Detection of Hallucination in LLAMA 65B responses on Trivia, Hotpot & TruthfulQA Datasets

Abstract

Large Language Models (LLMs) have rapidly gained prominence, highlighted by OpenAI's advancements with the Generative Pre-Trained Transformer (GPT) series. This success has not only validated the potential of LLMs but also spurred significant investment from both private and open-source sectors into their development.

Our study is designed to assess the effectiveness of these large language models (LLMs) by employing both human evaluation metrics and computational measures such as cosine similarity and edit distance. This dual approach provides a comprehensive analysis of the models' responses, enabling us to gauge their closeness to the ground truth from multiple perspectives.

Our research focuses on the performance impact of these models when tested against the TriviaWeb, TriviaWiki, TruthfulQA & HotpotQA datasets. By utilizing LLAMA 65b, we aim to provide insights into how dataset and metric selection influence outcomes.

Introduction

Large Language Models (LLMs) are now widely recognized as essential tools in various fields due to their impressive ability to produce text that resembles human writing. Nevertheless, the issue of hallucination within LLMs, where the generated text may lack factual correctness or logical consistency, poses a notable obstacle. In this paper, we examine the application of the LLaMA 65b large language model to generate responses compared against the ground truth using human evaluation metrics alongside computational measures such as edit distance and cosine similarity. These models,

notably diverse in their training data and parameter scales, provide an excellent basis for assessing LLM efficacy in realistic scenarios.

The LLaMA65b model, developed by Meta, is a part of the Large Language Model Meta AI (LLaMA) series designed to deliver high performance across various inference budgets while being trained exclusively on publicly available data. This makes it suitable for open-source applications, aiming to maximize accessibility and utility.

Our focus shifts towards an analysis of how both models perform on our four datasets. The effectiveness of the models is measured not only through qualitative human assessments but also quantitatively via edit distance and cosine similarity metrics. These metrics are critical as they quantitatively evaluate the similarity of LLM responses to the ground truth, providing a clear measure of model accuracy and coherence.

Levenshtein's (1966) introduction of edit distance provides foundational insights into text-based errors . Edit distance measures the number of operations required to transform the model's response into the ground truth, offering insights into the model's precision at a granular level.

Cosine similarity as stated by Bojanowski et al. (2017) assesses the cosine of the angle between the vector representations of the model's response and the ground truth.

These metrics, combined with human evaluations, furnish a comprehensive picture of model performance across various dimensions, supporting the development of more effective, nuanced LLMs.

Literature Review

Researchers in the field of natural language processing (NLP) have been exploring ways to detect hallucinations in Large Language Models (LLMs) using knowledge-based approaches, often leveraging advanced parametric models and libraries involved. They've made significant progress by fine-tuning these models on specific tasks. One common task involves generating the responses and comparing the generated text with the expected result on various performance metrics. This helps in determining the degree to which the response generated is hallucinated from ground reality.

Integration of HuggingFace

Our research is significantly enhanced by incorporating HuggingFace's innovative platform, which greatly boosts our abilities in natural language processing (NLP). By utilizing HuggingFace's extensive collection of pre-trained models and cutting-edge tools, we've improved our analysis capabilities and made our workflows more efficient.

HuggingFace's platform enables researchers and practitioners to easily access, deploy, and refine a wide range of language models. For our study, we specifically utilized the Llama 65b models. Through seamless integration with Hugging Face's Model Hub, we were able to tap into state-of-the-art large language models that were pre-trained on massive datasets. This significantly sped up our experimentation and analysis.

Additionally, HuggingFace's platform encourages collaboration and knowledge exchange within the NLP community, fostering innovation and interdisciplinary research. By leveraging the tools and resources provided by HuggingFace, we were able to deepen our understanding of hallucination detection in large language models and contribute to the responsible deployment of AI.

Methodology

The LLM Model architecture, loaded directly from the HuggingFace community, defines the specific implementation and mathematical structure of a language model, detailing its parameters and their interactions with inputs. Currently, the most successful Language Model architectures are often variations of the "decoder-only" Transformer architecture, and we have used the LLama 65B model to help with the research.

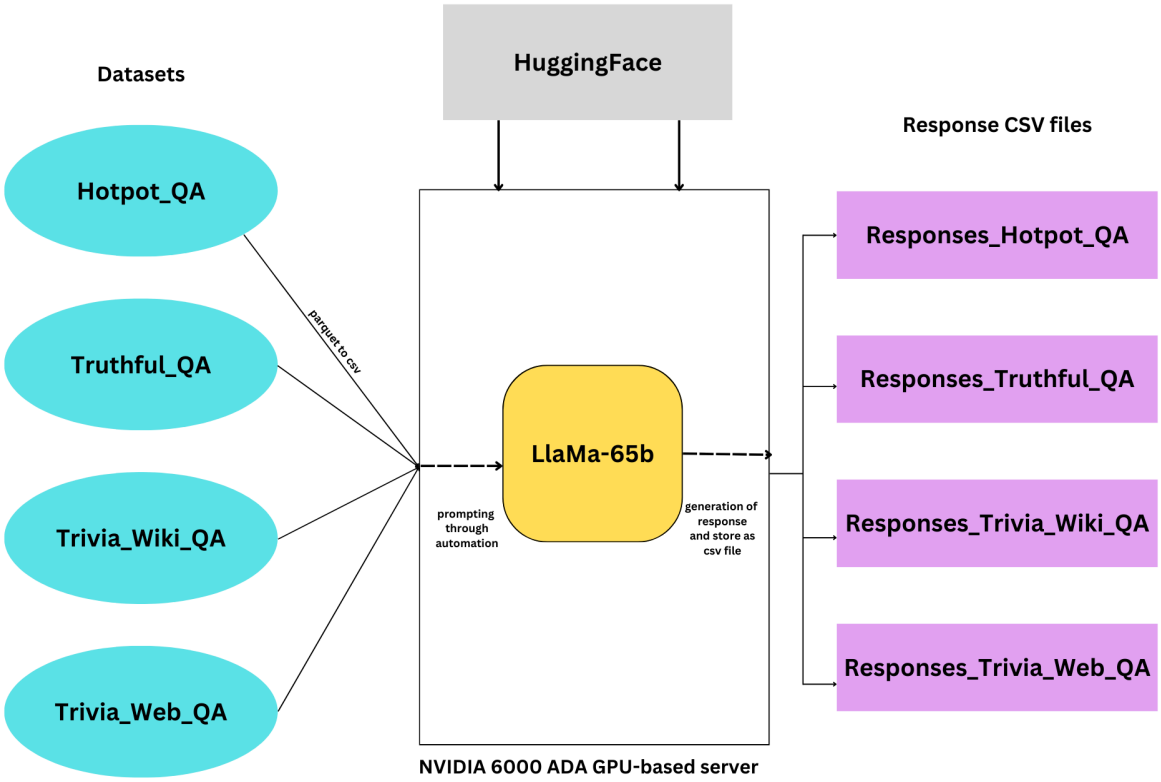
We chose 4 datasets, namely HotpotQA, TruthfulQA, TriviaWebQA and TriviaWikiQA comprise the examples and documents used to test the model, thus determining the specific patterns it exhibits in generating the response. To evaluate the LLama 65B model's performance and its ability to handle complex queries, we did not train or fine-tune the LLama 65B; instead, the datasets served as benchmarks to assess the model's response patterns using calculated metrics such as the edit distance and cosine similarity, and human evaluation.

Dataset Selection:

Dataset Name	Purpose
Hotpot QA	HotpotQA presents Wikipedia-based questions requiring analysis of multiple documents to enhance reasoning and accuracy. It introduces novel comparison questions for evaluating fact extraction and comparison capabilities.
TruthfulQA	The dataset contains questions designed to probe areas where common misconceptions or misinformation might exist, spanning various categories including health, law, finance, and politics. The questions also require the LLM to cite specific sources to back its claims
TriviaWeb	The dataset consists of web-sourced question-answer pairs from various sources.
TriviaWiki	TriviaQA Wiki dataset consists of the question-answer generated from wikipedia, which is a subset of Trivia QA dataset. It consists of 7.99K questions to help verify the factual correctness of response generated as per wikipedia.

Model Output:

Initially, we installed the mentioned models from Hugging Face onto our server. While the initial installation of the models may take some time, subsequent runs utilize a checkpoint of shards for efficiency. Next, we loaded the dataset onto the server. Subsequently, we fed the dataset into the models to generate the output. The entire process is depicted in the diagram below.



We've sampled 100 questions from each dataset and inputted these prompts into the Llama 65b model to generate responses. We then assessed the degree of hallucination in each response compared to the datasets using various performance metrics for analysis- like edit distance, cosine similarity and human evaluation.

Results

Human Evaluation

When Llama 65B was evaluated using human inference against the TriviaWeb, TriviaWiki, TruthfulQA & hotpotQA, the models were scored on how well the models answered compared to the best answer based on our defined scoring levels as shown

in the figures below. After scoring all the models for 100 questions, we calculated the score percentage of each model when evaluated with the four datasets. The scoring scheme of different benchmarks is given below as follows.

Model Output Type	**Score**
Completely Hallucinated	0
Somewhat related to the question	1
Answered correctly to some extent	2
Very close to the correct answer	3

The scores for the datasets are as shown :

Score	**TriviaWiki**	**TriviaWeb**	**TruthfulQA**	**HotpotQA**
0	14	24	15	38
1	2	4	4	13
2	1	11	2	2
3	83	61	79	47

TriviaWiki dataset had a high rate of accuracy, with 83 out of 100 answers being very close to the correct answers, indicating a strong performance by the model on this dataset.

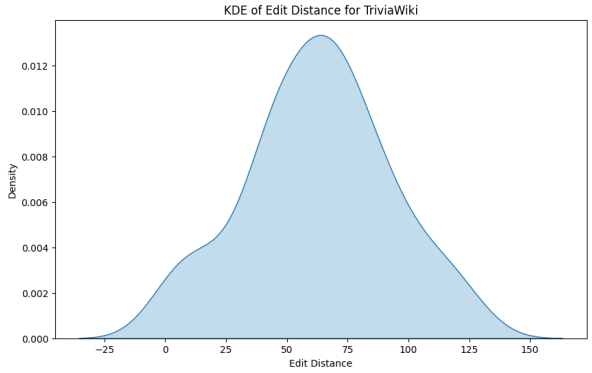
TriviaWeb, while generally adequate with 61% high accuracy responses, exhibited considerable inconsistency with 24% of answers completely unrelated to the questions.

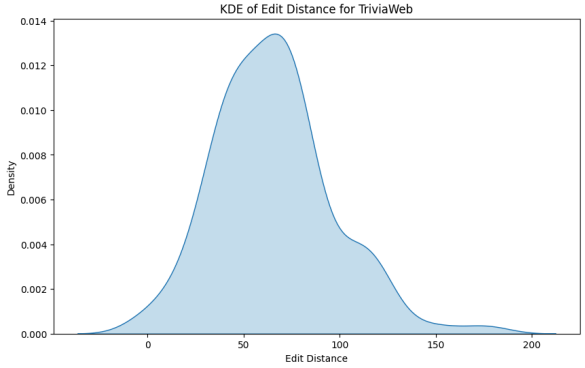
For TruthfulQA, the dataset demonstrates a substantial accuracy with 79 out of 100 answers being very close to the correct answers. We will see later on that there is variability between edit distance and other metrics for TruthfulQA.

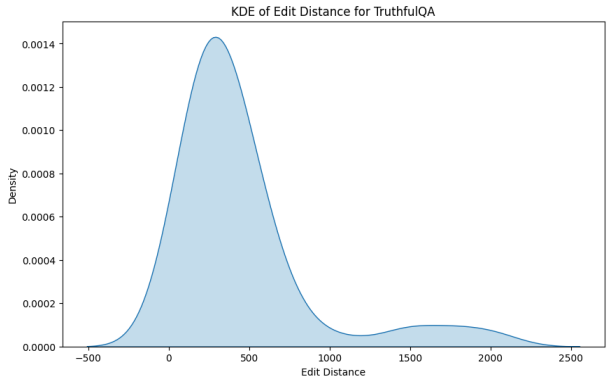
The HotpotQA dataset displayed more variability in the quality of responses. Although nearly half of the responses were very close to the correct answers, a significant portion (38%) were completely unrelated to the expected answers.

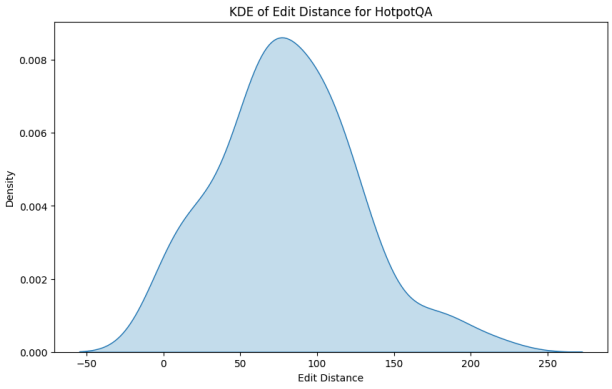
Edit Distance

Edit distance, which quantifies the number of edits required to convert the model's response into the ground truth, serves as one robust measure of textual similarity. A detailed statistical analysis and the distribution of these edit distances will be visualized through Kernel Density Estimation (KDE) plots in the subsequent section. This metric is particularly valuable in determining the precise level of adjustment needed for a model's output to align with verified data, as it values syntactic accuracy.









Edit Distance	TriviaWiki	TriviaWeb	TruthfulQA	HotpotQA
Mean	62.86	65.55	447.91	80.86
75th Percentile	83.00	79.00	521.00	108.50
50th Percentile	64.00	64.50	316.50	77.00
25th Percentile	44.00	43.00	213.50	52.25

Standard Deviation	29.62	30.48	432.35	45.68

The TriviaWiki dataset demonstrates relatively moderate variability in edit distances with a mean of 62.86 and a standard deviation of 29.62. The median value is 64, closely aligning with the mean, suggesting a symmetric distribution around this central value. The range from the minimum (0) to the maximum (128) edit distance highlights some cases of perfect matches and significant mismatches in the answers generated by LLAMA65b.

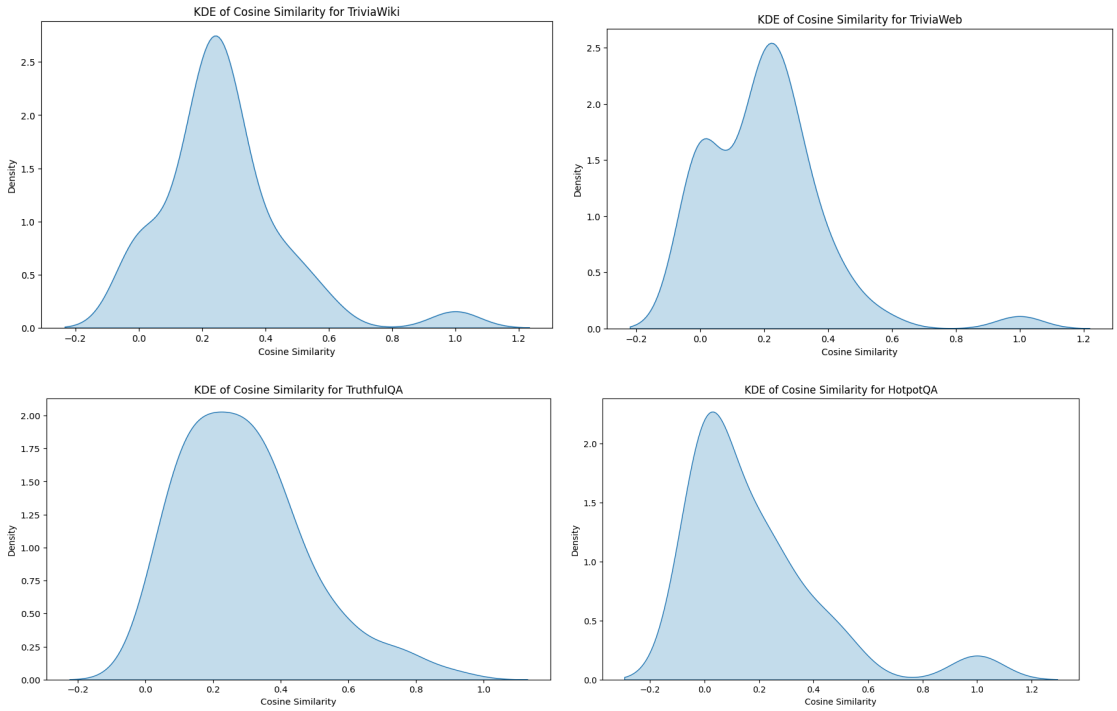
TriviaWeb shows a similar central tendency with a mean edit distance of 65.55, which is slightly higher than TriviaWiki, and a standard deviation of 30.48, indicating comparable variability in answer accuracy. The median is slightly higher at 64.5, with a narrower interquartile range from 43 to 79, suggesting that most answers cluster more closely around the median compared to TriviaWiki.

The TruthfulQA dataset starkly differs in its performance metrics, with a much higher mean edit distance of 447.91 and a very high standard deviation of 432.35. This indicates extremely large discrepancies between the generated answers and the correct answers, with a significant spread in the data. The median of 316.5 is substantially higher than the other datasets, and the maximum edit distance of 2038 points to some answers being vastly different from what was expected.

HotpotQA also stands out with a higher mean of 80.86 and a larger standard deviation of 45.68, which indicates greater textual divergence from the expected answers compared to TriviaWeb and TriviaWiki datasets. The data spread from a minimum of 0 to a maximum of 218, with the third quartile at 108.5, further highlights the variability and often substantial discrepancies between the generated and expected textual responses.

Cosine Similarity

Cosine similarity was utilized to measure the semantic proximity between LLAMA65b's responses and the ground truths. This metric, unlike edit distance, considers the angle between vector representations of text, providing insights into the semantic alignment of responses.



Cosine Similarity	TriviaWiki	TriviaWeb	TruthfulQA	HotpotQA
Mean	0.27	0.20	0.29	0.18
75th Percentile	0.30	0.28	0.39	0.25
50th Percentile	0.24	0.20	0.27	0.08
25th Percentile	0.18	0	0.15	0
Standard Deviation	0.20	0.18	0.19	0.25

The TriviaWiki dataset shows a mean cosine similarity of 0.2657 with a standard deviation of 0.1956, indicating moderate similarity on average between generated and expected answers. The distribution is relatively tight, with 75% of the data falling below a cosine similarity of 0.3032.

The TriviaWeb in comparison, exhibits a lower average cosine similarity of 0.2017 with a similar standard deviation to TriviaWiki (0.1834), suggesting less consistency in the similarity of responses. A noteworthy aspect is the lower quartile value at 0, which implies that 25% of the answers had very little to no similarity with the expected answers.

The TruthfulQA dataset contrasts from both TriviaWiki and TriviaWeb in terms of higher mean cosine similarity at 0.2911 and a slightly lower standard deviation of 0.1882, suggesting more consistent higher similarity scores. The median at 0.2682 is comparable to TriviaWiki, but the upper quartile reaches 0.3930, indicating a better overall alignment in the top 25% of responses compared to the other datasets.

The HotpotQA dataset shows a mean cosine similarity of 0.1794 with a standard deviation of 0.2467. A median of only 0.0799, much lower than any other dataset, reflects that more than half of the responses have weak semantic similarities. The 75th percentile at 0.2537 signifies that only the top quartile achieves only adequate levels of similarity to ground truth, underscoring the challenges the model faces in dealing with the complex question structures typically found in HotpotQA.

Calculated Metrics in Comparison With Each Other

Edit distance is highly sensitive to literal changes and is a harsh metric for textual divergence. In contrast, cosine similarity can reveal underlying thematic or contextual similarities even if the words or phrases used are quite different. Initially, high edit distances in the TruthfulQA dataset suggest a textual mismatch that might be misleading if considered alone, as the higher cosine similarities indicate that the generated answers are not entirely off the mark in terms of the broader context.

The combined analysis of edit distance and cosine similarity across these datasets reveals that while LLAMA65b often captures the conceptual underpinnings of the answers (as indicated by cosine similarities), it struggles with exact textual replication in more complex scenarios (as shown by edit distances).

Calculated Metrics in Comparison with Human Evaluation

Llama65b produced generally relevant and textually close responses as assessed by both computational and human metrics for TriviaWiki, as cosine similarity,, edit distance and human evaluated scores all suggest reasonable levels of semantic similarity between the LLM’s response and the ground truth. TriviaWeb dataset displayed relatively lower levels of performance compared to TruthfulQA in terms of cosine

similarity, however was more favored in terms of human evaluation and edit distance scores. The TriviaWiki dataset also consistently showed better results than TriviaWeb. Along the same lines, the HotpotQA dataset yielded relatively weaker scores across all metrics.

For the TriviaWiki dataset, there is a strong correlation between human judgments and computational metrics, where high human evaluation scores are matched with favorable edit distances and cosine similarities. This suggests that where the model performs well textually and semantically, it also aligns closely with human expectations of accuracy and relevance.

The HotpotQA and TruthfulQA datasets exhibit significant discrepancies between human and computational metrics. The TruthfulQA dataset shows high human evaluation scores despite large edit distances, suggesting that while the model's responses are contextually aligned with the ground truth, they may diverge significantly in terms of exact wording or detail. This underscores the potential limitations of edit distance as a sole metric for evaluating model performance.

Cosine similarity, especially in datasets with complex semantics like TruthfulQA, tends to align better with human judgment regarding the contextual accuracy of the model’s responses.

Conclusion

In conclusion, while the LLama65B model exhibits proficiency in capturing contextual accuracy, it demonstrates challenges with exact syntactic replication, often generating syntactically imprecise responses despite correct contextual understanding. Notably, the limitations of using edit distance as a sole metric for model evaluation became evident, highlighting the need for a multifaceted assessment approach. The performance of these datasets also underscore the challenges LLMs face with complex question structures and diverse factual content.

Future Work

Looking ahead, future research should explore a broader array of models, potentially including newer or differently trained LLMs, to enhance understanding and capabilities. Focusing on datasets that concentrate on specific themes or topics might yield more profound insights into specialized performance areas. Additionally, expanding this research to encompass multilingual datasets would offer valuable perspectives on the performance of LLMs across different linguistic contexts, broadening the scope and applicability of language model evaluations.

References:

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. ArXiv. /abs/2005.14165v4

Feng, S., Shi, W., Wang, Y., Ding, W., Balachandran, V., & Tsvetkov, Y. (2024). Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. *ArXiv*. /abs/2402.00367

Chiang, Y., Chou, C., & Riebesell, J. (2024). LLaMP: Large Language Model Made Powerful for High-fidelity Materials Knowledge Retrieval and Distillation. *ArXiv*. /abs/2401.17244

Łajewska, W., & Balog, K. (2024). Towards Reliable and Factual Response Generation: Detecting Unanswerable Questions in Information-Seeking Conversations. *ArXiv*. /abs/2401.11452

Rawte, V., Priya, P., Tonmoy, S. M., Zaman, S. M., Sheth, A., & Das, A. (2023). Exploring the Relationship between LLM Hallucinations and Prompt Linguistic Nuances: Readability, Formality, and Concreteness. *ArXiv*. /abs/2309.11064

Zhang, Y., Cui, L., Bi, W., & Shi, S. (2023). Alleviating Hallucinations of Large Language Models through Induced Hallucinations. *ArXiv*. /abs/2312.15710

Semnani, S. J., Yao, V. Z., Zhang, H. C., & Lam, M. S. (2023). WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia. *ArXiv*. /abs/2305.14292

A study titled "Rethinking Search: Making Experts out of Dilettantes" by Komeili et al. (2021) discusses how LLMs like GPT-3 are enhanced with retrieval-augmented techniques to improve knowledge retrieval and response accuracy in domain-specific contexts ar5iv.labs.arxiv.org/html/2306.05212

Chowdhery et al. (2022) detail the performance improvements in conversational LLMs through advancements in models like GPT-3.5, highlighting how these models handle natural language generation from factual data with increased efficiency and reduced errors ar5iv.labs.arxiv.org/html/2401.12599