FACtual enTailment fOr hallucInation Detection

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

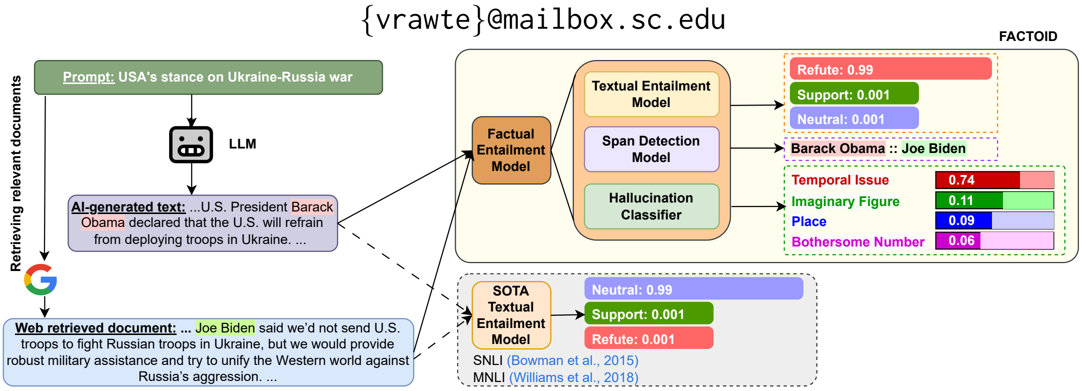


Figure 1: An illustration of traditional Textual Entailment (TE) vs.our proposed Factual Entailment (FE). In part A (top), we emphasize the limitation of the TE method (trained on standard entailment tasks like SNLI (Bowman et al., 2015) and/or MNLI (Williams et al., 2018), etc.) to recognize a case as a refute. In contrast, in part (B), the proposed Factual Entailment adopts a multitask learning approach that predicts an entailment score, hallucination type and the span of the entailment. FE therefore presents a novel approach to entailment that assists in identifying hallucinations. the retrieved document is a White House press release, could be see here: link

Abstract

The widespread adoption of Large Language Models (LLMs) has facilitated numerous benefits and applications. However, among the various risks and challenges, hallucination is a significant concern. In response, Retrieval Augmented Generation (RAG) has emerged as a highly promising paradigm to improve LLM outputs by grounding them in factual information. RAG relies on textual entailment (TE) or similar methods to check if the text produced by LLMs is supported or contradicted, compared to retrieved documents. This paper argues that conventional TE methods are inadequate for spotting hallucinations in content generated by LLMs. For instance, consider a prompt about the "*USA's stance on the Ukraine war*". The AI-generated text states, "*...U.S. President Barack Obama says the U.S. will not put troops in Ukraine...*" However, during the Ukraine-Russia war, the U.S. president is Joe Biden, not Barack Obama, which contradicts factual reality. Moreover, current TE systems are unable to accurately annotate the given text and identify the exact portion that is contradicted. To address this challenge, this paper introduces a new type of TE called "Factual

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

Entailment (FE).", aims to detect factual inaccuracies in content generated by LLMs while also highlighting the specific text segment that contradicts reality. We present $\mathcal{FACTOID}$ (FACTual enTAILment for hallucInation Detection), a benchmark dataset for FE. We propose a multi-task learning (MTL) framework for FE, incorporating state-of-the-art (SoTA) long text embeddings such as e5-mistral-7b-instruct, along with GPT-3, SpanBERT, and RoFormer. The proposed MTL architecture for FE achieves an avg. 40% improvement in accuracy on the $\mathcal{FACTOID}$ benchmark compared to SoTA TE methods. As FE automatically detects hallucinations, we assessed 15 modern LLMs and ranked them using our proposed *Auto Hallucination Vulnerability Index (HVI_{auto})*. This index quantifies and offers a comparative scale to evaluate and rank LLMs according to their likelihood of producing hallucinations. FACTOID dataset[1] and demo[2] are publicly available.

Contributions

- Introducing a new type of TE called "Factual Entailment (FE)", aims to detect factual inaccuracies in content generated by LLMs while also highlighting the specific text segment that contradicts reality. (cf. Sec. 1).
- Presenting $\mathcal{FACTOID}$ (FACTual enTAILment for hallucInation Detection), a benchmark dataset for FE (cf. Sec. 4).
- We propose an MTL framework for FE, yielding 30% improvement in accuracy on the $\mathcal{FACTOID}$ 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. 7).

1 FACTUAL Entailment: The Necessity

Large generative AI models like GPT (Brown et al., 2020; OpenAI, 2023), Stable Diffusion (Rombach et al., 2022), DALL-E (Ramesh et al., 2021, 2022), and Midjourney (Midjourney, 2022), face various challenges related to the risk of potential misuse. One such major challenge of Large Language Models (LLMs) is generating factually incorrect responses, which is referred to as *hallucination*. Recently, numerous techniques for mitigating hallucinations have been proposed, including i) Retrieval Augmented Generation (Peng et al., 2023; Vu et al., 2023; Kang et al., 2023; Gao et al., 2023), ii) Self Refinement through Feedback and Reasoning (Si et al., 2022; Mündler et al., 2023; Chen et al., 2023), iii) Prompt Tuning (Cheng et al., 2023; Jones et al., 2023), iv) Introducing a New Decoding Strategy (Chuang et al., 2023; Li et al., 2023), v) Utilization of Knowledge Graph (Bayat et al., 2023), vi) Introducing Faithfulness based Loss Function (Yoon et al., 2022; Qiu et al., 2023b), and vii) Supervised Finetuning (Elaraby et al., 2023; Tian et al., 2023; Qiu et al., 2023a).

Hallucination mitigation has received considerable research attention recently, with Retrieval Augmented Generation (RAG) being considered the most promising approach to eliminate hallucinations in LLM generation. The working principle of RAG involves providing a prompt p_1 to the LLM, which generates text t_1. Since the LLM's factual knowledge is limited to its training data, it retrieves relevant documents or information $(r_1, r_2, r_3, ..., r_n)$ from a repository or search engine. This retrieved information is then used as context when generating the text from the LLM. Recent research suggests that RAG can effectively mitigate hallucinations to a certain extent (). However, this area is still evolving, and we anticipate further progress soon. Nonetheless, we argue that before and after applying any mitigation technique, it's crucial to understand the hallucination rate. Automatic hallucination detection is essential in this regard.

A straightforward solution to this could be to utilize state-of-the-art textual entailment (TE) techniques and adapt them for hallucination detection. The three possible outcomes of any TE method are (i) support/entailment, (ii) contradict/refute, and, (iii) neutral/not enough information. However, we have empirically demonstrated that SoTA TE techniques have significant shortcomings in

[1] https://huggingface.co/datasets/aisafe/FACTOID
[2] https://huggingface.co/spaces/aisafe/FACTOID

terms of detecting factual errors in LLM-generated text. While the lack of entailment could signal the occurrence of hallucination, it should not be misconstrued as a definitive indicator of whether hallucination exists. For instance, what if both the first and second sentences are hallucinated? In that case, the fact that the sentences are entailed does not convey actionable insight as to whether hallucination is present. Similarly, the lack of entailment does not automatically mean that hallucination is occurring; it may simply indicate that the information provided is insufficient or that the texts are discussing different aspects of a topic. Therefore, a more nuanced approach is needed. This approach requires a combination of textual entailment recognition, factual verification, and span detection to mark the specific sections of both source and target text that contradict each other. One such scenario has been illustrated in Fig. 1.

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. The hallucination categories we consider are as follows:

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 any 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...

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

3 Choice of LLM

We have chosen 15 modern LLMs that consistently exhibit excellent performance across various NLP tasks, as 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 $\mathcal{FACTOID}$: Factual Entailment Dataset

We present $\mathcal{FACTOID}$ (FACTual enTAILment for hallucInation Detection), a benchmark dataset for FE containing total containing 2 million text pairs. Details are given in Table 2. $\mathcal{FACTOID}$ is a synthetic extension of HILT dataset introduced by (Rawte et al., 2023). HiLT comprises a total of 492K sentences, out of which 129K are annotated for hallucination, indicating that 364K sentences are factually correct. At this juncture, we aim to synthesize these 129K sentences further for the factual entailment (FE) task. To accomplish this, we devise hallucination category-specific techniques, as detailed below:

Bothersome Numbers (BN): The HiLT dataset contains 7275 sentences associated with number-related hallucinations. Our aim is to produce more negative samples for Factual Entailment (FE) by randomly adjusting numbers in these sentences. However, mere number changes might not consistently create valid entailment scenarios. To overcome this, we applied automatic paraphrasing techniques (explained in Section X). Numbers were detected using regular expressions and altered randomly within a range of $\pm 20\%$, as shown by the blue-marked numbers in the example. These paraphrased sentences effectively refute the originals.

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%.

Temporal Issue (TI): The HiLT dataset, containing about 7,500 sentences from the Time Wrap category of Factual Mirage, focuses on time-related hallucinations. Our goal is to expand negative samples for FE by randomly altering the entities of two individuals from different time periods within these sentences. Recent studies indicate that LLMs grasp linear representations of space and time across various scales (Gurnee and Tegmark, 2023), which inspired our experiment design. The experiment setup is semi-automatic, requiring human intervention.

We identified an entity and manually formulated a question: "*When did the Amber Alert program start?*" We posed this question to an LLM and received the response: "*The Amber Alert program officially began in 1996.*" Subsequently, we randomly selected a number between 50 and 150 and subtracted it from 1996 to determine the desired timeframe, which in this case (*let's assume*) was 1806. We then asked the LLM, "*Who was the President of the USA in 1806?*" and received the answer: "*Thomas Jefferson.*" We substituted "*Obama*" with "*Jefferson*" in all automatically generated paraphrases. We chose Llama for this task based on its usage in prior research (Gurnee and Tegmark, 2023). Although this process required manual intervention, we were able to manage the generation process with two student annotators over a two-week period, given the 7.5K sentences in the dataset.

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 Jefferson administration made the decision to suspend operations of the Amber Alert program.
Para §3 During the government shutdown in October 2013, under the Jefferson administration, the Amber Alert program halted.

Imaginary Figure (IF): The HiLT dataset contains 15K sentences focusing on person-related hallucinations, particularly from the Generated Golem category in Factual Mirage. Our aim is to enhance negative samples for Factual Entailment (FE) by randomly altering the names of individuals in these sentences. We utilize an automatic paraphrasing technique detailed in Section X for this task. Named Entity Recognition (NER) () helps us identify person names within prompts.

Then, leveraging a pre-trained Word2Vec-based (Mikolov et al., 2013) Euclidean distance measure, we locate other person names in close vector space proximity. An experimental Euclidean threshold guides this process.

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

Place (P): The HiLT dataset includes approximately 13K sentences related to location-related hallucinations, specifically from the Geographic Erratum category of the Factual Mirage dataset. Our objective is to create additional negative samples for Factual Entailment (FE) by randomly modifying the names of individuals mentioned in these sentences. We utilize similar techniques as those used for person names. Initially, we apply Named Entity Recognition (NER) () to identify location names within a given prompt. Subsequently, we utilize a pre-trained Word2Vec-based Euclidean distance measure to identify other location names that are distant in vector space. For this analysis, we establish an experimental Euclidean 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. Among the victims were a patient who was being transported and a family member accompanying them.

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

Hallucination classes: Given that $\mathcal{FACTOID}$ extends the HILT dataset, and since HILT already contains manually annotated categories, we simply transferred those categories directly to $\mathcal{FACTOID}$.

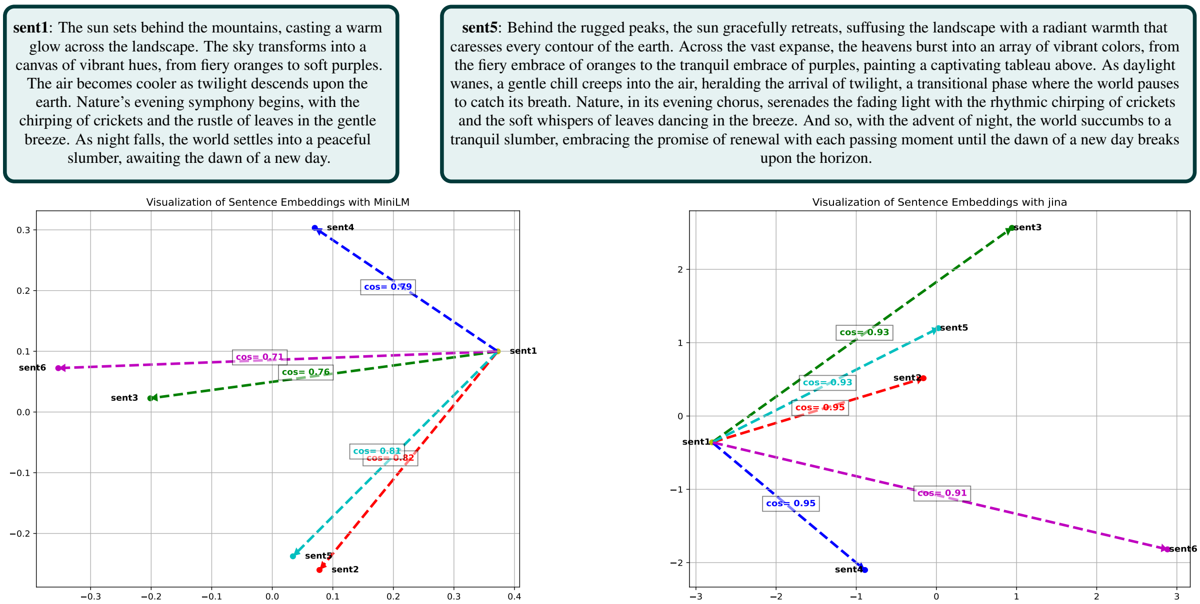
4.1 Automatic Paraphrasing

When choosing automatic paraphrasing, there are many other factors to consider for e.g., a model may only be able to generate a limited number of paraphrase variations compared to others, but others can be more correct and/or consistent. As such, we consider three major dimensions in our evaluation: *(i)* ***Coverage****: a number of considerable generations, (ii)* ***Correctness****: correctness in those generations, and (iii)* ***Diversity****: linguistic diversity in those generations*. We conducted experiments with three available models: (a) Pegasus (**?**), (b) T5 (T5-Large) (Raffel et al., 2020), and (c) GPT-3 (`text-davinci-003` variant) (Brown et al., 2020). Based on empirical observations, we concluded that GPT-3 outperformed all the other models. To offer transparency around our experiment process, we detail the aforementioned evaluation dimensions as follows.

Model	Coverage	Correctness	Diversity
Pegasus	32.46	94.38%	3.76
T5	30.26	83.84%	3.17
GPT-3	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-3 (`text-davinci-003`) is the most performant considering all three aspects.

A comprehensive discussion regarding Coverage, Correctness, and Diversity, along with the experimental setup for paraphrasing, is available in Appendix C.



(a) Vanilla sentence embedding.

(b) Longer sentence embedding.

Figure 2: 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.

4.2 *FACTOID*: Statistics

FACTOID extends the HiLT dataset synthetically. HiLT encompasses a total of 492K sentences, with 129K annotated for hallucination, leaving 364K sentences deemed factually correct. As we exclusively expand the hallucinated sentences through paraphrasing, the resulting *FACTOID* dataset may suffer from class imbalance. To address this, we also expanded the 364K factually correct sentences. A statistical overview of *FACTOID* is presented in Table 2.

	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: *FACTOID* dataset statistics.

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. In our work, we introduce several key contributions in terms of design choices, including the use of different LLMs for different tasks, employing long-text embedding, SpanBERT, RoFormer, and implementing specific loss functions as per the requirements of each task. Further details about these nuances are discussed below.

5.1 Long-Text High-Dimensional Embeddings

Long-text embeddings in NLP signify a transformative shift from traditional shorter embeddings, overcoming limitations and expanding application possibilities. Ranging from 768 to 4096 dimensions, these embeddings excel at capturing the semantics of extensive texts, enhancing document-level comprehension. They mitigate information loss by processing entire texts without truncation, preserving

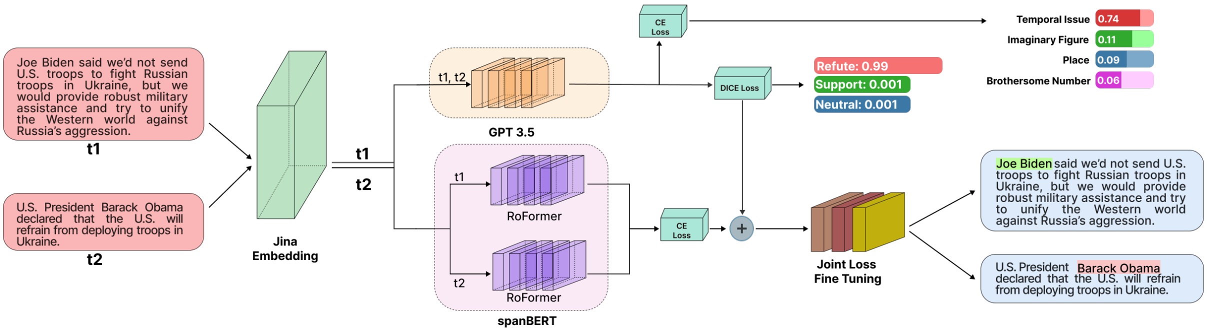


Figure 3: A summary of the overall multi-task learning framework for Factual Entailment. The framework encompasses three tasks: i) entailment, ii) span detection, and iii) hallucination classification.

crucial context and details. Notably adept at grasping long-distance relationships, they prove invaluable for tasks like question answering and textual entailment, enabling sophisticated analyses in thematic development, stylistic evolution, and sentiment tracking. This advancement in NLP unlocks new potentials, offering a deep understanding for tasks requiring both holistic context comprehension and nuanced topical insight. Since entailment is a classification task, we chose `e5-mistral-7b-instruct` based on its top classification performance reported on the MTEB Leaderboard (Muennighoff et al., 2022). Fig. 2 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:

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.

5.2 Introducing Span-based Textual Entailment

The example in Fig. 3 illustrates a case where an LLM, discussing the Russia-Ukraine war, incorrectly identifies *Barack Obama* as the US President instead of *Joe Biden*. Despite being deemed 'supportive' in textual entailment, the text contains a factual inaccuracy or 'hallucination.' To improve accuracy, the passage suggests refining text analysis by focusing on specific spans rather than entire sentences.

SpanBERT: It is specifically designed to understand and represent spans of text (Joshi et al., 2020), making it useful for tasks involving relationships between different segments of a document or passage. It also enhances the capabilities of BERT by considering the context of spans, enabling a more nuanced understanding of language structure and meaning.

RoFormer: Introduced in (Su et al., 2022), utilizes a rotation matrix to encode absolute position while incorporating explicit relative position dependencies in self-attention formulation. This approach, featured in RoFormer, imparts beneficial properties such as sequence length flexibility, diminishing inter-token dependency with increasing relative distances, and the ability to integrate relative position encoding into linear self-attention.

5.3 Loss Functions

We employed cross-entropy loss for span detection and hallucination type identification, while dice loss (Sudre et al., 2017) proved to be the best fit for entailment. Due to the significant imbalance in the support class, we opted for dice loss, known for its effectiveness in handling imbalanced datasets.

6 Performance of FE

Our empirical findings depicted in Fig. 8 illustrate that the proposed Factual Entailment (FE) outperforms the state-of-the-art textual entailment (TE) methods. Some key takeaways are listed below:

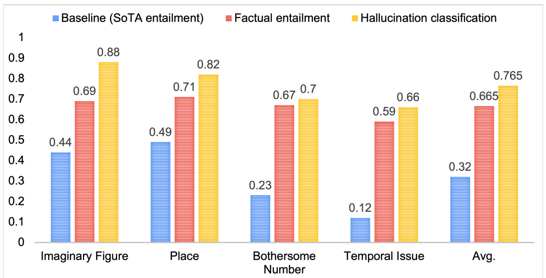


Figure 4: Results showing how FE performs better than TE at detecting hallucination in six different categories.

7 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 annotation for HVI assessment. In this study, we introduce an automated hallucination metric, HVI_{auto}, as defined in Eq. (1). By automating the detection and classification of hallucinations, it is now feasible to calculate HVI automatically. To compute HVI_{auto} 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 Factual Entailment (FE) was applied to the generated text to detect hallucinations and classify them into different types. When defining HVI_{auto}, we take several factors into account. We consider U as the total number of sentences we have in the corpus. Moreover, two/more LLMs can exhibit varying characteristics of hallucination, including person, location, time and number. For instance, if we have two LLMs and their total number of generated hallucinations in terms of sentences are the same, but LLM_1 produces significantly more time related hallucinations than LLM_2, we cannot rank them 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. The resulting HVIs are then ranked and scaled providing a comparative spectrum as presented in Fig. 6. Having damping factors enables easy exponential smoothing with a handful of data points, similar to z-score normalization (Wikipedia_zscore) and min-max normalization (Wikipedia_min_max). Finally, for ease of interpretability, HVI is scaled between $0 - 100$.

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

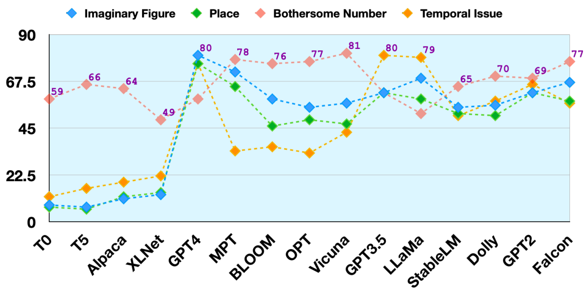


Figure 5: HVI for different hallucination categories across various LLMs.

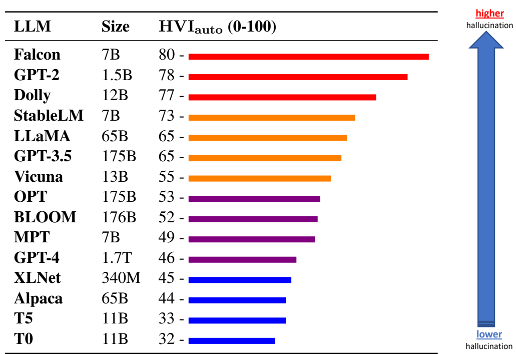


Figure 6: The HVI scale illustrates the hallucination tendencies exhibited by various LLMs.

Implications derived from $\mathrm{HVI}_{\mathrm{auto}}$

- Larger LLMs without RLHF (Ziegler et al., 2019) are prone to hallucination, as shown in Fig. 6.
- Number-related issues are widespread across most LLMs, although they appear notably lower in certain models such as XLNet and StableLM. The reasons behind this discrepancy remain unclear and warrant further investigation in the future.
- Hallucination categories such as Imaginary Figures and Temporal issues tend to increase with the size of LLMs.

8 Conclusion

The growing adoption and success of LLMs have been remarkable, yet they face a critical challenge: hallucination. While recent works have explored hallucination mitigation, automatic detection remains underexplored. To bridge this gap, we present $\mathcal{FACTOID}$, 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.

9 Discussion and Limitations

Discussion: On June 14th, 2023, the European Parliament successfully passed its version of the EU AI Act (European-Parliament, 2023). Following this, many other countries began discussing their stance on the evolving realm of Generative AI. A primary agenda of policymaking is to protect citizens from political, digital, and physical security risks posed by Generative AI. While safeguarding against misuse is crucial, one of the biggest concerns among policymakers is the occurrence of unwanted errors by systems, such as hallucination (source: https://cetas.turing.ac.uk/publications/rapid-rise-generative-ai).

Limitations: The empirical findings indicate that classifying temporal issues poses the greatest challenge, as shown in Figure 4. (Gurnee and Tegmark, 2023) claimed that LLMs acquire linear representations of space and time across various scales, it is expected that LLM 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.

10 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

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*.

Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyi, Samira Khorshidi, Fei Wu, Ihab F. Ilyas, and Yunyao Li. 2023. Fleek: Factual error detection and correction with evidence retrieved from external knowledge.

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*.

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.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. *arXiv preprint arXiv:2311.10081*.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Denvy Deng, and Qi Zhang. 2023. Uprise: Universal prompt retrieval for improving zero-shot evaluation.

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.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models.

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.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models.

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.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Knowledge distillation of large language models.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time.

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.

Erik Jones, Hamid Palangi, Clarisse Simões, Varun Chandrasekaran, Subhabrata Mukherjee, Arindam Mitra, Ahmed Awadallah, and Ece Kamar. 2023. Teaching language models to hallucinate less with synthetic tasks.

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.

Haoqiang Kang, Juntong Ni, and Huaxiu Yao. 2023. Ever: Mitigating hallucination in large language models through real-time verification and rectification.

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.

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.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.

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*.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

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*.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report.

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.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Yifu Qiu, Varun Embar, Shay B Cohen, and Benjamin Han. 2023a. Think while you write: Hypothesis verification promotes faithful knowledge-to-text generation. *arXiv preprint arXiv:2311.09467*.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. 2023b. Detecting and mitigating hallucinations in multilingual summarisation. *arXiv preprint arXiv:2305.13632*.

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.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Tonmoy, Aman Chadha, Amit P Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models–an extensive definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2310.04988*.

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.

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*.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.

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

Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. 2017. *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*, page 240–248. Springer International Publishing.

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`.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality.

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*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation.

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*.

Wikipedia_min_max. Normalization.

Wikipedia_zscore. Normalization.

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.

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.

Sunjae Yoon, Eunseop Yoon, Hee Suk Yoon, Junyeong Kim, and Chang Yoo. 2022. Information-theoretic text hallucination reduction for video-grounded dialogue. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4182–4193, Abu Dhabi, United Arab Emirates. 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.

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 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

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

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).

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-3.5-turbo-0301` performs the best, as reported in the table. Additionally, we prioritize a model that maximizes linguistic variations, and `gpt-3.5-turbo-0301` excels in this aspect. A plot illustrating diversity versus all chosen models is presented in **??**.

D FACTOID dataset creation

The process for creating the synthetic dataset is given in Algorithm 1,

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 {s1, s2, s3, s4, s5}
    for each similar entity s do
        replace the original entity with a similar entity
        generate 5 paraphrases {p1, p2, p3, p4, p5}
    end for
end for
```

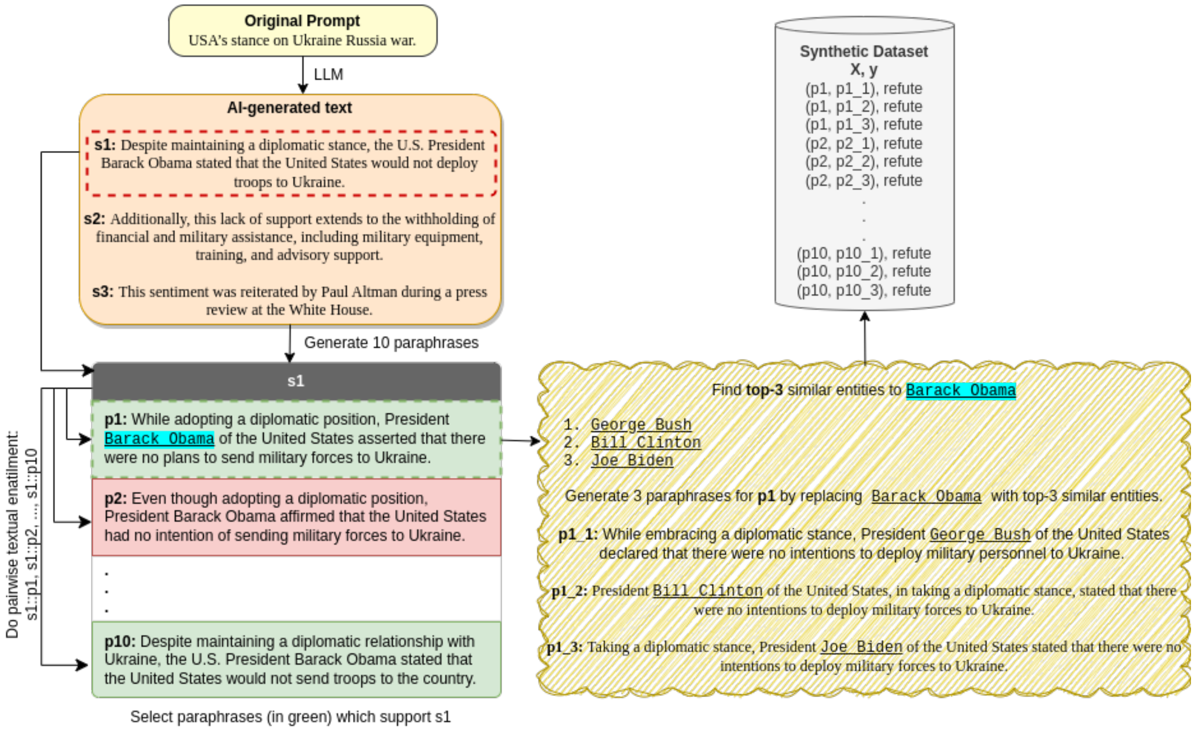


Figure 7: Process to generate synthetic data.

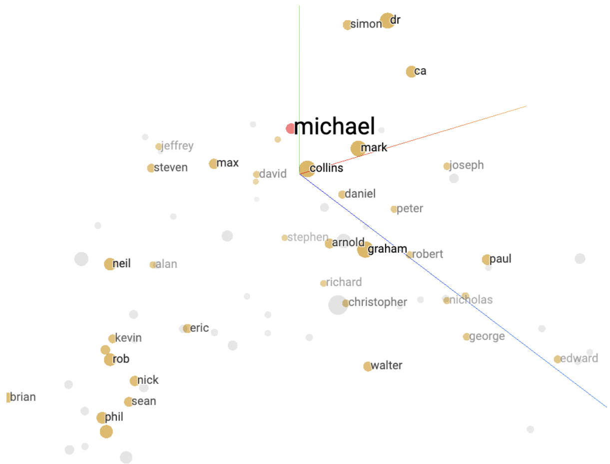


Figure 8: similar person names

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.

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.
Baseline (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