Large Language Models are Biased to Detect Hallucination across Languages

Anonymous ACL submission

Abstract

Hallucinations in generative AI, particularly in large language models (LLMs), have emerged as a significant concern. These models often facilitate multilingual operations, including querying and conversation. Yet, few research efforts have been devoted to understanding hallucinations in a multilingual context, specifically regarding the equitable treatment of supported languages, due to the lack of available benchmarks. Addressing this gap, this paper first proposes Poly-FEVER, a large-scale publicly accessible multilingual fact extraction and verification dataset for hallucination detection that covers 11 languages and more than 800K fact claims with diverse topics. We utilize Poly-FEVER to evaluate the hallucination detection capabilities of ChatGPT and LLaMA-2 series. Our investigation extends to exploring hallucination causes, employing Latent Dirichlet Allocation (LDA) for topic distribution analysis and web searches to assess resource imbalances. Furthermore, we propose a mitigation approach combining linguistic adjustments and resource-oriented strategies, including a trained LDA model and the Retrieval Augmented Generation (RAG) approach, to enhance the robustness and reliability of multilingual information verification in LLMs. Our findings highlight the critical need for multilingual benchmarks Poly-FEVER and demonstrate the potential of mitigation strategy to address biased detection abilities on hallucinations, thus contributing to the development of more equitable and reliable multilingual LLMs.

1 Introduction

LLMs, such as those in the GPT family, have exhibited remarkable proficiency across diverse domains including education, healthcare, and legal affairs. In these applications, the accuracy and factual integrity of the content generated by LLMs are critical, particularly in areas requiring precise guidance, like medical and legal advice. Despite their advancements, mainstream LLMs predominantly utilize corpora that are imbalanced in terms of demographic groups (Shah et al., 2019; Li et al., 2023). Language, an important facet of demographic backgrounds, remains relatively underinvestigated in the context of detecting hallucinatory content in LLMs, particularly from the perspective of fairness and equitable usability.

Previous research has largely focused on hallucinations in LLMs within widely spoken languages, such as English (Yao et al., 2023), Chinese (Cheng et al., 2023), and German (Sennrich et al., 2023). This focus has led to a thorough understanding of hallucinatory outputs' mechanisms and the development of mitigation strategies. One approach involves prompt engineering, which includes retrieval augmentation to ground content in external evidence (Lewis et al., 2020b), feedback loops for refining responses, and prompt tuning to adjust prompts during fine-tuning for desired behaviors. Another strategy is model development (Tonmoy et al., 2024), focusing on creating models inherently less prone to hallucinating through architectural changes, novel loss functions, and supervised fine-tuning using human-labeled data.

Despite the critical insights gained, the focus on major languages has marginalized the experiences and challenges of LLMs trained on or applied to less common languages. Moreover, these investigations often employ differing datasets, leading to an absence of a systematic approach to assessing hallucinations across languages with uniform input. The complexity of this issue is multifold. First, there is a scarcity of appropriate datasets for cross-linguistic studies. Second, accurately detecting hallucinations on a large scale can be challenging, particularly in topics intertwined with local cultural and linguistic contexts. Third, the generation of apparently plausible yet inaccurate content manifests due to biased training and fine-tuning referenced resources on languages of marginal transmission. Nonetheless, ensuring equitable performance of

LLMs across languages, especially those that are underrepresented, is crucial from the perspective of hallucination consistency.

In this paper, we investigate hallucination detection capabilities in LLMs across multiple languages by extending the fact verification dataset to a multilingual benchmark, Poly-FEVER. While fact verification and hallucination detection share similarities, they are distinct tasks. Fact verification entails verifying its accuracy against known knowledge sources (Murayama, 2021; Zhu et al., 2021). Hallucination detection, in contrast, focuses on **identifying inaccuracies without the necessity of evidence provided within the data**.

We summarize the contributions as follows:

1. We introduce Poly-FEVER, an extensive, publicly available dataset tailored for multilingual fact extraction and verification. It covers 11 languages and includes over 800,000 fact claims on various topics, designed for hallucination detection tasks.
2. We analyze hallucination detection capabilities in advanced language models of ChatGPT and the LLaMA-2 series (7B, 13B, and 70B versions), using Poly-FEVER with both language-wise and classification prompts.
3. We investigate the reasons behind hallucinations on a multilingual scale, employing LDA for topic distribution analysis and automated web searches to assess resource imbalances.
4. We propose a mitigation strategy to address linguistic discrepancies and resource imbalances, incorporating an LDA-based model and an RAG strategy to enhance information verification robustness and accuracy.

2 Related Work

The hallucinations in LLMs are classified by Huang et al. (2023) into two types: intrinsic and extrinsic hallucinations. Intrinsic hallucinations involve self-contradictions within the instruction, context, or due to logical inconsistencies, while extrinsic hallucinations entail the generation of factually inconsistent or fabricated content.

Hallucinations in LLMs arise from data inconsistencies, limited contextual awareness, and ambiguous prompts, leading to contradictory or inaccurate responses. This is due to conflicting information in training datasets and an over-reliance on nearby data or co-occurrence statistics (Bender et al., 2021; Weidinger et al., 2021). Evaluating these hallucinations involves comparing generated content with the source, using metrics based on entity and relation triples. Traditional n-gram metrics such as ROUGE and PARENT-T show limited human relevance (Lin, 2004; Wang et al., 2020). Entity hallucination precision, proposed by Nan et al. (2021), and a relation-based metric introduced by researchers in 2019 which computes relation tuple overlap using trained fact extraction models, are crucial in this process Goodrich et al. (2019).

To identify factual inaccuracies in outputs produced by LLMs, a straightforward approach is to **compare the content generated by these models with information from established and reliable knowledge sources.** This methodology aligns with the workflow of fact-checking tasks, as outlined by Guo et al. (2022). The evaluation of extrinsic AI hallucination, particularly in LLMs, encompasses the observation of various factors, including long-text generation, contextual conflicts, and over-inference scenarios (Chen et al., 2023; Galitsky, 2023; Min et al., 2023). It is crucial to develop hallucination evaluation benchmarks tailored to the identification of factual inaccuracies and the measurement of deviations from the original context in LLM-generated outputs. Nevertheless, assessing the consistency of text with observable facts often necessitates external tools (Chern et al., 2023).

By switching perspectives and presenting detailed claims to LLMs for factual validation, it becomes possible to effectively assess the presence and extent of hallucinations in the model's output. Most widely used fact-checking datasets (Wang, 2017; Thorne et al., 2018; Diggelmann et al., 2020; Wadden et al., 2020), despite their coverage of diverse fields such as entertainment, politics, culture, business, science, and biology, are primarily available in English. This English-language bias can restrict the scope of fact-checking tasks and the detection of misinformation to content written in English. While ongoing efforts are being made to address this gap by creating multilingual fact-checking datasets (Gupta and Srikumar, 2021), comparing different languages with varying claims poses a unique challenge. Most multilingual hallucination research on LLMs focuses on large-scale machine translation tasks, which often produce hallucinated translations, raising trust and safety concerns (Pfeiffer et al., 2023). Hence, a multilingual benchmark with identical claims in various languages is essential for comprehensive fact-checking and cross-linguistic comparison, aiding in the detection of hallucinations in LLMs.

3 Poly-FEVER Benchmark

3.1 Poly-FEVER Overview

We propose and develop the Poly-FEVER benchmark, which encompasses over 800,000 labeled claims in 11 languages. Poly-FEVER is compiled from three extensively utilized English fact-checking sources: FEVER (Fact Extraction and VERification) (Thorne et al., 2018), Climate-FEVER (Diggelmann et al., 2020), and SciFact (Wadden et al., 2020). FEVER contains over 185,445 Wikipedia-based claims categorized as *Supported*, *Refuted*, or *NotEnoughInfo*. Climate-FEVER, a specialized extension of FEVER, targets climate change claims, offering a curated set of statements verified against scientifically reliable sources. For FEVER and Climate-FEVER, we only include claims labeled as *Supported* and *Refuted*. SciFact focuses on verifying biomedical claims from scientific literature, providing annotations on whether research articles *Support* or *Refute* these claims. Claims are based on universal facts covering various subjects like Arts, Music, Science, Biology, and History. This ensures their relevance and applicability in a multilingual context.

The Poly-FEVER dataset, mirroring the structure of the original FEVER family datasets (including FEVER, Climate-FEVER, and SciFact), comprises four key fields:

- ***id***: Each claim is assigned a unique ID.
- ***label***: This denotes the annotated label for the claim, which can be either *SUPPORTS* or *REFUTES*.
- ***claim***: The content of the claim itself, presented in 11 different languages, showcasing the dataset's multilingual aspect.
- ***evidence***: A list of evidence sets. Each set includes tuples of *[Annotation ID, Evidence ID, Wikipedia URL, sentence ID]*. The *Annotation ID* and *Evidence ID* are primarily for internal tracking and do not contribute to the scoring process. They are useful for future debugging or correcting annotation issues.

These fields enable an assessment of claims in a multilingual context, an essential aspect of our research in evaluating the fact-checking capabilities of language models across diverse languages.

3.2 Language Selection

Besides English (en), Poly-FEVER contains claims in another ten selected languages, including, Mandarin Chinese (zh-CN), Hindi (hi), Arabic (ar), Bengali (bn), Japanese (ja), Korean (ko), Tamil (ta), Thai (th), Georgian (ka), and Amharic (am), ordered by the number of native speakers, aiming to conduct a thorough assessment of language bias in LLMs when detecting hallucination. These multilingual claims reveal the LLMs' limitation in adapting to users with a diverse range of languages.

The Poly-FEVER benchmark selects 11 languages, including English as a baseline, based on their tendency to induce hallucinations. These languages, like Tamil, Arabic, Thai, Vietnamese, Japanese, and Mandarin Chinese, pose challenges due to grammatical structures, ambiguity, polysemy, and homophones. Cultural and contextual understanding is crucial, especially in languages with diverse dialects and sociolects like Hindi, Korean, Japanese, and Amharic. We also consider script and orthography, focusing on languages with non-Latin scripts such as Chinese, Arabic, Thai, Hindi, Bengali, Tamil, Georgian, and Amharic.

3.3 Multilingual Claim Translation

We utilize the Google Cloud Translation service to extend the benchmark to 11 languages for translation purposes. In our pursuit of accurate translation for this study, we explored various methods, including different translation APIs like DeepL Translation and employing LLMs for translation tasks. Given the bilingual proficiency of several coauthors, we assessed translation quality across two languages, focusing on cultural and contextual nuances. Our evaluation revealed that Google Translate outperformed other methods in accuracy. We excluded certain APIs, like DeepL Translate, due to their errors in sentence structure and verb tense, such as rendering "The book was read by him" to "Book by him read." Additionally, we are concerned about potential hallucinations when using LLMs for translations, which could compromise the integrity of our data. Therefore, we chose the Google Cloud Translation service for its globally recognized precision. Specifically, this service is employed to translate over 80,000 English factual claims into our 10 chosen languages. The total expenditure for utilizing Google Cloud Translation amounts to $2,644 USD. We applied GPT Estimation Metric Based Assessment (GEMBA) (Kocmi and Federmann, 2023), which does not require human reference to the translated content, to evaluate 5% of our benchmark to gauge translation quality as shown in Table 1.

Lang.	zh	hi	ar	bn	ja	ko	ta	th	ka	am
AveScore	91.3	92.4	90.8	91.8	91.5	**93.0**	90.0	91.1	90.8	88.9

Table 1: Average scores for each language on translation quality evaluation of 5% Poly-FEVER benchmark. A score of zero means 'no meaning preserved' and a score of one hundred means 'perfect meaning and grammar'.

4 Multilingual Hallucination Detection

4.1 Experimental Setup

Our selection of ChatGPT and LLaMA-2 was driven by their extensive language support and significant influence in the AI field. Although LLaMA-2 is primarily designed for English, it includes 27 other languages(Touvron et al., 2023), prompting us to examine its non-English hallucination detection. Initial evaluations (see Table 2: columns 4, 5, 6) revealed its multilingual limitations, as its performance approximated random guessing. Therefore, we shifted focus to ChatGPT, which better aligns with our goal of improving LLM capabilities across languages.

We conduct evaluations of ChatGPT 3.5 Turbo, LLaMA-2 7B, and 13B on a server equipped with a 12-core CPU and dual NVIDIA A5500 GPUs, each with 24GB of memory. For the evaluation of the larger LLaMA-2 70B model, we use a more powerful machine comprising 2 NUMA nodes. Each of these nodes features a 20-core CPU and 4 NVIDIA V100 GPUs with 32GB of memory, interconnected via NVLink to ensure fast and efficient data transfer. We evaluate models using the PyTorch deep learning framework, version 2.0.1. Furthermore, the temperature of LLaMA-2 is set to 0, which is intended to yield deterministic results.

4.2 Prompts Design

We designed a language-wise prompt and a classification prompt with 11 translated versions for a fact-checking task to guide LLMs to respond with either *‘true’* or *‘false’*.

We observed that complex, information-rich prompts often confused the LLMs, causing overlooked restrictions and irrelevant responses. To improve comprehension and adherence, we designed a clear and concise general prompt to classify factual claims as *‘true’* or *‘false’* based on truthfulness.

General Prompt: Label the following claim with truthfulness into: *‘true’* and *‘false’*. No other characters, no punctuation, all lower size.

Nevertheless, LLMs are overzealous in providing descriptions and references, deviating from categorizing each claim with a binary *true’* or *false’* response. Therefore, the reconstructed language-wise prompt clarifies restrictions into three sentences by providing the overall task definition, answer formats, and forbidden responses. As LLMs worse perform in inconsistent languages between prompts and claims, the language-wise prompt is translated into 11 selected languages.

Language-wise Prompt: Assess the truthfulness of the user’s claim and provide a response. Use *‘true’* to indicate that the claim is true, *‘false’* to indicate that it is false. Your response should only consist of *‘true’* or *‘false’*, without any additional characters or punctuation.

Although LLMs are directed to evaluate claims as *‘true’* or *‘false’* without contextual analysis, their responses are integrated sentence structure. Concerned with classification accuracy, we abandoned simple traditional models to understand the contextual preference of *‘true’* or *‘false’* in LLM responses. The classification prompt guides LLMs to categorize responses. For consistency and to control for hallucinations, responses are translated into English before classification, ensuring binary labels are clear for subsequent analysis.

Classification Prompt: Classify the input as *‘true’* or *‘false’* based solely on the indicative words or phrases within it. Use ‘true’ for it contains affirming words like *‘Correct,’ ‘TRUE,’ ‘really,’* or *‘the truth.’* Use *‘false’* for it contains negating or contradictory phrases like *‘Fake,’ ‘False,’* or any form of correction or contradiction within the input. Respond with only *‘true’* or *‘false’* for the input, without any additional text, characters, or punctuation.

4.3 Self-Detection of Hallucinations in LLMs

We address a concern regarding the capacity of LLMs to identify and mitigate hallucinations in the text they generate. Despite the construction of a Poly-FEVER dataset aimed at detecting hallucinations, a gap exists in the literature concerning the effectiveness of these models in recognizing inaccuracies within their own outputs. This gap stems from the fact that the claims and labels in Poly-FEVER are not produced by LLMs, raising questions about the representativeness of the hallucinations that LLMs themselves produce.

Rephrase Prompt: Rephrase the following claim without changing its meaning. Ensure the essence and intent remain unchanged.

To bridge this gap, we instruct LLMs to rephrase dataset claims in multiple languages while keeping their original meaning intact. Claims generated by LLMs lack verifiable ground truth, which is essential for systematically assessing the model’s hallucination detection accuracy. Given this limitation, we rephrased 800,000 claims to simulate LLM-generated content while preserving each claim’s ground truth. This controlled evaluation offers a consistent baseline for comparing the performance of spontaneously generated claims without ground truth. Due to the lack of automated metrics, we randomly selected 100 claims in English and Chinese and found that their truthfulness remained unchanged. By doing this, we aim to evaluate whether LLMs can effectively detect hallucinations in their own generated text with similar accuracy to their performance on external datasets.

5 Multilingual Hallucination Mitigation

5.1 Hallucination Causes Exploration

To investigate the induction of hallucinations in multilingual fact-checking tasks, we employ Latent Dirichlet Allocation (LDA) (Blei et al., 2003), assessing its performance across 22 topics in 11 languages. In addition, web search is applied to each claim with each language to observe the training datasets' bias on different languages.

5.1.1 LDA on Topic Distribution

LDA, a prominent topic modeling technique in Natural Language Processing (NLP), treats claims as mixtures of different topics, each defined by its distribution of words. It assumes each claim is linked to a unique set of topic distributions, with every topic distinguished by a specific word distribution. The primary aim of LDA is to unveil these hidden topics within a corpus of claims, adjusting word distribution patterns during training to best match the observed claims. Upon evaluating topic classifications ranging from 0 to 50 on the Poly-FEVER benchmark, we found that 22 topics achieved the highest stable coherence scores.

To scrutinize the induction of hallucinations within multilingual fact-checking tasks, we tailor our preprocessing and topic modeling phases to leverage the strengths of LDA. Preprocessing includes standardizing text (lowercasing, tokenizing, correcting typos, removing stop and short words, lemmatizing, and stemming) to prepare the dataset for LDA's in-depth analysis.

By constructing a Gensim Dictionary and transforming the texts into a BoW corpus, further enhanced to a TF-IDF model, we tailored the LDA model to identify 22 distinct topics through 200 iterative passes. This approach is directly aligned with our purpose of uncovering the thematic structures that may influence the occurrence of hallucinations in fact-checking across languages. The detailed analysis is in section 6.3.

5.1.2 Web Search on References bias

As LLMs are black boxes for users, we utilized a Python-based automated web scraping tool to examine the presence of claims in Poly-FEVERon the web across 11 selected languages. This examination aims to identify potential biases in training datasets that cause imbalanced performance on the multilingual fact-checking task.

To simulate varied internet user environments and bypass potential search engine restrictions, we incorporated diverse user agents, thus mirroring the wide spectrum of real-world internet access points. Further enhancing the authenticity of our approach, we introduced randomized time intervals between search queries, mimicking human browsing behavior and avoiding anti-bot mechanisms.

We performed web searches by Google's search engine to count the number of search results as a measure of the claim's online presence. This approach allows for a nuanced understanding of how widely each claim is disseminated across different linguistic contexts on the web. The detailed analysis is in section 6.2.

5.2 Mitigation Process

To mitigate the phenomenon of hallucinations in LLMs during multilingual fact-checking tasks, we employed a two-fold approach that integrates advanced linguistic topic extraction and reference retrieval technologies, as shown in Figure 1.

We utilized a pre-trained LDA model to identify nuanced topics within each claim. Then, these identified claims are translated into English, leveraging the extensive resources available in English for accuracy enhancement. Finally, the Retrieval-Augmented Generation (RAG) (Lewis et al., 2020a) technique was applied to extract the top 5 most relevant documents from the wiki_dpr dataset, serving as factual references. In addition, after translating a non-English response into English, LLMs categorized the response with exact *'true'* or *'false'* directed by the classification prompt.

This strategy standardizes input for consistency and enhances LLMs' fact-checking with reliable data, diminishing hallucinations and boosting the accuracy and reliability of multilingual fact-checking in LLMs. The experiment result is described in section 6.4.

5.2.1 Multilingual Topic Inference via LDA

We leveraged a pre-trained LDA model to examine each claim, uncovering nuanced topics, which are politics, sports, aviation, American football, warfare history, equestrian, architecture and construction, automotive racing, soccer, and film and television programs. Following the identification of these nuanced topics, non-English claims were translated into English by the Google translator

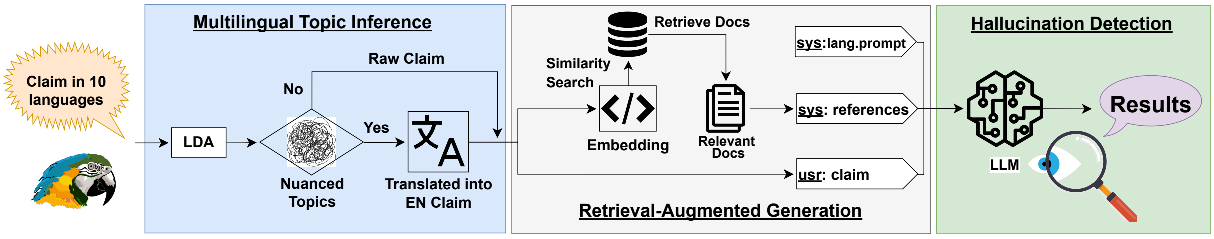

Figure 1: Multilingual hallucination mitigation process. Input claim with 11 languages' versions, utilize LDA to classify nuanced topics, employ RAG to provide references to LLMs, and output hallucination detection results.

to better simulate real-world application scenarios where original prompts are written in non-English. This selective translation approach, coupled with the targeted analysis facilitated by LDA, optimized our multilingual fact-checking process, ensuring that each claim is reviewed within the appropriate linguistic and thematic context.

5.2.2 Enhance Fact-Checking with RAG

We employed the RAG technique, leveraging its state-of-the-art capabilities to bolster the accuracy and relevance of responses produced by LLMs within our fact-checking framework. This system applies the Dense Passage Retrieval (DPR) mechanism (Karpukhin et al., 2020), which utilizes embeddings for document retrieval. RAG revolutionizes NLP by amalgamating generative models with an external knowledge retrieval component, enabling dynamic access to a vast corpus of information. This external augmentation enhances the model's internal knowledge base with pertinent external data during generation.

For external retrieval, we employ the wiki_dpr dataset, an extensive collection of 21 million Wikipedia passages, each adorned with DPR embeddings. These documents are segmented into 100-word, non-overlapping text blocks, optimizing the dataset for precise analysis and the evaluation of DPR's retrieval efficacy.

Leveraging Facebook AI Similarity Search (FAISS), we established an indexing framework based on the dataset's embeddings, streamlining the semantic retrieval of documents. Through the DPR Question Encoder, claims are transformed into semantically enriched embeddings. When these embeddings are matched against the FAISS index, the system identifies and retrieves the top 5 documents most relevant to the given claim. This retrieval process ensures the selection of documents that are semantically aligned with the claim. Consequently, the LLM is equipped with a rich input context that includes the original claim, constraints, and the substance of the retrieved documents.

6 Evaluation and Analytics

6.1 Hallucination on Fact-checking Task

Lang.	GPT	GPT Self.	L. 7B	L. 13B	L. 70B
en	**65.89%**	**61.88%**	**63.35%**	**64.27%**	**64.56%**
zh-CN	58.25%	53.94%	58.29%	59.88%	38.75%
hi	52.90%	58.10%	48.59%	54.33%	45.68%
ar	45.48%	58.80%	50.55%	54.97%	32.97%
bn	55.65%	58.73%	49.05%	52.30%	33.87%
ja	55.89%	58.95%	58.24%	59.57%	41.37%
ko	57.29%	60.14%	56.67%	58.74%	46.06%
ta	55.67%	59.98%	49.54%	50.86%	19.47%
th	56.82%	52.04%	53.90%	53.46%	48.09%
ka	57.37%	51.69%	47.39%	53.45%	46.15%
am	47.53%	47.06%	43.42%	48.64%	26.34%

Table 2: Comparison on accuracy of hallucination detection on fact-checking task by ChatGPT 3.5, ChatGPT 3.5 Self-Detection, LLaMA-2 7B, 13B, and 70B. Highest bolded, lowest underlined.

We deployed the same fact-checking process on ChatGPT 3.5, LLaMA-2 series to compare the hallucination detection abilities on multilingual claims. Specifically, we observed the self-detection ability of ChatGPT 3.5 by prompting it to rephrase the original claims and identify the validity of the generated context. In Table 2, English consistently shows the highest accuracy for all models. For other languages, accuracy rates around 50%, which is comparable to the expected outcomes of random guesses in binary-answer scenarios. It is also interesting that LLaMA-2 70B (see Table 8) outperforms ChatGPT 3.5 (see Table 5) in English Climate-FEVER and SciFact, yet it demonstrates inferior performance in the non-English versions of Climate-FEVER and SciFact.

Moreover, LLMs vary in self-detection accuracy across languages. The ChatGPT 3.5 model and its self-detection show performance differences, with the variant excelling in languages like Hindi due to targeted training. However, in English, the original model's broader training base provides superior accuracy, highlighting the impact of training scope and data diversity on self-detection capabilities.

6.2 Web Search Results

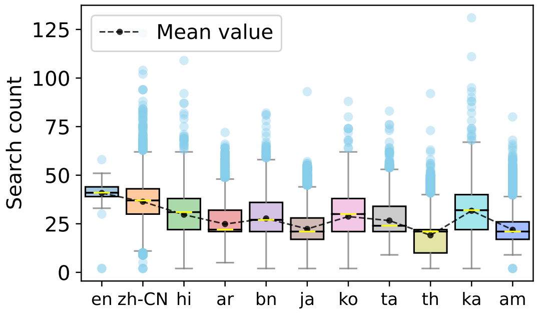

Figure 2: Web search distribution on multilingual claims. Middle 50% of search counts inside each box, mean values for every language are connected.

We employed a Python-based automated web scraping tool to analyze the web presence of claims in 11 languages, aiming to detect biases in training datasets that could lead to uneven performance in multilingual fact-checking tasks, which is described in the section 5.1.1.

Figure 2 compares the count of search results across 11 different languages, with an emphasis on identifying potential biases in the training datasets. Some languages, like English, show a relatively wide interquartile range (IQR), which contains the middle 50% of the data, indicating a high variability in the search count. Others, like Thai, have a much narrower IQR, indicating less variability. Languages like Amharic and Georgian have lower median and mean search counts, indicating less available content or fewer search results for these languages. This disparity could lead to an imbalanced performance in multilingual fact-checking, with better results in languages that have more content available, like English, and worse results in languages with less content.

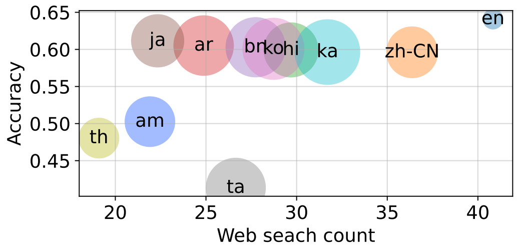

Figure 3: Detection accuracy on web search count. Bubble sizes depend on data variances to mean values on web search counts of each language.

Figure 3 compares relationships over web search counts with hallucination detection accuracy in various languages. The analysis indicates a relevance between a language's web search frequency and its fact-checking accuracy, with English and Chinese demonstrating high levels on both counts. Lesser-known languages like Amharic and Tamil, with low web search frequencies, exhibit reduced accuracy, indicating that limited data negatively affects model learning. Furthermore, the data suggests potential biases, with models possibly favoring languages that dominate web content, affecting their accuracy in languages with less online presence.

6.3 Fact-Checking Results with LDA

To investigate the induction of hallucinations, we utilized LDA to analyze detection performance over 22 topics in 11 languages, which is described in the section 5.1.1.

Figure 4 highlights LLMs generally perform best in English and struggle with Amharic and Thai. Topics such as Politics, Sports, Film/Television, and Warfare History prove challenging across languages due to their subjective nature, where personal biases and interpretations can obscure the distinction between fact and opinion. The dynamic nature of these fields, coupled with the need for specialized knowledge in areas like Architecture/Construction and Competitive Sports, complicates fact-checking. Historical contexts in Warfare History and Automotive Racing add another layer of complexity, as historical records can be biased or incomplete. Emotional ties to topics like American Football and Film/Television can bias information, while the subjective interpretation of data in Sports or Business/Finance makes objective verification difficult. The absence of universal standards in evaluating greatness in sports or the arts further complicates claim verification.

The standard deviation reveals varying degrees of biased hallucinations across languages. While topics 6, 7, 9, 10, 13, 15, and 17 show high average accuracy, there is significant variance, with languages like English and Chinese exhibiting higher accuracy. This variance underscores how hallucination biases differ among languages, reflecting the complex interplay between linguistic context and the accuracy of LLM predictions on specific topics.

Using LDA for linguistic topic extraction during the mitigation process, Table 3 shows the improvement in accuracy for non-English languages like Chinese, Arabic, Thai, and Amharic across nuanced topics 0, 2, 4, 5, 8, 11, 12, 14, 16, and 20. However, this approach results in decreased performance for Tamil on most of these nuanced topics, highlighting the method's variable impact

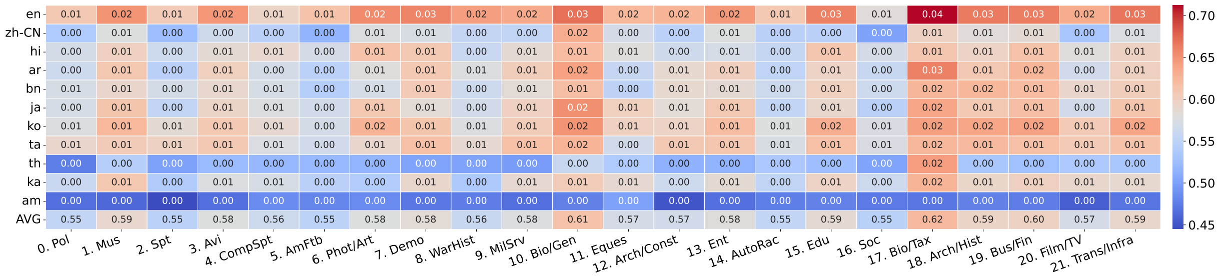

Figure 4: Average accuracy (by color) across topics by language with standard deviation (by annotation). The bottom row shows the average accuracy mean and standard deviation for one topic by considering all eleven languages.

Lang.	Poli	Sport	Comp	Football	WarHist	Equestr	ArchConst	AutoRace	Soccer	FilmTV
zh-CN	2.62	2.5	1.17	4.72	2.78	3.02	1.02	1.28	2.66	2.59
hi	1.97	0.14	-1.33	2.80	4.66	1.87	0.88	0.62	-0.39	1.52
ar	1.57	0.42	0	4.99	0.91	1.47	0.73	1.99	0.29	1.64
bn	0.86	0.28	-0.81	3.31	2.16	3.75	-0.58	-0.66	0.09	-1.13
ja	-0.35	-1.81	-1.38	1.11	1.21	-2.85	0.22	0.43	0.35	0.21
ko	0.4	-2.36	-0.76	1.08	1.87	-0.98	0.22	0.66	-0.83	-0.32
ta	-1.41	-4.31	-0.48	1.05	-1.62	0.33	-0.95	-1.89	-1.98	-3.00
th	6.95	6.05	2.9	4.09	6.65	4.07	6.43	2.27	3.79	3.93
ka	2.28	2.57	-0.6	2.83	4.16	1.39	0	0.76	0.67	0.25
am	7.71	7.44	3.76	5.05	6.98	4.24	5.41	3.60	4.41	7.09

Table 3: Percentage improvement of correct judgment on hallucination detection after translating nuanced topics into English. Only present selected topics on 10 languages.

Lang.	Original	RevPrompt	LDA	LDA+RAG
en	**65.89%**	**64.09%**	**63.48%**	**60.5%**
zh-CN	58.25%	59.61%	57.81%	53.74%
hi	52.90%	59.93%	60.03%	54.49%
ar	45.48%	60.52%	59.71%	55.67%
bn	55.65%	60.29%	59.57%	57.35%
ja	55.89%	61.16%	58.44%	55.44%
ko	57.29%	60.08%	60.46%	57.42%
ta	55.67%	41.34%	58.83%	56.89%
th	56.82%	48.04%	57.39%	53.37%
ka	57.37%	59.64%	59.44%	56.99%
am	47.53%	50.26%	54.24%	53.49%

Table 4: Accuracy of hallucination detection with original process, prompts revised, LDA, LDA+RAG

on different languages.

6.4 Multilingual Hallucination Mitigation

The detection accuracies of mitigating the multilingual hallucination described in section 5.2 are recorded step by step in Table 4. It is counterintuitive that intensifying efforts in English and Chinese fact-checking tasks degrades the performance of LLMs. Mitigation strategies aimed at reducing hallucinations hurt LLMs' fact-checking capabilities in these predominant languages. Specifically, the inclusion of topic information diminishes LLMs' comprehension, while the use of retrieved references constrains their access to extensive, potent internal data. In contrast, for languages spoken by smaller populations, employing LDA yields an average accuracy improvement of 4.83%. Combining LDA with RAG further enhances accuracy by an average of 1.76%. Arabic benefits most from the LDA approach, witnessing a 14.23% accuracy boost, and also shows significant gains from combining LDA with RAG. Hindi and Amharic experience considerable improvements with LDA, with accuracy increases of 7.13% and 6.71%, respectively, and positive outcomes from all strategies, underscoring their efficacy in hallucination mitigation. Bengali, Korean, and Tamil register modest gains with LDA, indicating the varied success of these strategies across languages. Conversely, Japanese, Georgian, and Thai exhibit minimal improvement or even slight accuracy declines with certain strategies, emphasizing the complex effects these methods may have, possibly influenced by linguistic traits or dataset characteristics. In summary, the mitigation methods mainly focus on balancing the performance of LLMs in English and other less frequently used languages, like Amharic with a maximized improvement of 6.71%, and Arabic with an improvement of 15.04%.

7 Conclusion

This paper proposes the publicly available Poly-FEVER, an innovative multilingual fact extraction and verification benchmark comprising over 800,000 factual claims in 11 widely spoken languages for hallucination detection in generative language models. Our in-depth investigation into the hallucination detection capabilities of LLMs, including ChatGPT and the LLaMA-2 series, illuminates the complexities of model performance and the effectiveness of language-wise and classification prompts. Furthermore, by exploring the underlying reasons for hallucinations through LDA and automated web searches, we reveal insights into topic distribution and resource imbalance issues. These findings guide the development of a targeted mitigation schema by integrating linguistic adjustments and resource-oriented strategies, such as a trained LDA model and the RAG approach for improving the accuracy and reliability of multilingual information verification.

8 Ethical Consideration

In the development of Poly-FEVER, we have placed a strong emphasis on ethical considerations, prioritizing data diversity, fairness, and environmental impact. Our benchmark encompasses a multitude of languages, with particular attention to low-resource languages, thereby promoting inclusivity and representation in the field of Large Language Model research. Furthermore, Poly-FEVER aims to guide the LLMs community towards ethical research practices by offering language diversity and topics. However, it is essential to view it as one criterion among others and encourage a broader examination of ethical implications. Moreover, the environmental sustainability of deploying large-scale computational resources and the importance of fostering collaboration within the research community for continuous improvement and ethical application of such technologies underline the multifaceted ethical landscape surrounding this innovative benchmark.

9 Limitation

This study, while providing valuable insights into the capabilities of LLMs in detecting and mitigating hallucinations across a range of languages, is subject to several limitations. The observed variations in the self-detection of hallucination in LLMs highlight the challenges in creating a standardized ability for hallucination detection across multiple languages. To enhance self-detection abilities, it may be necessary to adopt language-specific training approaches to LLMs. This could involve using larger, more diverse datasets for underrepresented languages or nuances of specific languages.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023. Complex claim verification with evidence retrieved in the wild.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. 2023. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Boris A Galitsky. 2023. Truth-o-meter: Collaborating with llm in fighting its hallucinations.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 166–175, New York, NY, USA. Association for Computing Machinery.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020a.

Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Taichi Murayama. 2021. Dataset of fake news detection and fact verification: a survey. *arXiv preprint arXiv:2111.03299*.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Francesco Piccinno, Massimo Nicosia, Xinyi Wang, Machel Reid, and Sebastian Ruder. 2023. mmt5: Modular multilingual pre-training solves source language hallucinations.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2023. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. *arXiv preprint arXiv:2309.07098*.

Deven Shah, H Andrew Schwartz, and Dirk Hovy. 2019. Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1072–1086, Online. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Biru Zhu, Xingyao Zhang, Ming Gu, and Yangdong Deng. 2021. Knowledge enhanced fact checking and verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3132–3143.

A Appendix

A.1 Hallucination on Multilingual Fact-checking Task

As shown in Table 5, for ChatGPT 3.5, it is observed that the model demonstrates great stability across all topic fields (general, climate, and science facts), with English consistently showing the highest judgment accuracy and Arabic showing the lowest judgment accuracy. The performance gap between the highest and lowest percentages for the three datasets are 23.11%, 38.22%, and 35.06%, indicating the greatest variability in the Climate-FEVER and the least in FEVER.

Lang.	FEVER	Climate-FEVER	SciFact
en	**65.89%**	**74.29%**	**71.57%**
zh-CN	58.25%	55.12%	54.83%
hi	52.90%	41.79%	42.14%
ar	45.48%	36.07%	36.51%
bn	55.65%	41.19%	43.15%
ja	55.89%	55.00%	55.56%
ko	57.29%	50.60%	54.11%
ta	55.67%	55.95%	43.58%
th	56.82%	46.31%	44.44%
ka	57.37%	50.36%	49.93%
am	47.53%	39.52%	37.95%

Table 5: Accuracy of hallucination detection by ChatGPT 3.5 on Poly-FEVER. Max and min values in each column are highlighted in bold and underlined.

LLaMA-2 (7B, 13B, and 70B) results are illustrated in Table 6, 7, and 8 respectively. As the size of the LLaMA-2 models increases, a noticeable bias towards different languages in LLMs becomes apparent. In the case of LLaMA-2 7B and 13B, Amharic exhibits the poorest performance in general fields, whereas Bengali demonstrates the weakest performance in categorizing science claims. LLaMA-2 70B displays significant variation in performance across different languages, particularly in terms of the lowest accuracy. In general and science topics, Tamil exhibits remarkably low estimation rates, registering only 15.00% and 10.17%, respectively.

Lang.	FEVER	Climate-FEVER	SciFact
en	**63.35%**	**77.70%**	**70.71%**
zh-CN	58.29%	58.59%	62.46%
hi	48.59%	41.92%	40.33%
ar	50.55%	46.53%	51.52%
bn	49.05%	46.31%	36.04%
ja	58.24%	61.23%	64.54%
ko	56.67%	52.36%	63.87%
ta	49.54%	46.56%	35.03%
th	53.90%	44.73%	48.49%
ka	47.39%	43.00%	37.31%
am	43.42%	48.15%	45.26%

Table 6: Accuracy of hallucination detection by LLaMA-2 7B on Poly-FEVER.

Lang.	FEVER	Climate-FEVER	SciFact
en	**64.27%**	**72.00%**	**72.44%**
zh-CN	59.88%	62.82%	66.57%
hi	54.33%	56.06%	57.33%
ar	54.97%	56.77%	61.41%
bn	52.30%	51.01%	52.25%
ja	59.57%	66.46%	67.65%
ko	58.74%	62.97%	67.36%
ta	50.86%	50.38%	61.02%
th	53.46%	45.57%	58.79%
ka	53.45%	55.00%	62.69%
am	48.64%	44.44%	53.68%

Table 7: Accuracy of hallucination detection by LLaMA-2 13B on Poly-FEVER.

Lang.	FEVER	Climate-FEVER	SciFact
en	**64.56%**	**78.42%**	**75.32%**
zh-CN	38.75%	49.72%	42.08%
hi	45.68%	45.96%	33.67%
ar	32.97%	30.03%	31.92%
bn	33.87%	19.46%	18.02%
ja	41.37%	51.27%	42.77%
ko	46.06%	47.88%	51.31%
ta	19.47%	19.85%	10.17%
th	48.09%	31.22%	30.65%
ka	46.15%	40.00%	33.03%
am	26.34%	11.11%	22.11%

Table 8: Accuracy of hallucination detection by LLaMA-2 70B on Poly-FEVER data.