*Note: Citations are missing in this proposal.*

# How does tokenization affect LLM hallucinations?


## Saaketh Raghava, Computer Science and Engineering, UC Merced



## Abstract

Large Language Models (LLMs) have emerged as powerful tools but can suffer from generating hallucinations. Hallucinations are a major hurdle in LLM adoption for safety critical tasks. This research investigates how different tokenization methods influence the frequency and type of hallucinations in LLMs. We use the LLama2 [citation missing] model and evaluate its performance on the HaluEval and TruthfulQA datasets under various tokenization schemes. The HaluEval dataset [citation missing] allows us to assess the model's ability to recognize hallucinations and categorize them based on established classifications. The TruthfulQA dataset enables us to measure how tokenization impacts the factual accuracy of generated text. This project aims to describe the relationship between tokenization and hallucinations in LLMs.


## Introduction and Statement of Need:

Generative AI has exploded throughout the world through products like LLMs and DALL-E. Generative AI models have shown remarkable promise in automating previously thought-impossible tasks. A survey of over 500 corporate leaders shows that 67% are prioritizing gen AI for their business within the next 18 months. However, there is a hesitation to adopt LLMs [citation missing] into many domains such as healthcare due to issues in LLMs like hallucinations. Hallucinations are one of the biggest obstacles in the deployment of LLMs in the field. Tokenization is the process of breaking down smaller units of text into tokens that can be understood by the model. However, there are many types of tokens such as character tokenization, word tokenization, phrase tokenization, sentence tokenization, byte pair encoding, unigram. Deciding which tokenization type to use depends on the use case of the LLM. Tokenization is extremely important [citation missing] for many NLP applications such as information retrieval in search engines, text preparation, sentiment analysis, and chatbots. Tokenization can have a significant impact on an LLM, but does that impact extend towards hallucinations? This research is essential to ensure the safe development of Generative AI models in the future. This project will use different types of tokenization on the Llama2 model and measure them against the HaluEval and TruthfulQA Datasets.

## Research Objectives and Plan

### Primary Objective:

The problem that I will be addressing is to investigate how different tokenization methods (sentence, word, phrase, character, and subword tokenizations like byte pair encoding and unigram) impact the frequency and type of hallucinations in llms.

### Methodology:

The HaluEval dataset is a large-scale benchmark for evaluating hallucinations in large language models. It contains over 35,000 samples across three tasks: question answering, knowledge-grounded dialogue, and text summarization. The dataset includes both automatically generated hallucinated samples as well as human annotated responses from the ChatGPT model. The samples cover a diverse range of topics from movies to climates. The HaluEval dataset is well-suited for this project because it provides a testbed for evaluating model performance in recognizing hallucinations. Furthermore, the many different

categories in hallucinations (i.e., comprehension, factualness, specificity, inference, extrinsic-soft, extrinsic hard, extrinsic-grouped, factual, non-factual, and intrinsic) (Zheng et al., 2023; Das et al., 2023; Cao et al., 2022) For my research, I plan to use Llama 2 model. (Touvron, et al., 2023) Llama 2 was chosen due to its customizability compared to other models like Claude 3 and GPT-4. Llama2 also works with the transformers library by Hugging Face which will allow me to specify tokenizers during inference. I will evaluate the model's performance on the HaluEval dataset under different tokenization schemes. This will involve experimenting with various tokenization approaches, such as word-level, sub word-level, and character-level tokenization. For each tokenization scheme, I will assess the model's ability to recognize hallucinations in the HaluEval samples. We will also test the types of hallucinations the model fails to recognize and then categorize them based on the different categories of hallucinations. Furthermore, I will assess the model on the Truthful QA Dataset. (Lin et al., 2022) The Truthful QA Dataset was picked because it consists of over 800 questions spanning various categories like health, law, finance, and politics. The questions are crafted to have verifiable answers which provides a standardized way to measure hallucinations. Using the llama 2 model, we can assess how different tokenization methods impact the factual accuracy of the generated text.

# Midterm and Final Goals:

## Midterm Goals:

- Review the current state of research into tokenization and LLM hallucinations
- Environment Setup: Set up the software environment of Llama2
- Pilot Testing: Conduct a pilot experiment
- Define the Evaluation Metrics: Define clear metrics for measuring hallucination rates and types.

## Final Goals (last XX weeks)

- Run the experiments on the complete HaluEval and TruthfulQA dataset using the chosen tokenization methods.
- Analyze the data to identify any statistically significant differences in hallucination rates, types, and factual accuracy between different tokenization methods

## Future Research

- Explore multiple models like GPT 3 and Mistral Large and see how they compare.
- Finetune the model to a specific safety critical area and see if the results are different

# References:

March 6, & 2023. (2023, March 6). IT Leaders Call Generative AI a "Game Changer" but Seek Progress on Ethics and Trust. Salesforce. https://www.salesforce.com/news/stories/generative-ai-research/

Ahmad, M., Yaramis, I., & Dutta, T. (n.d.). Creating Trustworthy LLMs: Dealing with Hallucinations in Healthcare AI. https://arxiv.org/pdf/2311.01463.pdf

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in answering questions faithfully? CoRR, abs/2304.10513.

Souvik Das, Sougata Saha, and Rohini K Srihari. 2023. Diving deep into modes of fact hallucinations in dialogue systems. arXiv preprint arXiv:2301.04449

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3340–3354.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods

Jinbiao Yang. 2024. Rethinking Tokenization: Crafting Better Tokenizers for Large Language Models

Llama 2: Open Foundation and Fine-Tuned Chat Models | Meta AI Research. (n.d.). Ai.meta.com. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/