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 270–273.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen Mckeown, and Tatsunori B Hashimoto. 2023. 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.

Irene Larraz, Rubén Míguez, and Francesca Sallicati. 2023. Semantic similarity models for automated fact-checking: Claimcheck as a claim matching tool. *Profesional de la información*, 32(3).

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 34586–34599. Curran Associates, Inc.

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

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*.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

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*.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Mamta, Asif Ekbal, and Pushpak Bhattacharyya. 2022a. Exploring multi-lingual, multi-task, and adversarial learning for low-resource sentiment analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

Mamta, Asif Ekbal, and Pushpak Bhattacharyya. 2022b. Exploring multi-lingual, multi-task, and adversarial learning for low-resource sentiment analysis. *Transactions on Asian and Low-Resource Language Information Processing*, 21(5):1–19.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023a. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

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

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023b. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6859–6866.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

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.

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.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

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*.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2022. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

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`.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama

2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Robert A Wagner and Michael J Fischer. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173.

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*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *Preprint*, arXiv:2403.18802.

Wikipedia_Fleiss's_Kappa. Fleiss's kappa.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2019. A unified multi-task adversarial learning framework for pharmacovigilance mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5234–5245, Florence, Italy. Association for Computational Linguistics.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2021. If you want to go far go together: Unsupervised joint candidate evidence retrieval for multi-hop question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4571–4581, Online. Association for Computational Linguistics.

Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. A new benchmark and reverse validation method for passage-level hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3898–3908, Singapore. Association for Computational Linguistics.

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.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. 2024. In search of truth: An interrogation approach to hallucination detection. *arXiv preprint arXiv:2403.02889*.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023a. SAC^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15445–15458, Singapore. Association for Computational Linguistics.

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. *Preprint*, arXiv:2205.01068.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023b. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. Association for Computational Linguistics.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2023c. Alleviating hallucinations of large language models through induced hallucinations. *arXiv preprint arXiv:2312.15710*.

Theodore Zhao, Mu Wei, J Samuel Preston, and Hoifung Poon. 2023. Automatic calibration and error correction for large language models via pareto optimal self-supervision. *arXiv preprint arXiv:2306.16564*.

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.

10 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 inpainting 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 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.

A Appendix

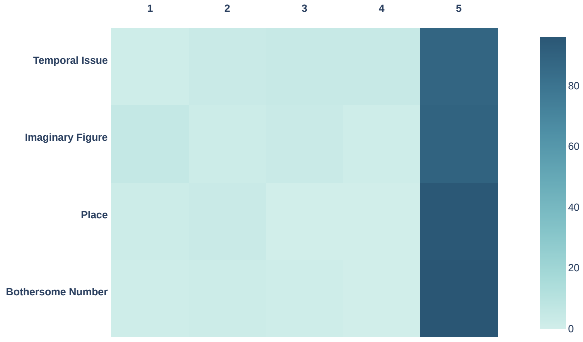


Figure 9: Heatmap of MOS scores with 100 manually assessed samples for each category by 5 annotators.

B Annotation Process, and agreement

In the initial in-house annotation phase, crowdsourcing platforms are acknowledged for their speed and cost-effectiveness in annotation tasks. Nevertheless, it's crucial to acknowledge that they may introduce noise or inaccuracies. To address this, prior to engaging crowdsourcing services, we conducted an in-house annotation process involving 1,000 samples.

C Paraphrasing

Coverage - Quantity of Significant Paraphrase Generations: Our aim is to create up to 5 paraphrases for each claim. Following the generation of claims, we employ the Minimum Edit Distance (MED) (Wagner and Fischer, 1974)—measured in words, not alphabets. If the MED exceeds ±2 for any paraphrase candidate (e.g., $c - p_1^c$) with the claim, we include that paraphrase; otherwise, we discard it. We assess all three models based on their ability to generate a substantial number of paraphrases.

Correctness - Accuracy in Paraphrase Generations: Post the initial filtration, we conduct pairwise entailment, retaining paraphrase candidates marked as entailed by (Liu et al., 2019) (Roberta Large), a state-of-the-art model trained on SNLI (Bowman et al., 2015).

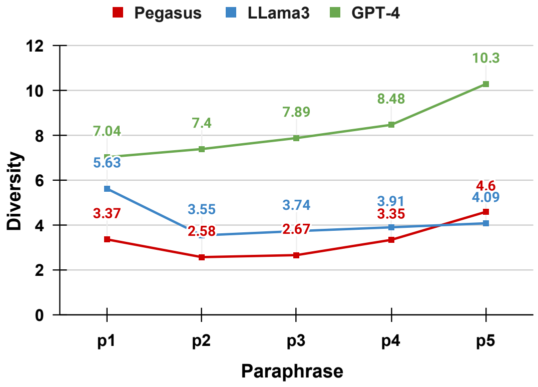


Figure 10: This figure shows the various parameters for generating paraphrases.

Diversity - Linguistic Variety in Paraphrase Generations: Our focus is on selecting a model capable of producing linguistically diverse paraphrases. We assess dissimilarities among generated paraphrase claims—for instance, $c - p_n^c$, $p_1^c - p_n^c$, $p_2^c - p_n^c$, and so on. This process is repeated for all paraphrases, averaging out the dissimilarity score. Lacking a specific dissimilarity metric, we use the inverse of the BLEU score (Papineni et al., 2002). This provides insight into how linguistic diversity is achieved by a given model. Our experiments reveal that gpt-4 performs the best, as reported in the table. Additionally, we prioritize a model that maximizes linguistic variations, and gpt-4 excels in this aspect. A plot illustrating diversity versus all chosen models is presented in Fig. 10.

D Dataset

The steps for creating positive-negative samples and the pipeline are shown in Algorithm 1 and Fig. 3.

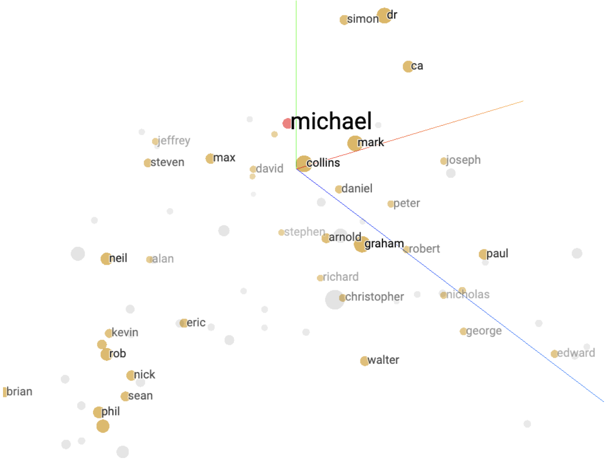


Figure 11: Similar person names.

E Longer embedding

Long-text embeddings are crafted to represent textual content and grasp the semantic essence of lengthy passages. In contrast to conventional embeddings for shorter texts that might face challenges in preserving context, longer text embeddings shine in capturing information from detailed articles, expansive books, or extensive documents. Defined by higher dimensions, usually spanning from 768 to 4096, they enable a nuanced understanding and the capture of relationships within extended textual contexts.

Model	Length
SFR-Embedding-Mistral	4096-dimensional embeddings over 32K tokens
e5-mistral-7b-instruct	4096-dimensional embeddings over 32K tokens
nomic-embed-text-v1	768-dimensional embeddings over 8K tokens
text-embedding-3-large	3072-dimensional embeddings over 8K tokens
jina-embeddings-v2-base	8192-dimensional embeddings over 8K tokens

Table 3: Examples of long-text embedding models.

E.1 Long-Text High-Dimensional Embeddings

In the realm of NLP, the advent of long-text embeddings marks a pivotal evolution from traditional, shorter embeddings, addressing critical limitations and broadening the application spectrum. Long-text embeddings, typically high dimensional ranging from 768 to 4096 dimensions, have emerged as a crucial innovation, primarily for their adeptness at encapsulating the semantics of extensive texts, ranging from detailed articles to comprehensive books. This capability significantly enhances document-level understanding, allowing for a more nuanced grasp of themes, narrative structures, argumentative patterns, etc. Moreover, the ability to process and analyze texts in their entirety without truncation reduces information loss, ensuring that vital context and intricate details are preserved. Long-text embeddings excel in capturing long-distance relationships and dependencies within texts, a feature that is instrumental for tasks requiring deep contextual interpretation such as question answering and textual entailment. Furthermore, these embeddings facilitate complex analyses, including thematic development, stylistic evolution, and sentiment tracking across lengthy documents, opening new avenues in literary analysis, historical research, and more. The shift towards longer text embeddings thus represents a significant leap forward in NLP, enabling more accurate, comprehensive, and sophisticated text processing and analysis, thereby overcoming the constraints posed by shorter embeddings and unlocking new potentials in understanding and leveraging large-scale textual data. This deep-rooted understanding offered by long-text embeddings is particularly beneficial for tasks that require a holistic understanding of the broader context, coupled with a nuanced understanding of the immediate topic at hand, to infer factual irregularities and thus detect hallucinations. Using the MTEB Leaderboard (Muennighoff et al., 2022), we identified the top-performing long-text embedding models as of this writing, with a max-token limit ranging from 8K to 32K.

The list of sentences is below:

sent1: “The sun sets behind the mountains, casting a warm glow across the landscape. The sky transforms into a canvas of vibrant hues, from fiery oranges to soft purples. The air becomes cooler as twilight descends upon the earth. Nature’s evening symphony begins, with the chirping of crickets and the rustle of leaves in the gentle breeze. As night falls, the world settles into a peaceful slumber, awaiting the dawn of a new day."

sent2: "As the sun dips beneath the silhouette of the mountains, its departing rays blanket the land with a comforting warmth, creating a picturesque scene. Gradually, the sky undergoes a breathtaking transformation, transitioning from the blazing brilliance of oranges to the soothing tones of purples, creating a mesmerizing spectacle overhead. With the fading light, a gentle coolness pervades the atmosphere, signaling the onset of twilight, a time when the earth enters a state of tranquil transition. Nature, in its evening rituals, orchestrates a harmonious symphony, with the melodious chirping of crickets and the gentle rustling of leaves accompanying the fading daylight. And so, as the darkness of night descends, the world surrenders to a serene slumber, patiently awaiting the emergence of a new dawn, heralding the promise of another day."

sent3: "Behind the rugged peaks, the sun gracefully retreats, suffusing the landscape with a radiant warmth that caresses every contour of the earth. Across the vast expanse, the heavens burst into an array of vibrant colors, from the fiery embrace of oranges to the tranquil embrace of purples, painting a captivating tableau above. As daylight wanes, a gentle chill creeps into the air, heralding the arrival of twilight, a transitional phase where the world pauses to catch its breath. Nature, in its evening chorus, serenades the fading light with the rhythmic chirping of crickets and the soft whispers of leaves dancing in the breeze. And so, with the advent of night, the world succumbs to a tranquil slumber, embracing the promise of renewal with each passing moment until the dawn of a new day breaks upon the horizon."

sent4: "The descent of the sun beyond the jagged peaks casts a golden glow upon the land, enveloping it in a serene embrace. Across the vast expanse of the sky, a kaleidoscope of colors emerges, transitioning from the fiery intensity of oranges to the gentle hues of purples and pinks, creating a breathtaking panorama. With the fading light, a sense of calmness descends, as the air grows cooler and the world prepares for the arrival of twilight. Nature, in its evening symphony, orchestrates a melodious chorus, with the chirping of crickets and the rustling of leaves providing the soundtrack to the fading day. And so, as night falls, the world settles into a tranquil slumber, eagerly anticipating the promise of a new beginning with the break of dawn."

sent5: "Behind the majestic peaks, the sun bids adieu, casting a warm glow that envelops the landscape in a comforting embrace. The sky transforms into a canvas of breathtaking beauty, with hues ranging from the fiery brilliance of oranges to the soft pastels of purples and pinks, creating a mesmerizing display. As daylight fades, a gentle coolness fills the air, signaling the arrival of twilight, a magical time when the earth transitions into a state of serene tranquility. Nature, in its nightly ritual, comes alive with the chirping of crickets and the gentle rustling of leaves, as if bidding farewell to the departing day. And so, as darkness descends, the world settles into a peaceful slumber, eagerly awaiting the dawn of a new day and the promise it brings."

sent6: "As the sun dips below the horizon, its fading rays cast a golden glow upon the land, imbuing it with a sense of warmth and serenity. Above, the sky transforms into a breathtaking tapestry of colors, with vibrant oranges giving way to soft purples and pinks, painting a scene of unparalleled beauty. With the onset of twilight, the air grows cooler, enveloping the world in a gentle embrace as it prepares for the night ahead. Nature, in its nightly symphony, fills the air with the soothing sounds of crickets chirping and leaves rustling, a melodic accompaniment to the fading light. And so, as night falls, the world settles into a peaceful slumber, eagerly anticipating the dawn of a new day and the endless possibilities it brings."

F Details of performance of FE

Entailment technique/ Hallucination Type	Imaginary Figure	Place	Bothersome Number	Temporal Issue	Avg.
Traditional entailment	0.44	0.49	0.23	0.12	0.32
Factual entailment	0.69	0.71	0.67	0.59	0.665

Table 4: Average overall performance improvement of FE across all four hallucination categories.