# HalluFormer: A Transformer-Based Framework for Detecting Hallucination in Large Language Models

**Sajid Mahmud**[1*], **Pawan Neupane**[1*], **Joel Selvaraj**[1*], **Bhanu Prakash Vangala**[1*], **Jianlin Cheng**[1†]

[1]Department of Electrical Engineering and Computer Science
[1]University of Missouri
Columbia, Missouri, USA
{smmkk, pngkg, jsbrq, bv3hz, chengji}@missouri.edu

## Abstract

Despite the impressive performance of Large Language Models (LLMs) in a variety of natural language processing tasks, they are still prone to producing information that is factually inaccurate, known as hallucination. In critical fields related to scientific and clinical domains that demand highly precise answers, the negative effect of this phenomenon is even more pronounced. To address this problem, we formulate the hallucination detection problem as a classification problem of assessing the consistency between questions, answers and retrieved knowledge contexts and propose HalluFormer, a transformer-based model for detecting hallucinations of LLMs. HalluFormer was trained and tested on the MultiNLI dataset. It achieves an F1 score of 0.9471 on the MultiNLI test dataset. On the blind ANAH test dataset, it achieves an F1 score of 0.7285, indicating it can generalize reasonably well to completely new data. The results demonstrate that transformer-based methods can be utilized to detect hallucinations of LLMs, paving the way for further research on improving the reliability of LLMs.

## Introduction

Large Language Models (LLMs) have revolutionized natural language processing, achieving state-of-the-art performance in tasks like translation, summarization, text generation, text classification, question-answering, etc.

Despite their ability to answer complex questions and generate knowledge-driven information for many domains, LLMs are prone to a phenomenon called hallucination where it generates answers which are not logically correct and cannot be verified using existing knowledge base. This phenomenon can have severe consequences when the domain of interest requires high precision and factual accuracy, like in the medical field or material design, etc. Detecting and eliminating hallucination is important for ensuring the trustworthiness of artificial intelligence (AI) systems which incorporate LLMs.

The focus of existing LLMs is primarily on overall accuracy and fluency, but hallucination often gets overlooked. This necessitates the development of generalized and independent AI methods for detecting hallucinations.

In this work, we propose a framework based on transformer (Vaswani et al. 2017), HalluFormer, for detecting hallucinations, which categorizes the responses produced by LLMs into two labels: hallucination and no hallucination.

The original user question, the LLM-generated response, and a corresponding knowledge text about the topic of the inquiry are the three inputs for HalluFormer. The Big Bird (Zaheer et al. 2020) tokenizer, which enables effective processing of lengthy sequences while maintaining contextual linkages, is used to concatenate and tokenize these inputs.

A classification head and a Big Bird Transformer encoder (Zaheer et al. 2020) make up the core components of HalluFormer. A multi-layer perceptron classifier is used to predict the hallucination labels after the encoder processes the tokenized inputs and creates contextualized embeddings. It is worth noting that, in addition to the Big Bird Transformer, other transformer-based encoders, such as BERT (Devlin et al. 2019), LongFormer (Beltagy, Peters, and Cohan 2020)), can also be used with HalluFormer.

According to our experiment, HalluFormer is able to rather accurately identify hallucinations, especially when LLM-generated response is contradictory to the ground truth knowledge. HalluFormer's ability to identify hallucination in long answers with long-range relationships is notable. This approach demonstrates the potential of transformer-based models to enhance hallucination detection.

## Related Work

Several studies have proposed methods to detect and mitigate hallucinations in LLM outputs. For instance, an entropy-based uncertainty estimator has been developed to identify hallucinated content by assessing the confidence levels of the model's predictions (Farquhar et al. 2024)). Additionally, the ReID discriminator has been introduced to effectively discern hallucinations in LLM-generated answers, demonstrating robustness across various datasets (Yuan, Xie, and Ananiadou 2024).

Comprehensive surveys have been conducted to categorize and analyze hallucinations in LLMs. One such survey presents a taxonomy of hallucinations, delving into their causes and proposing detection and mitigation strate-

[*]These authors contributed equally.
[†]Corresponding Author

gies (Huang et al. 2023) Another study provides a detailed examination of hallucination phenomena, offering insights into their definition, quantification, and potential remedies (Rawte et al. 2023).

Recent empirical studies have focused on understanding the sources of hallucinations and developing benchmarks for their evaluation. For example, a systematic study introduced the HaluEval 2.0 benchmark and designed a detection method to assess LLM hallucinations, analyzing factors contributing to their occurrence (Li et al. 2024).

Despite these advancements, challenges persist in fully understanding and mitigating hallucinations in LLMs. Ongoing research continues to explore various techniques, including enhancing model architectures, refining training processes, and developing post-processing methods to improve the factual accuracy and reliability of LLM-generated content.

In summary, addressing hallucinations in LLMs is a critical area of research, with efforts focusing on detection, mitigation, and leveraging models like Big Bird to enhance textual data.

## Methodology

We formulate the hallucination detection problem as a classification problem of assessing the consistency and agreement between users' questions, LLMs' answers (responses), and knowledge contexts related to the topics of the questions retrieved from authoritative sources (e.g., internet or data bases). The knowledge context is useful because it provides some factual evidence to detect the hallucination and can usually be obtained by search engines such as Google or other means. HalluFormer takes the triplet of the input to predict if there is hallucination or not.

HalluFormer is a Big Bird transformer-based hallucination detection model (Figure 1). It is made up of two parts, a pretrained Big Bird Transformer based on a Bert model and a classifier head which is essentially a multi-layer perceptron network. HalluFormer takes in a question, a related knowledge text (context) and an answer for the question produced by a LLM as the input, and it will predict if the answer is hallucinated or not. All three inputs (context, question, answer) are concatenated and passed through a tokenizer, which tokenizes the texts so that they can be used by HalluFormer.

### Data Collection and Processing

The MultiNLI (Williams, Nangia, and Bowman 2017) dataset contains 433,000 sentence pairs of premises and hypothesis annotated with textual entailment information. If the hypothesis follows the premise, then it is labeled as "entailment", if the hypothesis contradicts the premise, it is labeled as "contradiction" and if the hypothesis shows neither entailment nor contradiction, then it is labeled as "neutral". In this work, we are particularly interested in developing a model that can identify "contradiction" like hallucinations. We reorganized the MultiNLI dataset by re-labelling rows that contain "contradiction" as hallucination and "entailment" as no hallucination. We removed the rows containing "neutral" annotation as it is neither an entailment nor a contradiction.

HalluFormer is designed to take three inputs (knowledge context, question and answer), but the MultiNLI has only two inputs, the premise which can be considered as the context and the hypothesis which can be considered as the answer. Since the MultiNLI dataset does not contain questions, we augmented it by formulating a generic question such as "Given the premise, form a valid hypothesis that is consistent with the premise" and used LLMs to create 22 variations of the same question. One of these 22 question variations is randomly chosen for a pair of context (premise) and answer (hypothesis). The final processed MultiNLI dataset is split into train, validation and test sets, containing 235622, 26180 and 13393 samples respectively.

### Tokenizer

The Big Bird tokenizer is used to convert textual inputs into tokenized representations that can be processed by the neural network model. In the dataset, each data sample consists of a context component, a user query (question) and the corresponding LLM generated response (answer). The contextual relationships between these elements were needed to be preserved, thus the input is structured in such a way that the context is treated as its first sequence, while the question and answer are concatenated as the second sequence separated by a [SEP] token.

The tokenization is achieved using the Big Bird tokenizer. This involves splitting the input text into sub word tokens, adding special tokens such as [CLS] at the beginning and [SEP] at the appropriate segment boundaries, and truncating sequences that exceed a predetermined maximum length of 512. If the input sequence is shorter than the maximum length, padding tokens are added to maintain uniform sequence length across different inputs.

The outputs of the tokenizer are both input IDs and an attention mask. The input IDs correspond to a numerical representation of the tokenized words, while the attention mask is used to differentiate between actual tokens and padding tokens. The actual tokens which correspond to meaningful text are assigned a value of 1, and for padding tokens they are assigned a value of 0.

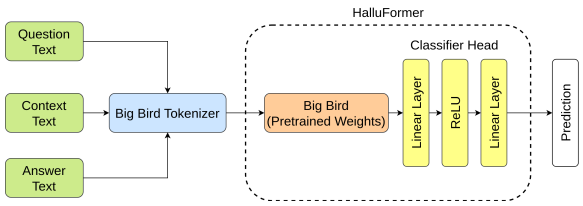



Figure 1: The architecture of HalluFormer

HalluFormer is designed to classify LLM generated answers/ responses into two categories: No Hallucination and Hallucination. The architecture consists of two parts: the Big Bird encoder and the classification head 1.

The Big Bird encoder serves as the feature extractor, by processing the token inputs and generating a contextualized numerical representation. Processing an input token and corresponding attention mask, the model computes a sequence

of hidden states. The [CLS] token embedding serves as the summary representation of the entire input, which is extracted from the final hidden state output of the encoder.

The classification head is a feedforward neural network applied to the extracted [CLS] embedding. It consists of a fully connected layer that maps the hidden size of Big Bird of 768 to an intermediate layer of size 256. A ReLU activation and a dropout layer with a probability of 0.3 is followed for the purpose of mitigating any possibility of overfitting. A final fully connected layer maps the intermediate representations to a vector of size 2, which corresponds to the previously discussed hallucination categories. The model produces raw logits, which are passed through a Softmax activation layer to get the class probabilities.

During the forward pass, the input consists of input_ids and attention mask. The Big Bird model processes these, and its output embedding associated with the [CLS] token is passed through the classification head for the purpose of generating class logits. This model architecture primarily leverages the self-attention mechanism of Big Bird to effectively model long-range dependencies as well as handling long sequences in an efficient manner. This makes it well suited for the task of hallucination detection in LLM generated responses.

### Training

The model was trained on the MultiNLI dataset with cross-entropy loss as the loss function and AdamW optimizer being used for the purpose of optimization with a fixed learning rate of 0.001.

After the completion of each epoch, the model undergoes a validation loop, where the model is being evaluated on the validation set. Several evaluation metrics are used to track the progress of the model, which include precision, recall and F1 score of hallucination detection. At the end of the validation phase, if the F1 score is the best to the point, the model weights are saved.

## Results

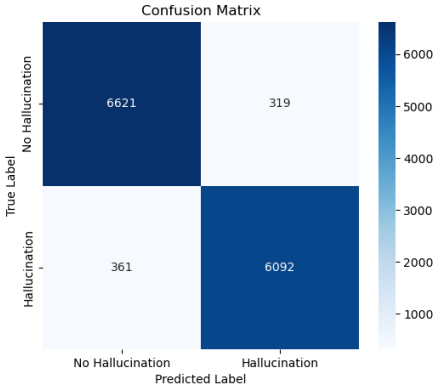
After HalluFormer was trained on MultiNLI training dataset, it was evaluated on the test set. In addition to evaluating on the MultiNLI dataset, we blindly evaluated the trained HalluFormer on the ANAH dataset (Ji et al. 2024).

The ANAH dataset contains sentence-level hallucination annotations for the LLM generated answers given the question and knowledge. We split each question and each LLM answer into separate rows along with their respective knowledge document text (context). In each row, if even one sentence in the answer contains hallucination based on the annotation provided, we consider it hallucination and if none of the sentences contains hallucination, then we consider it as no hallucination. Finally, we filtered the dataset to only contain examples in English language and selected only rows where the inputs (context, question, answer) combined contains less than 512 tokens. The final processed ANAH dataset contains 683 rows on which we blindly evaluated HalluFormer.

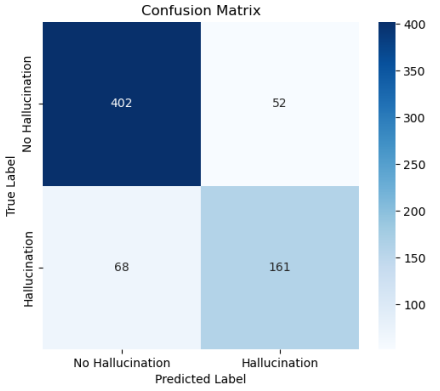
| Dataset | Precision | Recall | F1 |
|---|---|---|---|
| MultiNLI test | 0.9502 | 0.9441 | 0.9471 |
| ANAH | 0.7559 | 0.7031 | 0.7285 |

Table 1: Performance metrics on different datasets

In Table 1, the precision, recall and F1 scores are reported for HalluFormer on the two different test datasets – MultiNLI and ANAH. The model achieves a very high detection accuracy on the MultiNLI test dataset. As this dataset only contains hallucination of "contradiction" type, the results indicate that HalluFormer is quite effective in identifying such kind of hallucination in LLM generated responses. The ANAH dataset has a domain shift, containing hallucinations of other types not seen in the training data, resulting in a decrease of detection accuracy (e.g., F1 score drops from 0.9471 on the MultiNLI dataset to 0.7285 on the ANAH dataset). However, a decent F1 score of 0.7285 indicates that HalluFormer generalizes reasonably well to the new data. Using more data containing various types of hallucinations to train it should further enhance its performance.



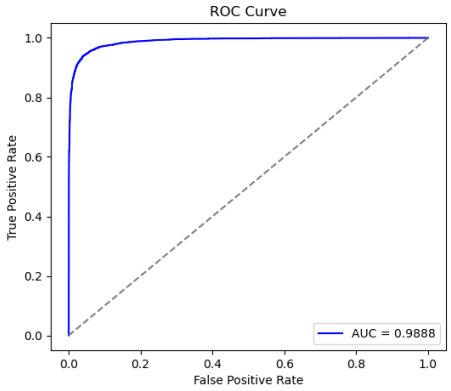

(a)



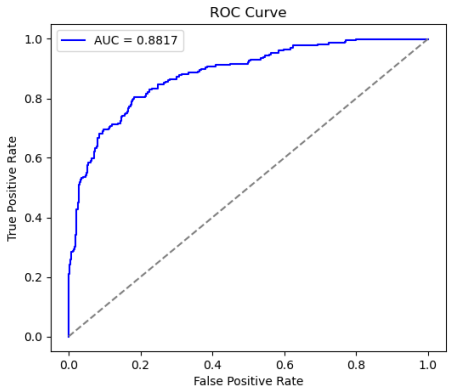

(b)

Figure 2: Confusion Matrices showcasing the performance of HalluFormer on (a) MultiNLI dataset and (b) ANAH dataset

In Figure 2a the confusion matrix shows that the HalluFormer is performing well on the MultiNLI dataset, with the majority of the data points correctly classified. Only a small number of false positives (319) and false negatives (361) are seen, indicating that HalluFormer is effective in correctly classifying "contradiction" like hallucination from non-hallucinated responses. Figure 2b shows the performance on the ANAH dataset, where performance is reasonable but weaker than on the MultiNLI dataset. With the model misclassifying 52 non-hallucinated responses as hallucination, and 68 hallucinated responses as non-hallucination, there is a decrease in the recall and precision. This difference in performance is likely due to a domain shift, considering ANAH may contain hallucinated responses which are unlike "contradiction".




(a)




(b)

Figure 3: ROC curves of the performance of HalluFormer where (a) is on the MultiNLI dataset with an AUC score of 0.9888 and (b) is on the ANAH dataset with an AUC score of 0.8817.

In Figure 3a, analyzing the ROC curve on the MultiNLI dataset, we can clearly see a near perfect classification performance from HalluFormer, with an AUC score of 0.9888. Conversely, Figure 3b shows a good but lower performance, as the curve seen here does not rise as sharply as in Figure 3a. The AUC score on the ANAH dataset is 0.8817, which despite being less than the one seen in the MultiNLI dataset, is still pretty good.

## Impact of Training on HalluFormer's Performance

As HalluFormer uses the pretrained Big Bird transformer to process the input, it is necessary to check whether the training of HalluFormer helps improve its performance over the pretrained transformer itself without any further downstream training. For this analysis, we compared the untrained HalluFormer with the trained one on the two test datasets.

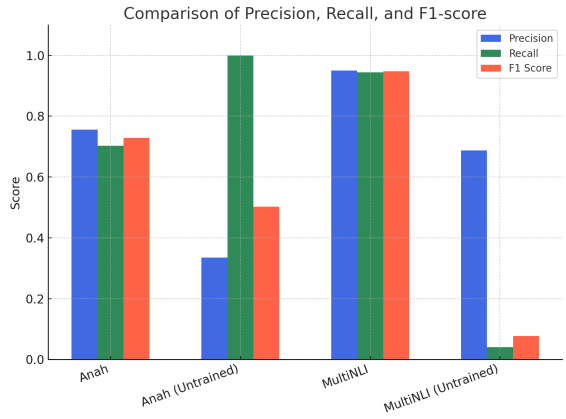



Figure 4: Comparison of precision, recall and F1 scores between the untrained and trained HalluFormer on the two datasets

Figure 4 shows a bar chart comparison of the precision, recall and F1 scores across the two datasets: ANAH and MultiNLI for the trained and untrained HalluFormer. It demonstrates a clear improvement in performance with training being done. In the ANAH dataset, precision rises from 0.3353 to 0.7559, while the recall drops moderately from 1.0 to 0.7031, achieving a higher, balanced F1 score. The decrease in recall indicates that the untrained model classifies nearly every example as hallucination, whereas upon training the model is more competent in successfully distinguishing between non-hallucinated and hallucinated responses.

For the MultiNLI dataset, the improvement is even more apparent, as the untrained model has a poor recall of 0.0408, which jumps to 0.9441 and the precision becomes 0.9502, effectively getting an F1 score of 0.9471 after training. This solidly demonstrates that the further training of the pretrained Big Bird along with the classifier head in HalluFormer makes it significantly more competent in identifying hallucination in LLM responses.

## Conclusion

In this work, we formulate the hallucination detection problem as a problem of assessing the consistency between questions, answers, and knowledge contexts. Under this conceptual framework, we develop HalluFormer, a Big Bird

transformer-based language model combined with a multilayer perceptron-based classifier head, for detecting presence of hallucination in LLM generated responses. Using the MultiNLI and ANAH dataset, we showcase the effectiveness of HalluFormer's ability to detect hallucination, in particular hallucination that contains contradictory answers in comparison to the knowledge. This finding hints at the viability of such models as they possess the capability of capturing long-range contextual understanding, which may be an important aspect to focus on for the goal of identifying hallucination.

Despite such early promising results, one of the next challenges would be to scale this approach on larger and more diverse datasets, preferably spanning different domains and hallucination types to further improve its generalizability of HalluFormer. Based on the available GPU memory, a fixed token length of 512 was optimally chosen. Further experiments can be conducted with longer token lengths, which will enable the model to analyze more context information and detect hallucination in LLM-generated answers that span multiple paragraphs.

Moreover, it is necessary to improve the explainability and interpretability of hallucination predictions. Further analysis on this could shed light upon the core reason of hallucination occurring in LLM responses in the first place.

Finally, by addressing these challenges, future advancements could lead to a more robust, reliable, interpretable, and domain-adaptable hallucination detection models, ensuring great usability in AI applications.

## Acknowledgments

This work is partly supported by a subaward from Arizona State University entitled "Accelerating Material Design and Manufacturing through Artificial Intelligence and Machine Learning".

## References

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.

Ji, Z.; Gu, Y.; Zhang, W.; Lyu, C.; Lin, D.; and Chen, K. 2024. Anah: Analytical annotation of hallucinations in large language models. *arXiv preprint arXiv:2405.20315*.

Li, J.; Chen, J.; Ren, R.; Cheng, X.; Zhao, W. X.; Nie, J.-Y.; and Wen, J.-R. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*.

Rawte, V.; Chakraborty, S.; Pathak, A.; Sarkar, A.; Tonmoy, S. I.; Chadha, A.; Sheth, A.; and Das, A. 2023. The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Williams, A.; Nangia, N.; and Bowman, S. R. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Yuan, C.; Xie, Q.; and Ananiadou, S. 2024. Temporal relation extraction with contrastive prototypical sampling. *Knowledge-Based Systems*, 286: 111410.

Zaheer, M.; Guruganesh, G.; Dubey, K. A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33: 17283–17297.