# Hallucination Detection - A comparative study

Raghavendra Kulkarni
*Department of Computer Science*
*Indian Institute of Technology*
Hyderabad, India
cs23mtech11016@iith.ac.in

Supriya Rawat
*Department of Computer Science*
*Indian Institute of Technology*
Hyderabad, India
cs23mtech11019@iith.ac.in

Madhumitha V
*Department of Artificial Intelligence*
*Indian Institute of Technology*
Hyderabad, India
ai23resch11004@iith.ac.in

***Abstract*—The field of Natural Language Processing has discovered new dimensions of research and progress since the introduction of the revolutionary Transfomer-based Large Language Models. With their powerful capabilities of processing humongous text inputs like news articles, Twitter posts, scientific journals, literature novels, and other text sources, the large language models fit in a wide range of applications. However, the large language models tend to generate grammatically sound yet factually inaccurate or non-sensical text data when used in real-world scenarios. This phenomenon is called Hallucination. Detecting hallucinations in an LLM-generated text is an active field of research in Natural language Processing today. In this work, we present a comparative study of three techniques used to detect factually incorrect text by applying them to four different large language models with varying sizes measured in the number of trainable parameters. We aim to investigate the impact of the number of parameters of a large language model on its tendency to hallucinate.**

***Index Terms*—Hallucination, Large Language Models, SelfCheckGPT**

## I. Introduction

The introduction of Large Language Models has proven to be significant progress towards success as they find applications in a wide range of real-world tasks like sentiment analysis, web query completion, and predictive typing in emails and smartphones. However, these benefits of large language models pose new substantial challenges, like foul language detection in an LLM-generated text, violation of data privacy, copyright issues in LLM text generation, and others. One such challenge is to detect factually incorrect text generated by an LLM, called a Hallucination. Since large language models output the next word or sentence based on a probabilistic distribution controlled by the previously generated text, they tend to deviate from the context or generate irrelevant and false information as the length of the generated output text increases. This phenomenon results in misinformation or misguidance, causing harm to a user's trust and reliability of the system. The Hallucination detection task has been an open problem in Natural Language Processing. Research works are in progress to explore new techniques for detecting hallucinations from machine-generated text.

The Literature Survey section explores and summarizes the work done in this field of research. After defining the Problem Statement, we discuss the three scoring methods of interest. The Implementation section gives the details of dataset generation and implementation of the scoring system. Finally, the Results and Conclusion section compares and summarizes the scoring methods on different large language models varying in size, measured in the number of parameters.

## II. Literature Survey

The need for reliability and user trust in a system has necessitated the research work towards Hallucination detection. To detect hallucinations efficiently, we must study the different types of hallucinations possible in an LLM-generated text. A formal world framework defines hallucinations as the inconsistencies between a computable LLM and a computable ground truth function, showing that the complete elimination of hallucination from an LLM is impossible[1]. A taxonomy of Hallucination categorizes it into two classes: Factuality Hallucinations and Faithfulness Hallucinations[2]. A comparative analysis of Conversational Large Language Models shows that few-shot prompting and other post-processing and fine-tuning techniques improve the capabilities of large language models in triple verbalization[3]. An LLM evaluation framework quantitatively evaluates interactive LLMs like ChatGPT in multitasking, multilingual, and multimodal aspects. Human collaboration with the LLM improves its performance on summarization and machine translation in a multi-turn prompt engineering fashion[4]. An analysis of Hallucination in the medical generative QA systems proposes an interactive self-reflection methodology, incorporating knowledge acquisition and answer generation and improving the factuality, consistency, and entailment of the generated answers through a feedback process[5]. The Chain-Of-Verification (CoVE) method eliminates hallucinations by drafting an initial response for a query and then verifying atomic claims of the response independently through question generation and refining the response by combining the verified claim responses[6]. A lightweight knowledge-free framework, HALOCHECK quantifies the severity of hallucinations using knowledge injection and teacher-student approaches to alleviate hallucinations in low-parameter LLMs like BLOOM-7B[7]. Another study finds that the activations of some hidden layer of an LLM are enough to predict the final response. The LLM model trained on these hidden activations can estimate the final output by learning to reject a sample and avoid generating a false result[8].

## III. Problem Statement

"To conduct a comparative study of Hallucinated Text generated by various Large Language Models with different ML parameters using the SelfCheckGPT model."

We choose the following four Large Language Models for the comparative study:

- GPT-3 (175B parameters)
- LED (164M parameters)
- BART-base (137M parameters)
- T5 (78M parameters).

This work aims to conduct a comparative study of three techniques for detecting hallucination: N-Gram scoring, BERT scoring, and MQAG scoring, and investigate the impact of the number of parameters of a large language model on its tendency to hallucinate.

## IV. SelfCheck-GPT Scoring

In this section, we discuss the three scoring systems proposed by the SelfCheck-GPT. The principle idea behind the SelfCheckGPT scoring model is that, if a response is non-factual or hallucinated, the other responses generated by the LLM, for the same query but at a different set of parameters, won't be consistent with the original response[9].

So, for each of the scoring method, we have a query $Q$, an LLM response $R$ for the query $Q$ and $N$ sample responses $\{S^1, S^2, S^3, ....., S^N\}$ taken from the same LLM as of $R$ by varying parameters like temperature, sampling method, etc.

### A. N-Gram Scoring

We use the sample responses $\{S^1, S^2, S^3, ....., S^N\}$ as a training corpus to train a Unigram model $M$. We also use the LLM Response $R$ in the training corpus as a smoothing factor to avoid the possibilities of any zero probabilities. Now we calculate the probability of model $M$ generating the response $R$ as

$$S^{Avg}_{n-gram}(i) = -\frac{1}{J}\sum_{j} \log p_{ij}$$

### B. BERT Scoring

Let $BERT(i, j)$ denote the BERT-Score between the sentences $s_i$ and $s_j$. In this scoring, we find the BERT score for every sentence $r_i$ from the response $R$ with the most similar sentence $s_j$ from each sample $S^n$. The sentence hallucination score of sentence $r_i$ is given by the equation:

$$BERT(r_i) = 1 - \frac{1}{N}\sum_{n=1}^{N} \max_{k} BERT(r_i, s^n_k)$$

### C. MQAG Scoring

In this scoring, we generate a question $q$ and four corresponding choices $o_1, o_2, o_3$ and $o_4$ from every response sentence $r_i$. Then, we give the question and four choices to a Question Answering system prompting it to answer the question using the sample responses $\{S^1, S^2, S^3, ....., S^N\}$ as the context. If $R$ is factual, then the sample responses will be in alignment with the response $R$ and hence the QA system will be able to correctly answer the question $q$ using these sample responses as context.

Let $a_R$ be the true answer from the response $R$ for the question $q$ and let $a_{S^n}$ be the answer given by the QA system using the sample responses as the context. Then the consistency score between $a_R$ and $a_{S^n}$ is calculated as the MQAG score of the response sentence $r_i$.

$$S_{QA}(i) = E_q[S_{QA}(i, q)]$$

## V. Methodology

The Fig.1 below shows the methodology of our comparative study.

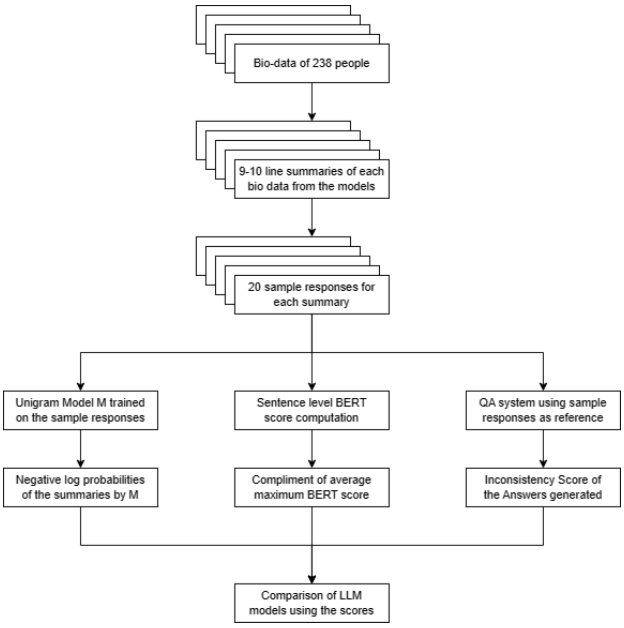


Fig. 1. The Methodology

### A. Dataset Description

We use the SelfCheckGPT dataset for our implementation. The dataset contains a complete Bio-data of 238 people, summary of these Bio-data generated by GPT-3, an annotation for each sentence of the summaries categorizing them as **major_inaccurate**, **minor_inaccurate** and **accurate** and finally 20 sample summary responses generated by GPT-3 by varying the parameters.

### B. Dataset Generation

We take the bio-data of 238 people and prompt the other three LLMs: T5, BART-base and LED, to generate a summary response $R$. We then vary the parameters of the LLMs to set the temperature to 1.2 and use top-50 method for sampling. With this settings, we prompt the LLMs to generate 20 more sample responses $S$ for the bio data. These are used to calculate the N-gram, BERT and MQAG scores. Finally, we compare the performance of the LLMs based on each scoring method.

## VI. Results

We divide the analysis part into two sections. In the first part, we compare the sentence level hallucination scores of GPT-3 model with the human annotated ground truth labels. In the second part, we compare the passage level hallucination scores of the other three LLMs with GPT-3 model.

### A. Sentence Level Hallucination Score Comparison

The Fig.2 below shows the comparison of average Unigram Scores of GPT-3 summary sentences against each ground truth annotation labels. Since we are computing the average negative log probability, we expect the major inaccurate, i.e., factual but out of context sentences to have higher unigram scores compared to the other two labels, and the results seem to be in alignment with this expectation.

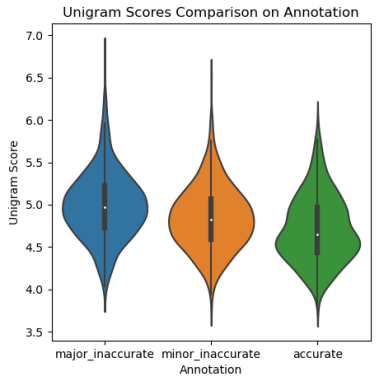


Fig. 2. Unigram Scores Comparison on Annotation

The Fig.3 below shows the comparison of average BERT Scores of GPT-3 summary sentences against each ground truth annotation labels. Since we subtract the average BERT score of the sentences from 1, we expect the major inaccurate to have higher BERT scores compared to the other two labels, and the results seem to be in alignment with this expectation.

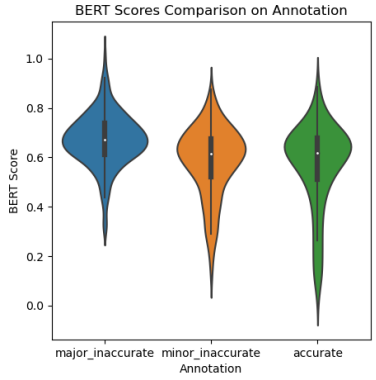


Fig. 3. BERT Scores Comparison on Annotation

The Fig.4 below shows the comparison of average MQAG Scores of GPT-3 summary sentences against each ground truth annotation labels. Since we compute the inconsistency of the answers $a_R$ and $a_{S^n}$, we expect the major inaccurate to have higher MQAG scores compared to the other two labels. But the results of MQAG scores seem to deviate from this expectation.

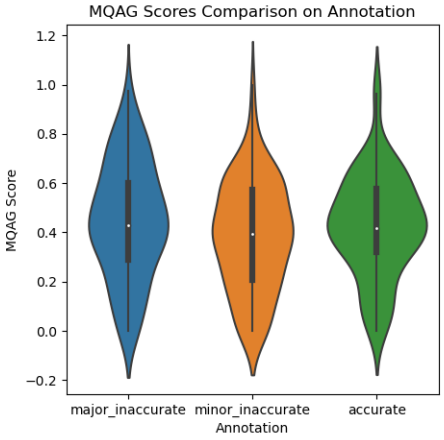


Fig. 4. MQAG Scores Comparison on Annotation

### B. Passage Level Hallucination Score Comparison

We now compare the Passage Level Hallucination scores for the summaries generated by the LLMs.

First we determine which score gives a distribution for in alignment with the ground truth annotations of for the GPT-3 model. To compute this, we assign scores 1.0, 0.5 and 0.0 for the three annotations **major_inaccurate**, **minor_inaccurate** and **accurate** respectively. Then we take the average of each score over all the sentences in each summary and plot the distribution against the ground truth annotation scores. Fig.5 below shows this plot. We see that the Unigram score matches the most among all the three scores, with the distribution of the ground truth annotation scores at the passage level.


Fig. 5. Passage Level LLM Comparison

Now we compare the passage level hallucination scores across the models. We have chosen four models with varying size measured in number of trainable parameters. In this section, we compare the hallucination score range and distribution of these LLMs with each other. Due to the limited computational resources, we restricted the number of passages for the plots to 100 for the Unigram and BERT scores and 30 for the MQAG score instead of considering all the 238 passage summaries.

Fig.6 shows the Plot of the Average Unigram scores per Passage and Fig.7 shows the Violin Plot distribution of the same. We see that the GPT-3 model with 175B parameters shows a high average and wide range distribution of hallucination scores, whereas the other models with lesser parameters (in the order of millions) show a narrow range distribution for Unigram score.

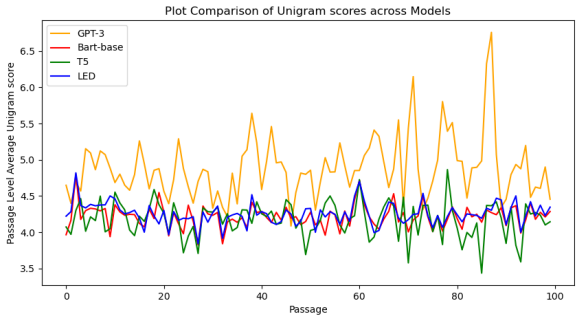


Fig. 6. Plot Comparison of Unigram scores across Models

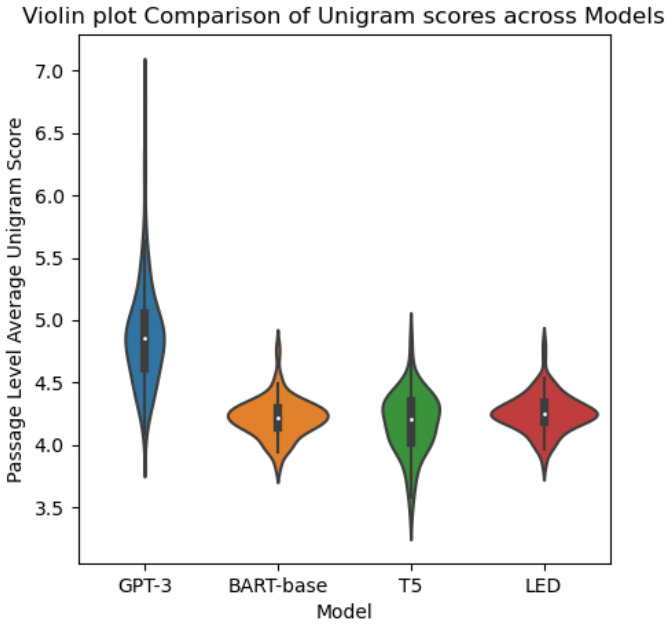


Fig. 7. Violin plot Comparison of Unigram scores across Models

Fig.8 shows the Plot of the Average BERT scores per Passage and Fig.9 shows the Violin Plot distribution of the same. We see that both the GPT-3 model with 175B parameters and T5 model with 78M parameters show a high average and wide range distribution of hallucination scores, whereas the other two models BART-base and LED with intermediate (137M and 164M respectively) parameters show a narrow range distribution for BERT score.

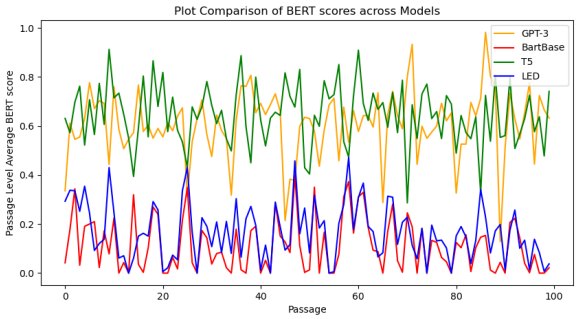


Fig. 8. Plot Comparison of BERT scores across Models

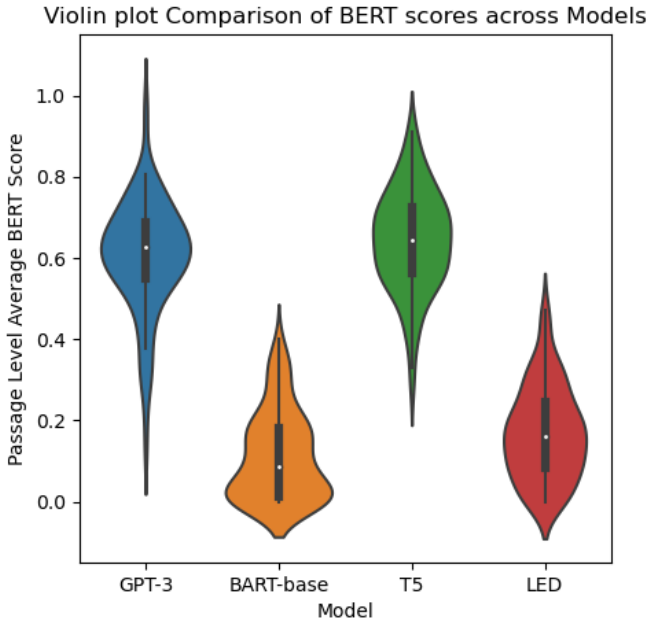


Fig. 9. Violin plot Comparison of BERT scores across Models

Fig.10 shows the Plot of the Average MQAG scores per Passage and Fig.11 shows the Violin Plot distribution of the same. We see that both the GPT-3 model with 175B parameters and T5 model with 78M parameters show a high average and wide range distribution of hallucination scores compared to the other two models BART-base and LED with intermediate (137M and 164M respectively) parameters for MQAG score.

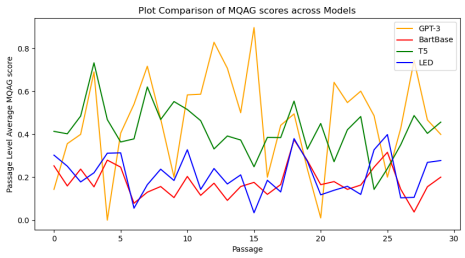


Fig. 10. Plot Comparison of MQAG scores across Models

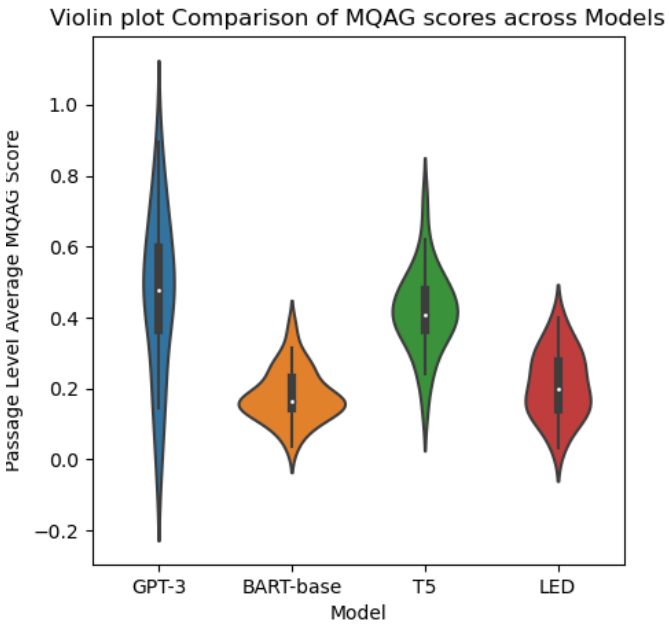


Fig. 11. Violin plot Comparison of MQAG scores across Models

## VII. Conclusion

The objective of this work was to compare the hallucinating tendency of different LLMs against their size in terms of number of parameters they are trained with. It is evident that the hallucinating tendency does not vary linearly with the size of the model. As the inferences justify, too smaller models (with very few parameters $<$ 100M like T5) and too larger models (with very high parameters $>$ 100B like GPT-3) both are likely to hallucinate more.

This work also analyses the different scoring variants of SelfCheckGPT for hallucination detection. We find that n-gram performs better among the three variants.

## References


[1] Z. Xu, S. Jain, and M. S. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models," *ArXiv*, vol. abs/2401.11817, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267069207.

[2] L. Huang, W. Yu, W. Ma, *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ArXiv*, vol. abs/2311.05232, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265067168.

[3] P. Schneider, M. Klettner, E. Simperl, and F. Matthes, "A comparative analysis of conversational large language models in knowledge-based text generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, Y. Graham and M. Purver, Eds., St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 358–367. [Online]. Available: https://aclanthology.org/2024.eacl-short.31.

[4] Y. Bang, S. Cahyawijaya, N. Lee, *et al.*, "A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity," in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, J. C. Park, Y. Arase, B. Hu, *et al.*, Eds., Nusa Dua, Bali: Association for Computational Linguistics, Nov. 2023, pp. 675–718. DOI: 10.18653/v1/2023.ijcnlp-main.45. [Online]. Available: https://aclanthology.org/2023.ijcnlp-main.45.

[5] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards mitigating LLM hallucination via self reflection," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1827–1843. DOI: 10.18653/v1/2023.findings-emnlp.123. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.123.

[6] S. Dhuliawala, M. Komeili, J. Xu, *et al.*, "Chain-of-verification reduces hallucination in large language models," *ArXiv*, vol. abs/2309.11495, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:262062565.

[7] M. S. Elaraby, M. Lu, J. Dunn, X. Zhang, Y. Wang, and S. Liu, "Halo: Estimation and reduction of hallucinations in open-source weak large language models," *ArXiv*, vol. abs/2308.11764, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261076218.

[8] A. Azaria and T. Mitchell, "The internal state of an LLM knows when it's lying," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 967–976. DOI: 10.18653/v1/2023.findings-emnlp.68. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.68.

[9] P. Manakul, A. Liusie, and M. J. F. Gales, *Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models*, 2023. arXiv: 2303.08896 [cs.CL].