Hallucination Signatures in Multilingual LLMs

Nishanth Mitta Sathish nmittasathis@umass.edu
Samruddhi Bhabad sbhabad@umass.edu
Kathiresan Palaniappan kpalaniappan@umass.edu

1 Introduction

Large Language Models (LLMs) such as GPT-3.5 and mT5 have achieved remarkable multilingual capabilities, yet they frequently *hallucinate* (produce fluent but factually incorrect information.) While hallucination behavior is well-documented in English, little is known about how it varies across languages. Multilingual LLMs are trained on uneven data distributions, and the imbalance in linguistic resources may affect both the rate and type of factual hallucinations they generate.

This project investigates whether hallucinations differ systematically by language. We hypothesize that language resource availability and linguistic structure shape hallucination patterns, influencing both their frequency and qualitative signatures. For example, models might exhibit higher hallucination rates in low-resource languages due to weaker factual grounding or sparse entity exposure.

Our study also explores whether hallucinations carry **linguistic precursors**—signals such as rare entity mentions, excessive specificity, or abrupt confidence spikes—that can be used to predict false outputs. Finally, we examine whether models can **self-correct**, either preemptively (through confidence calibration) or post-generation (via retrieval-augmented verification).

Research Questions:

- **Primary RQ:** Do hallucination rates and types systematically differ by language for multilingual LLMs?
- **Secondary RQ1:** Are certain hallucination types (invented vs. misremembered) more common in low-resource languages?
- **Secondary RQ2:** Can confidence-based or self-retrieval methods predict or reduce hallucinations, and does that reliability vary across languages?

This investigation aims to deepen our understanding of factual reliability in multilingual NLP systems and build toward language-robust hallucination detection and mitigation tools.

2 Related work

Recent surveys such as Ji et al. (Ji et al., 2023) and Zhao et al. (Zhao et al., 2023) have outlined the growing concern around hallucinations in text generation systems. They emphasize that hallucinations often arise from model overconfidence, incomplete factual grounding, or representational bias. Bang et al. (Bang et al., 2023) conducted a multilingual evaluation of ChatGPT across 40 languages and observed large accuracy drops in low-resource languages, indirectly suggesting higher hallucination potential.

A major step forward in multilingual hallucination analysis was made by Islam et al. (Islam et al., 2024), who carried out the first large-scale evaluation of hallucinations across 30 languages and 11 large language models. They introduced the MFAVA-Silver dataset, created by automatically inserting different types of hallucinations into factual data and translating them into multiple languages. Importantly, they validated this synthetic approach by showing a strong correlation between their automatically generated (silver) and human-verified (gold) annotations, confirming that synthetic multilingual data can reliably capture real hallucination patterns. This finding supports our proposed data creation method, where we also generate and translate controlled hallucination examples from Wikipedia-based sentences.

Islam et al. (Islam et al., 2024) also proposed a quantitative framework for estimating true hallucination rates by adjusting for a detector's precision and recall, which provides a solid foundation for measuring hallucination frequency more accurately. Their analysis revealed several key trends: smaller models hallucinate more, models that support more languages tend to hallucinate more often, and surprisingly, the overall hallucination rate does not strongly depend on how resource-rich a language is. These results motivate our focus on Indian languages, as we aim to examine whether such trends hold within a typologically related family of languages that vary in resource availability.

Cross-lingual transfer studies by Xu et al. (Xu et al., 2023) highlight that models perform inconsistently when knowledge must be transferred across typologically distant languages. Similarly, Shuster et al. (Shuster et al., 2022) point out that unbalanced pretraining and tokenization biases can cause "confabulations," where models fabricate language-specific facts.

While these works have focused mainly on detection and estimation, recent methods like Manakul et al. (Manakul et al., 2023) explore **self-verification** pipelines, where models assess their own outputs for factual correctness. However, these methods have mostly been tested in English. Our work extends beyond detection by combining insights from Islam et al. (Islam et al., 2024) with self-retrieval and confidence calibration experiments, aiming to understand not only *how* hallucinations differ across languages, but also *whether* models can identify and correct their own false outputs. Islam et al. (Islam et al., 2024) also conducted one of the largest multilingual hallucination studies, comparing 11 models across 30 languages. They found that hallucination patterns depend not only on model size but also on multilingual coverage and tokenization quality. Interestingly, they observed that low-resource languages do not always show higher hallucination rates, suggesting that other linguistic and architectural factors may play a role. These results motivate our focus on Indian languages as a coherent family for deeper analysis.

3 Our approach

We design a multilingual hallucination evaluation pipeline combining factual QA benchmarking, custom dataset, hallucination signature analysis, and self-correction experiments.

3.1 Cross-lingual Benchmarking

We will construct factual prompts across English, Hindi, Kannada, Marathi, Tamil, and other Indian languages using vanilla or translated aligned question-answer datasets (MKQA, XQuAD, TyDi QA, MMLU). Each model, mT5, mmT5, GPT-3.5 multilingual, and optionally multilingual LLaMA or Qwen models, will be queried with equivalent prompts. Responses will be judged for factual correctness via both human review and evidence-based retrieval from Wikipedia or Wikidata.

3.2 Self-RAG and Confidence Calibration

We will evaluate whether models can predict or correct their own hallucinations:

- **Pre-generation detection:** Using model confidence scores and calibration curves to flag potentially unreliable responses before they are output.
- **Post-generation self-RAG:** Using a retrieval-augmented pass where the model re-queries Wikipedia or MKQA evidence to validate or revise its earlier answer.

We aim to measure how effective these mechanisms are across languages and whether self-retrieval reduces hallucination rates more in high-resource or low-resource languages.

3.3 RAG-Fusion for Hallucination Mitigation

We will implement **RAG-Fusion**, an enhanced retrieval-augmented generation (RAG) framework designed to mitigate hallucinations through multi-query retrieval and intelligent document reranking. Unlike standard RAG pipelines that rely on a single query, RAG-Fusion generates multiple semantically diverse query variants for each input. These variants are used to perform parallel vector searches, capturing a broader range of relevant evidence. Retrieved documents are then combined using *Reciprocal Rank Fusion (RRF)*, which prioritizes sources that consistently rank highly across multiple query formulations.

This multi-perspective retrieval process ensures that the model accesses more complete and cross-validated contextual information before generation, reducing the likelihood of unsupported or fabricated claims. In our setup, RAG-Fusion will operate as part of the **post-generation verification pipeline**: after a model outputs an answer in a given language, we will generate several verification queries in both the source and target languages, retrieve evidence from language-specific Wikipedia dumps, and use the fused, reranked results to verify or flag potential factual inconsistencies.

We hypothesize that RAG-Fusion's ability to highlight documents that appear consistently across diverse query perspectives will improve factual grounding, particularly in low-resource or linguistically diverse settings. We will compare its performance against traditional single-query RAG systems to evaluate gains in cross-lingual hallucination detection and mitigation.

3.4 Baselines

To evaluate the effectiveness of our proposed self-retrieval and calibration methods, we will compare them against several meaningful baselines. These baselines are designed to represent simple or widely used approaches to multilingual factual QA, providing a reference point for improvement.

- **Frequency baseline:** Predicting the most common or majority answer per question, regardless of language or context. This helps measure how much performance gain comes from actual model reasoning rather than statistical bias.

- **Translation pipeline baseline:** Translating all queries to English, generating answers using a strong English-only model, and translating the responses back into the target language. This baseline tests whether multilingual reasoning adds value beyond a purely English-centric workflow and reveals how translation quality influences factual accuracy.

- **Confidence-only baseline:** Applying a simple confidence threshold to decide whether a model's answer should be accepted or flagged as unreliable, without any retrieval or verification step. This serves as a minimal self-assessment method and highlights the additional benefit of retrieval-augmented self-correction.

3.5 Schedule

1. Literature review and dataset alignment (1 week)

2. Model setup, prompt translation, and preprocessing (2 weeks)

3. Response generation and hallucination annotation (2 weeks)

4. Self-RAG implementation and calibration analysis (2 weeks)

5. Final evaluation, visualization, and report writing (1 week)

4 Data

We will be using two kinds of datasets:

4.1 Manually curated dataset

Inspired by the methods proposed by Abdaljalil et al. (Abdaljalil et al., 2025), we have decided to curate our own dataset. We will query facts from biographies on Wikipedia in English and introduce hallucinations of three different kinds, namely entity-level, relation, and sentence-level hallucinations. We will then translate these hallucination-ridden sentences to other languages. As a control, we shall also posses hallucination-less sentences.

4.2 Publicly available multilingual QA datasets

We will use publicly available multilingual QA datasets suitable for factual recall:

- **MKQA (Longpre et al., 2021):** 26-language aligned factual QA benchmark.

- **TyDi QA (Clark et al., 2020):** Diverse typologically representative QA dataset.

- **XQuAD/XQuAD-R (Artetxe et al., 2019):** Parallel SQuAD-derived multilingual datasets.

- **MMLU (Hendrycks et al., 2020):** consists of 15,908 multiple-choice questions, spanning across 57 subjects, from highly complex STEM fields and international law to nutrition and religion.

Language-specific Wikipedia dumps will serve as retrieval sources for self-RAG experiments. All data are open-access and suitable for academic use.

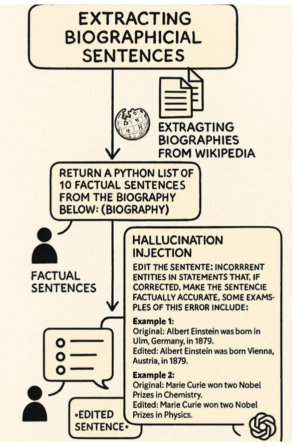
EXTRACTING BIOGRAPHICIAL SENTENCES
EXTRAGTING BIOGRAPHIES FROM WIKIPEDIA
RETURN A PYTHON LIST OF 10 FACTUAL SENTENCES FROM THE BIOGRAPHY BELOW: (BIOGRAPHY)
FACTUAL SENTENCES
HALLUCINATION INJECTION
EDIT THE SENTENTE: INCORRRENT ENTITIES IN STATEMENTS THAT, IF CORRECTED, MAKE THE SENTENCIE FACTUALLY ACCURATE, SOME EXAMS-PLES OF THIS ERROR INCLUDE:
Example 1:
Original: Albert Einstein was bom in Ulm, Germany, in 1879.
Edited: Albert Einstein was born Vienna, Austria, in 1879.
Example 2:
Original: Marie Curie won two Nobel Prizes in Chemistry.
Edited: Marie Curie won two Nobel Prizes in Physics.
EDITED SENTENCE

Figure 1: Manual data creation methodology

5 Tools

We will use `HuggingFace Transformers` for model access, `LangChain` for retrieval and prompting workflows, and `evaluate` for factuality scoring. Model experiments will run on Google Colab Pro GPUs. GPT-3.5 multilingual or open-weight multilingual LLaMA models will be accessed via available APIs. For text preprocessing, we will use `spaCy`, `langdetect`, and `deep-translator`. Analysis and visualization will be performed using Python libraries such as `pandas` and `matplotlib`.

6 AI Disclosure

- We used ChatGPT (GPT-5) for formatting, language refinement, and section organization based on our original notes.
- Despite the refinement offered by ChatGPT, we made changes to the proposed text to better fit our writing style.

References

Abdaljalil, S. et al. (2025). HalluVerse25: Fine-grained multilingual benchmark dataset for llm hallucinations.

Artetxe, M., Ruder, S., and Yogatama, D. (2019). On the cross-lingual transferability of monolingual representations.

Bang, Y., Cahyawijaya, S., Lee, N., Su, D., Xu, Y., et al. (2023). A multilingual evaluation of chatgpt.

Clark, J. H. et al. (2020). TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring massive multitask language understanding.

Islam, M., Rahman, M., and Ahmad, W. (2024). Multilingual hallucination evaluation in large language models.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Longpre, S., Lu, Y., and Daiber, J. (2021). MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Manakul, P., Laban, P., and Augenstein, I. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Shuster, K., Ju, D., Roller, S., Dinan, E., and Weston, J. (2022). Language models that seek for knowledge: Modular search and generation for dialogue and qa.

Xu, H., Liu, Y., Wang, Y., and Xue, N. (2023). Cross-lingual transfer in multilingual transformers: An empirical study. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhao, W., Li, J., Wang, R., Xu, J., and Lin, C. (2023). Hallucination in large language models: A survey.