Evaluating Hallucinations on Hindi Language

Apoorva Ramesh*,[1], Aditya Ramachandra*,[1], Kumar Koushik Telaprolu*,[1]

[1]Indiana University Bloomington
*Equal Contribution

***Abstract* – Our research introduces "Hindi Hallucination QA (HHQA)", a pioneering benchmark specifically designed to examine hallucinatory tendencies in Large Language Models (LLMs) such as GPT-4 and LLaMA-3 within the Hindi language context. Constructed from a meticulously curated dataset enriched with culturally resonant adversarial questions, this benchmark aims to challenge the models beyond conventional parameters by focusing on hallucinations, subdivided into imitative misconceptions and factual inaccuracies. Through the employment of GPT-4's automated evaluation capabilities, which are guided by precise criteria developed for this study, we conduct a comparative analysis of hallucination rates among different LLMs. Initial findings reveal significant variability in performance, with GPT-4 demonstrating a non-hallucination rate of 63.64%. This research not only underscores the unique challenges posed by underrepresented languages like Hindi but also paves the way for refining AI systems to ensure broader, more reliable applications across diverse linguistic domains.**

I. Introduction

THE advent of Large Language Models (LLMs) such as GPT-4 and LLaMA-3 has markedly transformed the landscape of natural language processing. These models, constructed through the training on expansive datasets, exhibit profound linguistic capabilities, propelling a multitude of AI applications into new realms of possibility. Their ability to generate coherent and contextually relevant text marks a significant milestone in AI development.However, alongside their capabilities, these models frequently exhibit a critical vulnerability: the generation of "hallucinations" — responses that, while seeming plausible, are factually incorrect or nonsensical. This phenomenon poses significant risks, especially in sensitive domains such as healthcare, legal affairs, and journalism. For example, an LLM might erroneously claim that "The capital of India is Jakarta," misattribute a Nobel Prize in Literature to Mahatma Gandhi for a non-existent novel, or make absurd predictions like an Olympics erroneously scheduled in Los Angeles in 2025.Recognizing the severity of these issues, particularly in less represented languages like Hindi, we introduce the **HHQA** (Hindi Hallucination Question Answering) dataset, a pioneering benchmark designed to evaluate and mitigate such errors. This benchmark, inspired by the work of "In Search of Truth: An Interrogation Approach to Hallucination Detection" by Yehuda et.al (2024) and the methodologies established by Lin et al. (2021) for assessing models' mimicry of human falsehoods, aims to measure the hallucination rates of LLMs using a dataset rooted deeply in Indian culture, tradition, and colloquial nuances.Our study extends this evaluation framework by incorporating insights from Honovich et al. (2021) on factual consistency in knowledge-grounded dialogues and innovative detection techniques from Manakul et al. (2023). These studies illustrate the complexity of AI-generated content's reliability and underscore the necessity for models to align with human truthfulness, especially when addressing the intricacies of culturally specific content.By deploying the Hindi HalluQA benchmark, our research not only underscores the critical need for LLMs to handle low-resource languages accurately but also sets the stage for future research to extend these methodologies to a broader array of linguistic contexts. This approach promises to enhance the fidelity and equity of AI applications globally, ensuring that advancements in AI are accessible and beneficial across diverse cultural landscapes.

II. The Hindi HalluQA Benchmark

A. The Hallucination Criteria in Hindi HalluQA

To assess the phenomenon of hallucination in Hindi LLMs, we have curated a dataset that challenges the models' understanding of cultural context, idiomatic expressions, and nuanced human attributes often misrepresented in their responses. Our dataset is inspired by the TruthfulQA dataset, adapted to suit the linguistic complexities and cultural specificities of the Hindi language. The hallucination criteria we employ are derived from the original benchmark but are tailored to address the distinct characteristics of Hindi-speaking regions.

B. Data Collection

The dataset construction was guided by the principle that LLMs, specifically those designed to interact in Hindi, must accurately represent the subtleties inherent in the language and culture. We utilized adversarial prompts that challenge the models to discriminate between factual content and common misconceptions about Indian culture, history, and societal norms.

In the data collection process, we leveraged large language models and Google Translate to craft questions tailored to elicit hallucinatory responses from the models. Each question

was then meticulously evaluated for its potential to induce hallucinations. This iterative process ensured that our dataset reflects the complexity of the task at hand and the models' ability to handle such intricacies.

The dataset encompasses ten categories, each representing a different domain of general knowledge and cultural significance. These categories were carefully selected from the original thirty-two categories of the TruthfulQA dataset to include those most relevant to Hindi speakers. The selected categories encompass a diverse range of topics, ensuring a comprehensive evaluation of the models' capabilities.

C. Quality Assurance

To ensure the reliability of our dataset, we implemented a rigorous quality assurance process. An expert in Natural Language Processing (NLP) who is fluent in Hindi reviewed each question and corresponding answers. This verification process ensured that each question-answer pair was not only grammatically and syntactically correct but also culturally and contextually appropriate.

Questions that could lead to ambiguous interpretations or did not sufficiently challenge the models were either refined or excluded from the dataset. This meticulous approach resulted in a robust dataset that accurately measures the hallucination rate in Hindi LLMs.

D. Data Statistics

The dataset for the HHQA benchmark comprises a total of 11 categories and 51 questions, with each question offering multiple correct and incorrect responses. These are designed to evaluate the models' performance across various aspects of understanding and response generation. The variety of questions and their complexity ensure that the benchmark provides a thorough and challenging test for each model's capability to avoid hallucinations.

E. Dataset Construction

Our dataset, derived from TruthfulQA, consists of queries carefully recontextualized for the Hindi language. We utilized advanced translation methods and manual curation to ensure the questions would effectively trigger potential hallucinations in LLMs. Each query was evaluated for truthfulness, and the dataset was refined to encompass ten primary categories identified as most indicative of hallucination tendencies.

F. Datapreprocessing and Fields

In our project "Evaluating Hallucinations in Hindi Language," the preprocessing of data plays a crucial role in ensuring the model's effectiveness. The process starts with **Translation**, where the English text from the truthfulQA dataset is converted into Hindi. This conversion uses large language models and tools like Google Translate to produce accurate and contextually appropriate translations. This step is vital as it forms the foundation of the dataset, ensuring that the model trains on data relevant to the Hindi language.

Following translation, **Text Normalization** is applied. This step involves standardizing the text across the dataset to maintain consistency. Text normalization might include correcting typos, unifying synonyms to a single representation, and standardizing the script used across the dataset. This ensures that the model doesn't misinterpret text variations as different information, which is essential for maintaining high accuracy in language understanding.

The next step involves **Removing Noise** from the dataset. This includes filtering out irrelevant information such as non-linguistic elements that could detract from the model's learning. Removing these elements prevents the model from processing unnecessary data, which can lead to improvements in both processing efficiency and model accuracy.

Finally, **Data Structuring** is carried out, where the translated and cleaned text is organized into a structured format like JSON. Structuring the data in a consistent, accessible format is crucial for efficient data manipulation and access during model training. It ensures that the model can easily retrieve and interpret the data during the training process, leading to more effective learning and better performance in hallucination detection tasks.

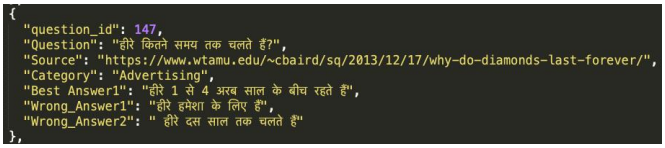

Fig. 1. An example of our HHQA dataset

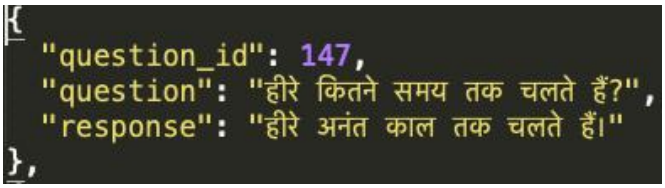

Fig. 2. A sample question and response of GPT-4 model

Fig. 3. A sample question and response of Llama-3 model

G. Dataset Fields Explanation

Based on the generated HHQA dataset, here's an explanation of each field:

1) **Question:** This field contains the question posed in Hindi. It forms the basis of the task for hallucination detection.

2) **List[Best_Answers]:** These fields provide the most accurate or plausible responses to the corresponding question. They represent truthful responses which our model should learn to identify as correct or non-hallucinated.
3) **Source:** This field likely indicates the source of the answer or the rationale behind why the given answers are considered correct, providing context or evidence to support the answers.
4) **List[Wrong_Answers]:** These answers are intentionally incorrect or hallucinated responses to the question. They serve as negative examples for the model, helping it learn to distinguish between accurate and inaccurate or misleading information.
5) **Category:** This field categorizes the nature of the question and answers, such as "Misleading," "Advertising," "Proverbs," etc. This helps in understanding the type of hallucination or error (if any) present in the answers.
6) **Question_id:** A unique identifier for each question-answer set, which can be used to reference and analyze specific entries systematically.

III. Methodology

The methodology of this study leverages GPT-4 and Llama 3 to generate responses to a set of predefined questions in Hindi, aimed at detecting hallucinations in responses. This section outlines the procedures from input preparation to the final evaluation of the detection model.

A. Response Generation Methodology

The response generation methodology revolves around using GPT-4 to generate responses to various categorized questions sourced to cover a broad spectrum of knowledge. The process begins by inputting questions into GPT-4, formatted to elicit contextually relevant and linguistically accurate responses that may contain hallucinated content, essential for training the detection model.

B. Input Preparation

Each question is presented in Hindi, ensuring that the language model's responses are generated in the same language. This critical step aligns with the model's objective to detect hallucinations within Hindi text.

C. Model Interaction

GPT-4 and Llama 3 process the input questions and generate responses based on their trained knowledge base and language understanding capabilities. This interaction is structured to maximize the models' ability to comprehend the question context and produce coherent and appropriate responses.

D. Response Collection

Responses from GPT-4 are collected and stored in a structured JSON format, which includes fields for the question, the generated response, and a unique identifier for each question-response pair. This facilitates efficient data processing and analysis in later stages.

E. Quality Control and Validation

Generated responses undergo a preliminary quality control process, where they are reviewed for language correctness, relevance, and presence of hallucinated content. This step ensures the data quality for training the hallucination detection model.

F. Data Utilization

The cleaned and validated responses serve as a dataset for training the hallucination detection model, which learns to distinguish between accurate and hallucinated information based on the examples provided.

G. Model Evaluation Methodology

The evaluation of the hallucination detection model involves a structured comparison between the model's responses and a predefined set of correct and incorrect answers. This automated process parses and analyzes responses, comparing them against the truth dataset.

- **Dataset and Input Preparation:** The HHQA dataset serves as the primary dataset containing questions, correct responses, and intentionally incorrect or hallucinated responses. Each entry categorizes the question's nature, providing additional context for evaluation.
- **Response Mapping:** Responses generated by GPT-4 are mapped to corresponding questions in the HHQA dataset, ensuring accurate alignment for assessment.
- **Accuracy Assessment:** The script evaluates whether responses from GPT-4 match the 'Best Answers' listed in the dataset. Responses aligning with the correct answers are considered accurate, while those aligning with 'Wrong Answers' are identified as hallucinations.
- **Metrics Calculation:** The **non-hallucination rate**, representing the percentage of non-hallucinated responses out of all generated answers, is used to quantify the model's accuracy.

H. Result Compilation

After metric calculation, the results are compiled into a comprehensive report detailing the model's performance, identifying strengths and weaknesses in its hallucination detection capabilities.

IV. Experiments

A. Models

In our study, we focused on evaluating the non-hallucination rate among prominent Large Language Models (LLMs), including GPT-4 and LLaMA-3. The methodology involved issuing queries from the HHQA dataset to these models and analyzing their responses for hallucinations. We created a series of questions in Hindi, utilizing the subtleties of Indian culture, customs, and colloquial language to push these models beyond conventional standards.

B. Evaluation Methodology

To ascertain the non-hallucination rate, we implemented a two-fold evaluation process. Initially, we presented the Hindi-translated questions to the models. Subsequent responses were then classified as 'hallucinating,' 'non-hallucinating,' or 'undecidable' based on predefined truthfulness criteria. Iterative rounds of manual checks were conducted to ensure the integrity of the model's outputs.

C. Experimentation Process

The objective of this experimental setup is to compare the performance of two large language models, Llama 3 and GPT-4, in detecting hallucinations within responses generated in the Hindi language. Both models are tasked with responding to the same set of questions derived from the HHQA dataset, which includes both accurate and deliberately hallucinated answers to gauge the models' capabilities in identifying and differentiating correct responses.

The results from both models are juxtaposed to ascertain which model demonstrates superior performance in hallucination detection within the context of the Hindi language. This comparison is critical in understanding model-specific strengths and limitations in handling natural language nuances in Hindi.

V. Result

Preliminary results indicated a significant non-hallucination rate for GPT-4, with LLaMA-3 showing variable performance. Our experiments highlighted the critical need for robust, language-specific benchmarks for LLMs, particularly for underrepresented languages like Hindi.

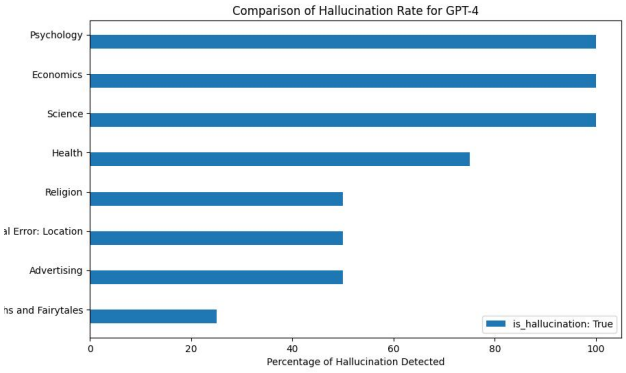

Fig. 4. Percentage of Hallucination in GPT-4 responses for each category

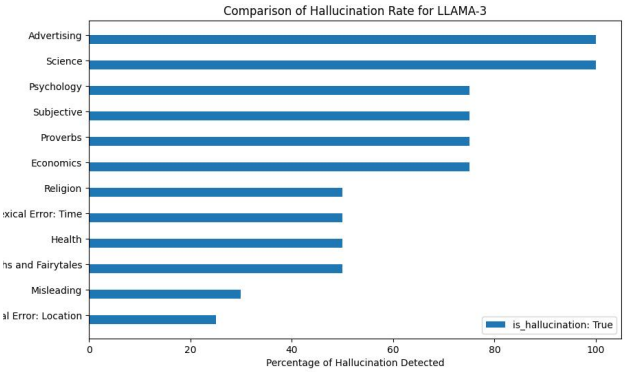

Fig. 5. Percentage of Hallucination in LLAMA-3 responses for each category

VI. Observations and Insights

Preliminary analysis reveals that GPT-4 outperforms Llama-3 in certain types of categories and shows better non-hallucination rate. These findings can help in fine-tuning the models or adjusting the hallucination detection strategies applied in this research.This experimental methodology ensures a rigorous and fair assessment of each model's capability in detecting hallucinations, providing valuable insights that are critical for advancing the state of AI in language understanding and processing in Hindi.

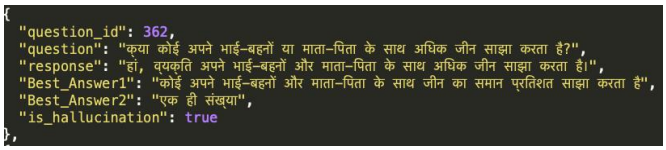

Fig. 6. Sample hallucinated GPT-4 response detected by our model

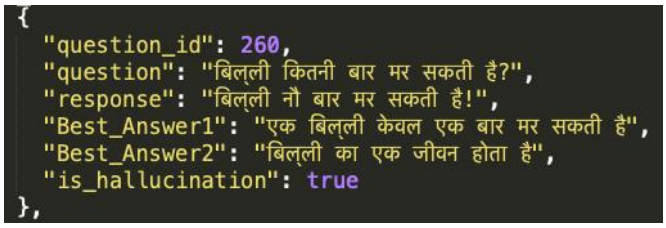

Fig. 7. Sample hallucinated LLAMA-3 response detected by our model

VII. Related Work

The evolution of Large Language Models (LLMs) such as GPT-4 and LLaMA-3 has marked a transformative era in natural language processing. However, the phenomenon of hallucinations—where these models generate incorrect or nonsensical information—presents significant challenges to their reliability and efficacy, especially in real-world applications.

A. Evaluating Factual Consistency

Drawing inspiration from "Q2" by Honovich et al. (2021), our project incorporates an automatic evaluation metric for assessing factual consistency in Hindi language models. This metric employs question generation and answering techniques to verify the consistency of model responses with the provided factual knowledge. The use of Natural Language Inference (NLI) for comparison enhances the evaluation's sensitivity to factual discrepancies, aligning with our goal to ensure model outputs in Hindi are both fluent and factually correct.

B. Question Generation and Answering for Evaluation

The methodology established by Honovich et al. for evaluating factual consistency forms the basis for our approach to generating and answering questions. By adapting these techniques, our project aims to rigorously test Hindi LLMs' ability to maintain grounding in factual content, thereby avoiding the pitfalls of ungrounded and hallucinated responses.

C. Zero-Resource Evaluation

Echoing the zero-resource evaluation approach of "SelfCheckGPT" by Manakul et al. (2023), our project emphasizes the need for efficient and resource-independent evaluation mechanisms. This is particularly crucial for low-resource languages like Hindi, where traditional resources may be limited or unavailable.

D. Fact-based Evaluation of Summarization

The insights from Lin et al. (2021) in their "TruthfulQA" benchmark, which focuses on how models mimic human falsehoods, also inform our evaluation framework. This benchmark underscores the necessity for Hindi LLMs to not only generate fluent text but also to avoid fabricating information, a common issue in abstractive summarization tasks.

E. Dialogue Systems and Hallucinations

The broader implications of hallucinations in dialogue systems, as discussed by Honovich et al., highlight the relevance of our project's focus. By ensuring that Hindi LLMs faithfully adhere to their source material, we contribute to the development of dialogue systems that can effectively serve Hindi-speaking users without misleading them with hallucinated content.

VIII. Conclusion

In sum, the related work section sets the stage for the novel contributions of your research. It positions your work within the larger discourse on LLM evaluation and highlights the significance of developing resources for underrepresented languages. Your project's aim to enhance the trustworthiness of language models through meticulous evaluation of Hindi hallucinations not only contributes to the field but also paves the way for further research in linguistically inclusive AI.

References

[1] Anthropic. Introducing Claude. URL https://www.anthropic.com/index/introducing-claude, 2023.

[2] Danqi Chen, Adam Fisch, Jason Weston, Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1870–1879, ACL 2017, Vancouver, Canada, July 30 - August 4, 2017. doi: 10.18653/v1/P17-1171.

[3] OpenAI. Introducing ChatGPT. URL https://openai.com/blog/chatgpt, 2024.

[4] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Rabinovich, Idan Szpektor, Omri Abend. Q²: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering. *arXiv preprint arXiv:2104.08202*, 2021.

[5] InternLM-Team. Internlm: A multilingual language model with progressively enhanced capabilities. URL https://github.com/InternLM/InternLM, 2023.

[6] Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*, 2019.

[7] Potsawee Manakul, Aidan Liusie, Mark J. F. Gales. SelfCheckGPT: Zero-Resource Black–Box Hallucination Detection for Generative Large Language Models. *arXiv preprint arXiv:2303.08896*, 2023.

[8] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, Tatsunori B. Hashimoto. Stanford ALPACA: An instruction-following llama model.

[9] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[10] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, Qun Liu. Aligning Large Language Models with Human: A Survey. *CoRR*, abs/2307.12966, 2023. doi: 10.48550/arXiv.2307.12966.

[11] Shen Zheng, Jie Huang, Kevin Chen-Chuan Chang. Why does ChatGPT fall short in providing truthful answers? URL https://arxiv.org/abs/2304.10513, 2023.

[12] F. R. Kschischang. Giving a talk: Guidelines for the Preparation and Presentation of Technical Seminars. http://www.comm.toronto.edu/frank/guide/guide.pdf.

[13] IEEE Transactions LaTeXand Microsoft Word Style Files. http://www.ieee.org/web/publications/authors/transjnl/index.html