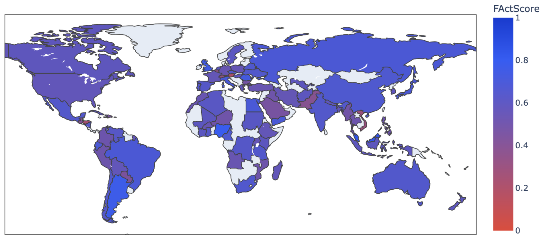
Multi-FAct: Assessing Multilingual LLMs' Multi-Regional Knowledge using FActScore

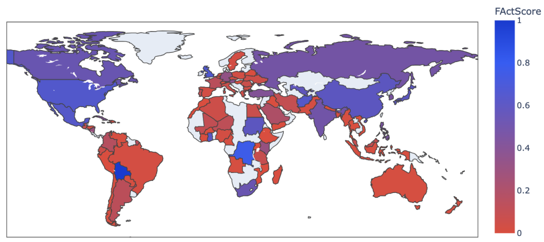
Sheikh Shafayat, Eunsu Kim, Juhyun Oh, Alice Oh
School of Computing, KAIST
Daejeon, Republic of Korea
{sheikh.shafayat, kes0317, 411juhyun}@kaist.ac.kr,
alice.oh@kaist.edu

Abstract

Large Language Models (LLMs) are prone to factuality hallucination, generating text that contradicts established knowledge. While extensive research has addressed this in English, little is known about multilingual LLMs. This paper systematically evaluates multilingual LLMs' factual accuracy across languages and geographic regions. We introduce a novel pipeline for multilingual factuality evaluation, adapting FActScore (Min et al., 2023) for diverse languages. Our analysis across nine languages reveals that English consistently outperforms others in factual accuracy and quantity of generated facts. Furthermore, multilingual models demonstrate a bias towards factual information from Western continents. These findings highlight the need for improved multilingual factuality assessment and underscore geographical biases in LLMs' fact generation[1].



(a) Factuality Heatmap of GPT-3.5 Biography Generation in **English**



(b) Factuality Heatmap of GPT-3.5 Biography Generation in **Korean**

Figure 1: While GPT-3.5 in English knows about the presidents of all countries almost uniformly, its knowledge in Korean is much smaller and mostly limited to East Asia and North America. Here color represents FActScore (bluer is better).

1 Introduction

Large Language Models (LLMs) are susceptible to factuality hallucination, a phenomenon in which the generated text contradicts established world knowledge (Huang et al., 2023; Zhang et al., 2023). Despite extensive research focusing on hallucination and factuality of LLMs in free-form generation, these endeavors have predominantly concentrated on English (Huang et al., 2023; Min et al., 2023; Mishra et al., 2024). Consequently, there exists a notable gap in our comprehension of the factual accuracy of LLMs when producing content in various languages. As highlighted by Kang et al. (2024), current metrics for detecting hallucination are inadequate in multilingual settings. The extent to which LLMs exhibit factual hallucination remains unclear across different languages. Similar to other capabilities, there may be a discernible decline in performance when LLMs are tasked with non-English contexts (Ahuja et al., 2023; Bang et al., 2023).

In this paper, we address this gap by systematically evaluating the factual accuracy of multilingual LLMs across different languages and geographic regions. We explore the following research questions: **R1:** Do multilingual models, such as ChatGPT, exhibit uniform factual accuracy across all languages in free-form generation? **R2:** Do these models demonstrate higher factual precision for content that aligns closely with the language (e.g., discussing a president of a country whose official language is used for generation)?

To address these questions, we introduce a novel pipeline tailored for evaluating factuality in a multilingual setting. Our approach first adapts the FActScore (Min et al., 2023) to accommodate multiple languages, and we make this pipeline openly

[1] Source code and data available at: https://github.com/sheikhshafayat/multi-fact

accessible as open-source. Using this pipeline, we then assess the factual accuracy of nine languages in biography generation tasks, ensuring geographical diversity through a curated list of subjects.

Through our analysis, we observe two significant findings. First, English consistently maintains an advantage in both factual accuracy and the quantity of generated facts compared to other languages when generating identical content. Second, content produced by multilingual language models tends to exhibit a stronger performance for factual information originating from Western regions, such as America and Europe, across the languages. These results emphasize the need for enhanced assessment methods in evaluating multilingual factual accuracy and underscore the geographical biases in the fact generation capabilities of LLMs.

2 Methodology

2.1 Task Selection

Min et al. (2023) select biography generation as their task because biographies contain factual information readily verifiable from existing sources like Wikipedia, and confirm through human evaluation that English Wikipedia is suitable for biography verification tasks.

Topic Selection We select "National Leaders of Each Countr" as the topic for factuality measurement, as it is fairly common across various cultures and languages to have well-known and well-described national leaders such as presidents and prime ministers. Specifically, we choose to focus on the president or head of state in the year 2015. This year is randomly selected from those prior to the emergence of LLMs. To determine our samples of the topic, we select countries by compiling a list of the 20 most populous nations from each continent.[2] The lists of countries and their corresponding leaders used in our experiments are detailed in Table 2 of the Appendix. We choose a total of 9 languages, divided among high, middle and low resource languages: English, German, French, Spanish, Arabic, Swahili, Chinese, Korean and Bengali. We make sure to include at least three languages that are widely spoken in each continent.[3]

[2]America, Asia, Europe, and Africa; America consisted of North and South America; Australia and New Zealand were also grouped with America.

[3]**Africa:** Arabic, Swahili, and French; **Europe:** Spanish, French, German; **Asia:** Korean, Bengali, Chinese; **America:** Spanish and French. English is widely spoken all over the world, and kept as the baseline.

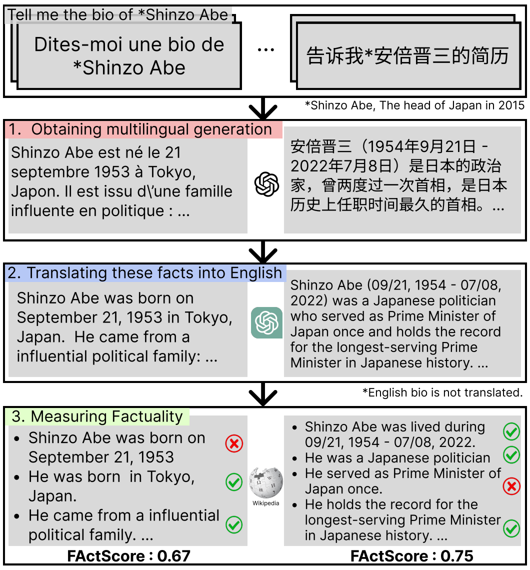


Figure 2: Pipeline of our methodology. For details of step three, fact verification, please refer to the original FActScore (Min et al., 2023).

.

2.2 Multi-FAct Pipeline

We introduce a novel pipeline for automatically measuring factuality of a multilingual LLM for multi-regional knowledge. This pipeline evaluates the biographies of presidents or heads of 80 countries across 9 languages.

The pipeline as shown in Figure 2 is structured into three main stages: 1) Obtaining multilingual generations (§ 2.3), 2) Translating these facts into English using GPT-3.5 (§ 2.4), 3) Measuring Factuality(§ 2.5).

2.3 Generating Facts

Following Min et al. (2023), we prompt the model to generate a biography of designated topics.

Model We use GPT-3.5 (`gpt-3.5-turbo-0613`) and GPT-4 (`gpt-4-1106-preview`) for the generation of facts. All experiments were carried out between January and February 2024. The model temperature was set at 1.0.

Prompt We translate the following prompt into 9 languages: `Write a biography of {name}`. The prompts were translated by GPT-4 and verified by native speakers' manual evaluation. We also transliterate each person's name into corresponding languages and validated by native speakers.

2.4 Translating Generated Facts

To compare the biographies written in the eight non-English languages, we translate the generated content from the original languages to English using GPT-3.5, specifically, `gpt-3.5-turbo-0125`. In principle, any translation model that can generate coherent translation would work. We do not use commercially available machine translators like Google Translate because manual evaluation revealed that they can not maintain the consistent gender of the person throughout the whole text.

We translate the generation and verify facts in English, rather than doing fact verification in corresponding languages with respective Wikipedia articles, for several reasons. Firstly, Wikipedia differs in size and scope for each language (Wikipedia contributors, 2024), which makes it difficult to compare cross-lingual factuality if the knowledge base is different. Moreover, some important steps in the original FActScore pipeline work best in English (e.g., RAG, NPM are designed primarily for English). We, however, do some quantitative analysis on the effect of LLM-based translation on FActScore evaluation and find that translation by GPT-3.5 does not affect FActScore of texts significantly (see appendix A.2).

2.5 Measuring Factuality

To evaluate the accuracy of the generated model's response M, we use the FActScore metric. This involves breaking down the translated generation into atomic facts—short sentences conveying a single piece of information, as defined in the FActScore paper. The accuracy of these atomic facts, denoted as A, is verified against the corresponding English Wikipedia article, serving as the reference knowledge source C.

$$
\begin{aligned}
\#\text{ of Correct Facts}(M) &= \mathbb{1}(\text{a is supported by } C) \\
\text{FActScore }(M) &= \frac{1}{|A|}\sum_{a \in A} \#\text{ of Correct Facts}(M)
\end{aligned}
\tag{1}
$$

We replace all proprietary models of the original FActScore pipeline with open-source models, ensuring cost-effectiveness while maintaining the quality of our system. We decompose LLM responses M into atomic facts using Mistral-7B (Jiang et al., 2023). The verification step also uses Mistral-7B along with RAG and NPM (Min et al., 2022) for the best performance. The details statistics can be found in Appendix A.2.

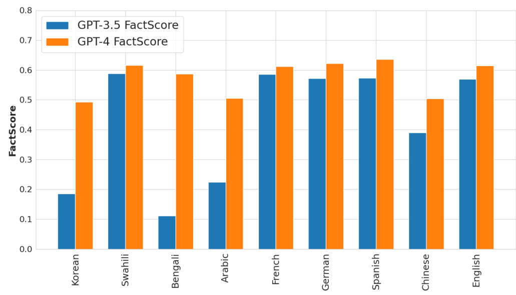


Figure 3: FActScore for each language of GPT-3.5 and GPT-4 for each language.

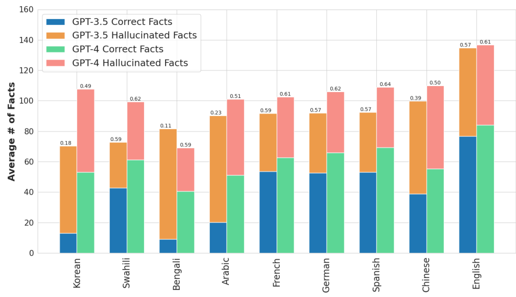


Figure 4: Number of correct and hallucinated facts by GPT-3.5 and GPT-4 for each language. The number on top of each bar is the FActScore . Note that in some languages ChatGPT generates shorter responses, means even though they might be generating not very good responses (or refusing to respond), their FActScore is higher.

3 Results

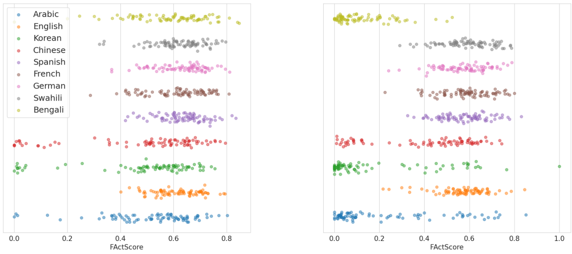
3.1 Does the Factuality Differ Across Languages?

Figure 3 displays the outcomes of measuring factuality across various languages using FActScore as a metric, uncovering significant variations among languages.

English, Spanish, French, German, and Swahili exhibit notably higher FActScore for both GPT-3.5 and GPT-4. Figure 5 shows the distribution of FActScore by language. Note that even in GPT4, some languages have long tail distribution in FActScore .

Figure 4 compares factuality across languages through the average number of correct and hallucinated facts, highlighting a significant gap in the quantity of facts generated between English and other languages. Remarkably, even with similar FActScore as shown in Figure 3, English surpasses other languages in generating a higher number of correct facts, thus delivering more useful responses.

These observations emphasize the influence of



(a) GPT-4 (b) GPT-3.5

Figure 5: FActScore Distributions by language between GPT-4 and GPT-3.5

Language	Africa	America	Asia	Europe	Mean	STD
Spanish	0.624	**0.641**	0.627	0.640	0.633	0.087
German	0.622	**0.642**	0.616	0.628	0.627	0.094
Swahili	**0.643**	0.637	0.585	0.610	0.619	0.096
English	0.607	**0.632**	0.613	0.607	0.615	0.086
French	0.595	**0.655**	0.609	0.594	0.613	0.110
Bengali	0.579	0.574	0.585	**0.589**	0.582	0.149
Arabic	0.493	**0.559**	0.485	0.522	0.515	0.167
Korean	0.481	**0.501**	0.490	0.476	0.487	0.208
Chinese	0.453	0.502	0.479	**0.514**	0.487	0.230

Table 1: Overall FActScore by language with mean and standard deviation (STD). The continents with the highest and second-highest values for each language are emphasized in bold and underlined, respectively.

output length differences on the number of correct and hallucinated facts across languages, despite similar FActScore values. English and other high-resource languages (e.g., Chinese, Spanish) typically produce more facts due to longer outputs. In contrast, low-resource languages like Bengali yield fewer facts because of shorter responses, even when their FActScore are comparable to English (0.58 for Bengali, 0.61 for English). This discrepancy further widens the factuality gap between low-resource and high-resource languages. Hence, the denormalization of output length in multilingual contexts suggests that evaluating factuality requires considering both FActScore and the number of correct facts for a thorough assessment.

3.2 Does the Factuality *Geographically* Differ Across Languages?

In this section, we explore whether the factuality of models varies with the geographic distribution of the language under study. Since GPT-4 performs better in low resource language as seen in figure 3, from here on, our analysis is conducted using only the outputs from GPT-4.

Table 1 presents the overall FActScore by language and continent. Interestingly, languages with higher mean values tend to have lower standard

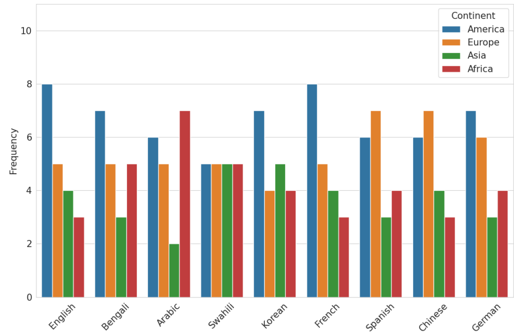


Figure 6: Continental distribution of top 20 countries that had the highest FActScore in each language. The right way to interpret this plot is, say for German, we take the twenty most accurate presidents' biographies and then look into which continents they belong to.

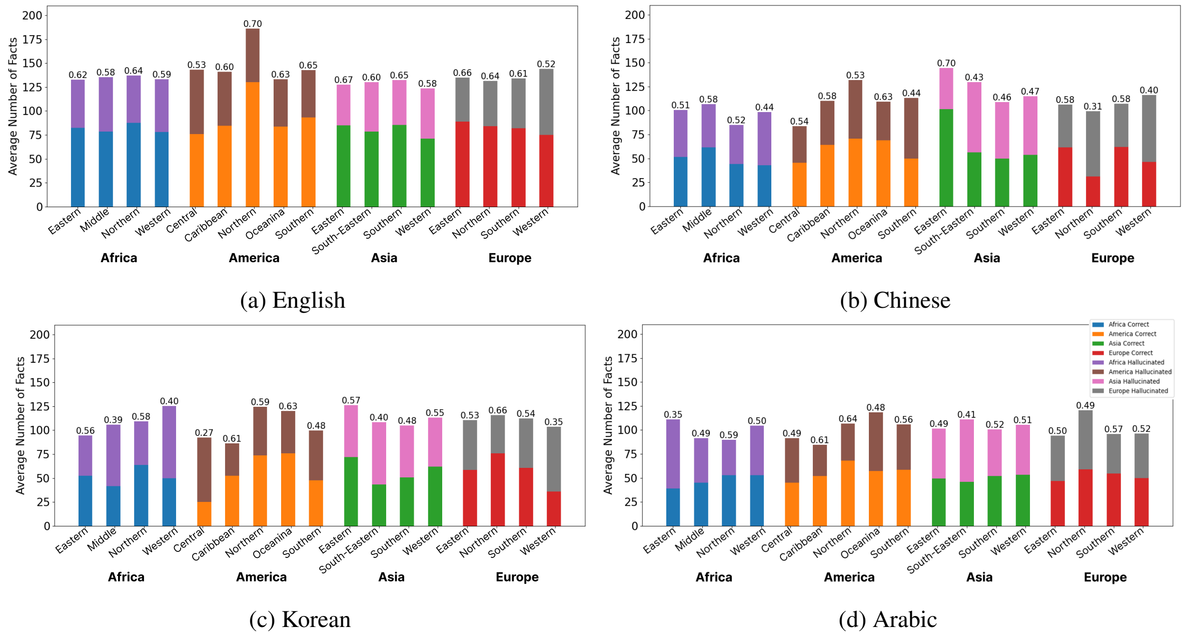
deviations, indicating that these languages maintain a relatively uniform level of factual accuracy regardless of the geographical area. Western languages such as English, Spanish, and German notably score high across continents, while Chinese, although a high-resource language, exhibits low factual precision. Additionally, among the nine languages evaluated, six demonstrate their highest performance in content related to the American continent, highlighting a prevalent American-centric bias in GPT's knowledge across most languages.

To conduct an in-depth analysis, we counted which continent the top-K FActScore for each language pertains to, thereby mapping out the distribution across continents. Figure 6 presents the findings for when K equals 20, representing the top 25 percentile of the dataset.

Across analyzed languages, America and Europe are the primary focal points, underscoring their global influence. Even in languages like Korean and Chinese, distinct from these continents, the most accurate outputs primarily relate to American and European topics. For Arabic, while the highest FActScore is linked to Africa, likely reflecting its geographical roots, the second highest pertains to America. This trend highlights a Western-centric bias in the model's factual content distribution across languages.

3.3 Geographical Biases in Factual Precision Across Subregions

In this section, we examine how languages with notable variations in factuality across continents exhibit distinct geographical biases. Further analysis involves breaking down the continents into



(a) English (b) Chinese

(c) Korean (d) Arabic

Figure 7: Fine-grained geographic distribution of languages with the highest standard deviation (Chinese, Korean, Arabic) and the least standard deviation (English) is given[4]. The top bar represents the average number of hallucinated facts, and the bottom one denotes correct facts. The number on top of each bar represents FActScore .

sub-regions to compare the number of correct facts and FActScore among three languages: Chinese and Korean, which exhibit he highest standard deviation (SD), and English, which demonstrates the lowest SD among the four continents, as detailed in Table 1 of § 3.2.

As shown in Figure 7, Chinese stands out in Eastern Asia, achieving the highest number of correct facts among all regions analyzed, indicating a pronounced bias towards its main linguistic area. Korean, also categorized in Eastern Asia, shows comparable levels of factuality in Eastern Asia to that of regions associated with America and Europe, underscoring geographical biases in the model's factual precision. English, while displaying a bias towards North America, its main region, exhibits relatively even factuality across other regions. This analysis underscores the geographical biases in factuality, with languages showing preferential accuracy towards regions where they are predominantly spoken (Figures for the other language can be found in Figure 10 in Appendix A.3.)

3.4 Is there a correlation between FActScore across languages?

Figure 8 shows the language-wise distribution of FActScore grouped by continents. English, Spanish, French and German have the most correlation with each other and these languages have very sim-

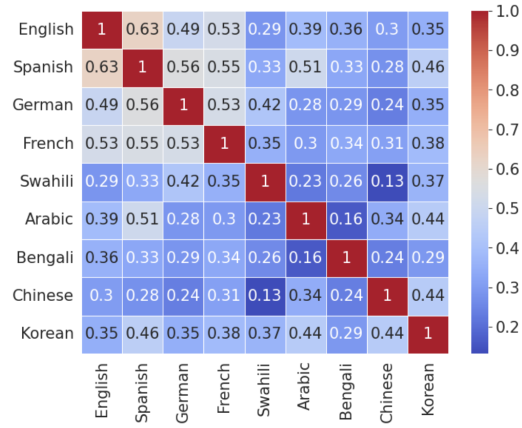


Figure 8: FActScore correlation matrix computed over all the FActScore values per language.

ilar FActScore . One interesting observation is that, Swahili, even though has very similar FActScore with English, Spanish, French, and German, (table 1) does not correlate with those languages as much. This likely reflects the fact that these two languages have slightly different knowledge distributions, which we also observed in figure 6.

[4]Please note that we discarded South Africa and Uzbekistan from this analysis because they corresponded to only elements in their respective geographic categories, while other categories represent averages of two or more countries. Details of which countries belong to which category can be found in appendix A.1

4 Qualitative Analysis

Model Generations with Low FActScores Qualitative analysis on model generations, and generations with low FActScore reveals some interesting patterns. Firstly, LLMs provide much more informative outputs in English than in other language, which is also quantified in figure 4. For example, in the biography of Park Geun-hye, the former president of South Korea, GPT-4 generates the following sentence in English: "Park Geun-hye studied electrical engineering at Sogang University in Seoul and later pursued further studies at the University of Grenoble in France" while when asked in Korean, it only mentions "She studied in France (...)". This highlights the need to develop a metric that measures beyond simple factuality of each atomic sentence, but the informativeness, or specificity of it. Some generation failures in Korean (the outliers in Figure 5) reveal that GPT-4 suggests there's no famous person by that name and proceeds to generate a biography template. Interestingly, the English version of GPT-4 never refuses to generate a biography of the leaders in our list.

Comparison of GPT-3.5 & GPT-4 The qualitative difference between GPT-3.5 and GPT-4 is also striking. For example, while GPT-3.5 struggles to even generate fluent and coherent texts in Bengali, GPT-4 Bengali generations are mostly fluent and accurate, however the generations were short and lacking specific details. One interesting example regarding difference between GPT-3.5 and GPT-4 comes from Arabic. In GPT-3.5, when asked about Pakistani former president Mamnoon Hussain, GPT-3.5 hallucinates a biography nonexistent "ambitious young man from Saudi Arabia", while GPT-4 rightfully identifies and writes the biography of former Pakistani president. GPT-4 also refused to answer more often than GPT-3.5; For example, in Chinese, it refuses to answer the biography of Sushir Koirala and claims there is no famous person of this name, however goes on to generate a fictional biography of an Indian person (with a disclaimer). GPT-4, too, does not always refrain from answering, when it doesn't know. For example, when asked about Ollanta Humala in Chinese, it goes on to describe Ollantay Tambo, an archeological site in Peru.

5 Discussion & Limitations

In this study, we introduce a unique approach to examining the distribution of geo-culturally based knowledge using atomic factuality in long-form LLM generation. We observe that LLMs exhibit a Western-centric bias, which remains evident even in languages from non-Western regions. This discovery prompts intriguing inquiries regarding the representation of multilingual knowledge within LLMs. For instance, do LLMs learn to represent more facts about South America in Bengali by being exposed solely to content in English or Spanish? Or, to what extent is multilingual data required to acquire factual knowledge in a specific language? However, due to the proprietary nature of models like GPT-4, answering these questions proves to be challenging. Future research should delve deeper into this issue with language models that also have open datasets. Recent studies such as Dolma (Soldaini et al., 2024) and OLMo (Groeneveld et al., 2024) represent a step forward in this regard, although their focus is limited to monolingual English corpora and models.

One limitation of our current study pertains to small sample bias. We examine only a single sample from each country, which may not provide a comprehensive representation, especially considering that automated FActScore sometimes yields slight overestimations or underestimations. Consequently, caution is needed when interpreting individual scores. For instance, assertions such as "In Korean, the biography of the prime minister of Malaysia is 5% more accurate than that of the Romanian president" should be made with caution. In this paper, we present all findings as averages across languages, continents, or regions to mitigate sample bias. Future research should include multiple individuals, not only political leaders, from each country to alleviate this limitation.

Another limitation of our methodology is the varying durations national leaders have been in power, potentially biasing internet corpora in their favor. One solution could be to prompt models to generate a "country biography" or "capital biography" (e.g., "Write me an essay on Addis Ababa"). However, we opted not to pursue this in our study, as the original FActScore research did not include a human evaluation baseline for non-biographical domains.

6 Related Work

LLM Factuality Evaluation The assessment of factual accuracy in natural language texts predates the advent of LLMs (Guo et al., 2022). Our Multi-FAct pipeline follows the previous fact evaluation research, which involves validating claims based on external resource such as Wikipedia article (Thorne et al., 2018; Krishna et al., 2022; Zhong et al., 2019) or Google search (Chern et al., 2023). Our work is heavily inspired and based on Min et al. (2023), which suggests to measure the factuality of a long-form generated output of LLMs by breaking it down into atomic facts. Another line of work focuses on evaluating factual accuracy solely based on model output or internal states, eliminating the need for external knowledge bases like Wikipedia (Manakul et al., 2023; Azaria and Mitchell, 2023; Dhuliawala et al., 2023). While such methods offer resource efficiency, they often sacrifice accuracy compared to approaches leveraging external resources for fact verification.

Multilingual Factuality Evaluation In the realm of multilingual factuality, Gupta and Srikumar (2021) presents the X-Fact dataset for multilingual fact-checking. Similarly, Thorne et al. (2018) have developed frameworks for multilingual fact-extraction and verification evaluation. Additionally, Aharoni et al. (2022) introduce mFACE, a dataset aimed at evaluating multilingual factual consistency. While these contributions significantly advance the understanding of multilingual factuality, they primarily focus on the assessment of factual accuracy in existing articles or benchmarks rather than examining the factual accuracy in the generative outputs of models. In this vein, Kang et al. (2024) explore the efficacy of existing metrics designed to detect hallucinations of LLM-generations, proposed for English texts, in a multilingual setting and point out they fall short in multilingual contexts. Our study extends the endeavor of developing a factuality evaluation metric that can be applied in multiple languages by introducing a Multi-Fact pipeline.

Geo-culture biases of LLMs There has been considerable research regarding cultural and linguistic bias in LLMs (Hovy and Yang, 2021; Cao et al., 2023; Hershcovich et al., 2022; Huang and Yang, 2023; Jin et al., 2023). A concurrent work (Mirza et al., 2024) examines geographic knowledge gaps in LLMs and finds that GPT models show higher factuality in the Global North over the Global South, consistent with our findings. Manvi et al. (2024) also reach a similar conclusion. However, none of the studies concerns free-form generation with non-English languages.

7 Conclusion & Future Work

In this paper, we introduce a simple pipeline to evaluate factuality of LLMs in multilingual settings and demonstrate its effectiveness with the national leaders' biographies in 80 countries around the world. Our findings reveal that multilingual models, such as GPT-4, tend to produce more factual content in high-resource languages, with notably longer outputs in English. Additionally, we observe a pervasive Western-centric bias across all languages; on average, even Asian and African languages demonstrate better performance for American and European leaders. However, a more detailed geographical analysis indicates that languages still reflect some aspects of their local geographic distribution, albeit to a lesser degree.

While factuality might be an interesting way to look into a model's knowledge about a certain person, our current implementation treats every fact equally. FActScore does not differentiate between the informational value of facts. For instance, the statements "Person X attended MIT" and "Person X attended a famous university" are treated as equivalent, though the former is obviously more informative and useful. Future research should aim to distinguish between specific, valuable facts and generic, less informative ones.

Another area for future investigation could involve examining the consistency of model-generated facts across different languages; for example, whether the model's depiction of Barack Obama's childhood in English aligns with its portrayal in Chinese, or whether it keeps its focus solely on his political endeavors.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2022. mface: Multilingual summarization with factual consistency evaluation. arXiv preprint arXiv:2212.10622.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023.

MEGA: Multilingual evaluation of generative AI. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 4232–4267, Singapore. Association for Computational Linguistics.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Yang Trista Cao, Anna Sotnikova, Jieyu Zhao, Linda X. Zou, Rachel Rudinger, and Hal Daume III au2. 2023. Multilingual large language models leak human stereotypes across language boundaries.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. arXiv preprint arXiv:2307.13528.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models.

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. arXiv preprint arXiv:2106.09248.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 588–602, Online. Association for Computational Linguistics.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 7591–7609, Singapore. 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.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. arXiv preprint arXiv:2301.08745, 1(10).

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2023. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries.

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. 2024. Comparing hallucination detection metrics for multilingual generation. arXiv preprint arXiv:2402.10496.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. Proofver: Natural logic theorem proving for fact verification. Transactions of the Association for Computational Linguistics, 10:1013–1030.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. arXiv preprint arXiv:2402.02680.

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.

Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wen-tau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Nonparametric masked language modeling. arXiv preprint arXiv:2212.01349.

Shujaat Mirza, Bruno Coelho, Yuyuan Cui, Christina Pöpper, and Damon McCoy. 2024. Global-liar: Factuality of llms over time and geographic regions. arXiv preprint arXiv:2401.17839.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. arXiv preprint arXiv:2401.06855.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Wikipedia contributors. 2024. Wikipedia:multilingual statistics — Wikipedia, the free encyclopedia. [Online; accessed 24-February-2024].

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2019. Reasoning over semantic-level graph for fact checking. arXiv preprint arXiv:1909.03745.

A Appendix

A.1 List of Countries and Presidents

Table 2 lists all the names of the countries, presidents and their Geographic locations. Please note that, we changed Australia and New Zealand to Oceania in the UN geoscheme for plotting purpose.

A.2 Multi-FAct Implementation Details

We replicated the original FActScore implementation and replaced InstructGPT for atomic fact generation with Mistral-7B. We also used Mistral 7B for the retrieval-based verification step. The final version uses retrieval + Mistral-7B + NPM. NPM threshold was set to 0.3.

A.2.1 Replication of original FActScore

We provide a comparison between our implementation with the results reported in the paper. For all the following experiments, we used a subset of ChatGPT-generation human-annotated factual data from the original paper.

Note that the Human FActScore value in the table 3 is different from what is provided in the paper as we averaged over our subset only. We replicated the best-performing method (retrieval + NPM) on this dataset reported in the paper and marked it as Error (Original). Our higher accuracy is likely due to Mistral 7B's better capabilities and our better hyperparameter tuning.

Please note that there is some stochasticity in the automated evaluation of FActScore , which is why we need to be cautious of individual scores, and only work with scores averaged over a large number of samples.

A.2.2 Evaluation of Multi-FAct Pipeline

The basis of our pipeline relies crucially on the GPT3.5-based translation (Step 3 in figure 2). It is imperative to see whether this translation step hurts the automated estimation of FActScore in a multilingual setting.

Evaluation Method: The basis of our evaluation method is that the factuality of a text is translation invariant, ie, factual knowledge in a text should not change when it is translated to other languages. To evaluate this, we first take the human-annotated examples provided in original FActScore , and translate them into four languages using GPT4. We use GPT4 as previous studies (Jiao et al., 2023) revealed GPT4 to be a good translator and better

Country	Continent	Region	President
Ethiopia	Africa	Eastern Africa	Hailemariam Desalegn
Tanzania	Africa	Eastern Africa	Jakaya Kikwete
Kenya	Africa	Eastern Africa	Uhuru Kenyatta
Uganda	Africa	Eastern Africa	Yoweri Museveni
Mozambique	Africa	Eastern Africa	Filipe Nyusi
Madagascar	Africa	Eastern Africa	Hery Rajaonarimampianina
DR Congo	Africa	Middle Africa	Joseph Kabila
Angola	Africa	Middle Africa	José Eduardo dos Santos
Cameroon	Africa	Middle Africa	Paul Biya
Egypt	Africa	Northern Africa	Abdel Fattah el-Sisi
Algeria	Africa	Northern Africa	Abdelaziz Bouteflika
Sudan	Africa	Northern Africa	Omar al-Bashir
Morocco	Africa	Northern Africa	Abdelilah Benkirane
South Africa	Africa	Southern Africa	Jacob Zuma
Nigeria	Africa	Western Africa	Muhammadu Buhari
Ghana	Africa	Western Africa	John Mahama
Côte d’Ivoire	Africa	Western Africa	Alassane Ouattara
Niger	Africa	Western Africa	Mahamadou Issoufou
Burkina Faso	Africa	Western Africa	Michel Kafando
Mali	Africa	Western Africa	Ibrahim Boubacar Keïta
Australia	America	Australia and New Zealand	Tony Abbott
New Zealand	America	Australia and New Zealand	John Key
Haiti	America	Caribbean	Michel Martelly
Cuba	America	Caribbean	Raúl Castro
Dominican Republic	America	Caribbean	Danilo Medina
Mexico	America	Central America	Enrique Peña Nieto
Guatemala	America	Central America	Otto Pérez Molina
Honduras	America	Central America	Juan Orlando Hernández
Nicaragua	America	Central America	Daniel Ortega
United States	America	Northern America	Barack Obama
Canada	America	Northern America	Justin Trudeau
Brazil	America	South America	Dilma Rousseff
Colombia	America	South America	Juan Manuel Santos
Argentina	America	South America	Cristina Fernández de Kirchner
Peru	America	South America	Ollanta Humala
Venezuela	America	South America	Nicolás Maduro
Chile	America	South America	Michelle Bachelet
Ecuador	America	South America	Rafael Correa
Bolivia	America	South America	Evo Morales
Paraguay	America	South America	Horacio Cartes
Uzbekistan	Asia	Central Asia	Islam Karimov
China	Asia	Eastern Asia	Xi Jinping
Japan	Asia	Eastern Asia	Shinzo Abe
South Korea	Asia	Eastern Asia	Park Geun-hye
Indonesia	Asia	South-Eastern Asia	Joko Widodo
Philippines	Asia	South-Eastern Asia	Benigno Aquino III
Vietnam	Asia	South-Eastern Asia	Trng Tn Sang
Thailand	Asia	South-Eastern Asia	Prayut Chan-o-cha
Myanmar	Asia	South-Eastern Asia	Thein Sein
Malaysia	Asia	South-Eastern Asia	Najib Razak
India	Asia	Southern Asia	Narendra Modi
Pakistan	Asia	Southern Asia	Mamnoon Hussain
Bangladesh	Asia	Southern Asia	Sheikh Hasina
Iran	Asia	Southern Asia	Hassan Rouhani
Afghanistan	Asia	Southern Asia	Ashraf Ghani
Nepal	Asia	Southern Asia	Sushil Koirala
Turkey	Asia	Western Asia	Recep Tayyip Erdoğan
Iraq	Asia	Western Asia	Fuad Masum
Saudi Arabia	Asia	Western Asia	Salman of Saudi Arabia
Yemen	Asia	Western Asia	Abdrabbuh Mansur Hadi
Russia	Europe	Eastern Europe	Vladimir Putin
Ukraine	Europe	Eastern Europe	Petro Poroshenko
Poland	Europe	Eastern Europe	Andrzej Duda
Romania	Europe	Eastern Europe	Klaus Iohannis
Czech Republic	Europe	Eastern Europe	Miloš Zeman
Hungary	Europe	Eastern Europe	János Áder
Belarus	Europe	Eastern Europe	Alexander Lukashenko
Bulgaria	Europe	Eastern Europe	Rosen Plevneliev
United Kingdom	Europe	Northern Europe	David Cameron
Sweden	Europe	Northern Europe	Stefan Löfven
Italy	Europe	Southern Europe	Sergio Mattarella
Spain	Europe	Southern Europe	Mariano Rajoy
Greece	Europe	Southern Europe	Prokopis Pavlopoulos
Portugal	Europe	Southern Europe	Marcelo Rebelo de Sousa
Germany	Europe	Western Europe	Joachim Gauck
France	Europe	Western Europe	François Hollande
Netherlands	Europe	Western Europe	Mark Rutte
Belgium	Europe	Western Europe	Charles Michel
Austria	Europe	Western Europe	Heinz Fischer
Switzerland	Europe	Western Europe	Simonetta Sommaruga

Table 2

Metric	Human FS	FS (ours)	Error (ours)	Error (Original)
Value	0.626 ± 0.238	0.637 ± 0.207	-0.011 ± 0.129	-0.0383 ± 0.121

Table 3: Summary of Replication. Error refers to the difference between human FActScore and FActScore determined by our method. Error (Original) refers to the error by the best-performing method reported on the paper. We also report the standard deviation alongside each metric.

than GPT3.5 and commercially available machine translators. Manual inspection by native speakers of a subset of our languages also confirmed this.

Language/Dataset	FActScore
Human eval (EN)	0.626 ± 0.238
EN	0.637 ± 0.207
KOR	0.625 ± 0.211
BN	0.600 ± 0.216
FR	0.638 ± 0.200
SW	0.654 ± 0.196

Table 4: FActScore evaluation Across Languages. EN refers to FActScore evaluation on original English human annotation; other languages refer to FActScore evaluation on GPT4 translation of the human annotation.

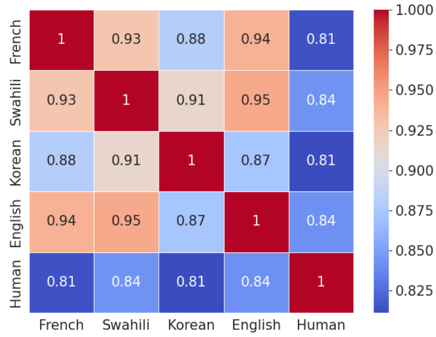


Figure 9: Correlation of FActScore evaluated on translated human-annotated data.

We then applied our multilingual FActScore pipeline (Figure 2) on this GPT4 translated text. The premise is that if our pipeline (which involves GPT3.5 translation[5]) provides similar FActScore as the original English annotated data, it might provide good evidence that having a translation step in our pipeline doesn’t hurt FActScore estimation significantly.

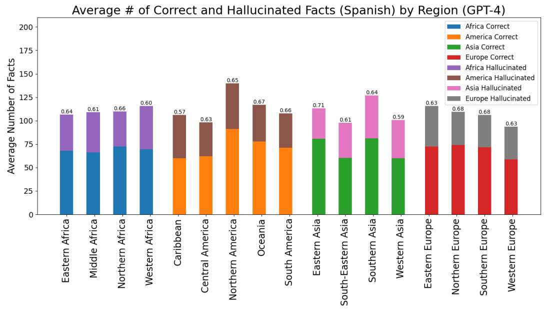
[5]We did not use GPT4 translation in Multi-FAct pipeline to reduce API costs

The results are provided in Table 4. We see that running Multi-FAct on non-English data does not significantly degrade FActScore estimation.

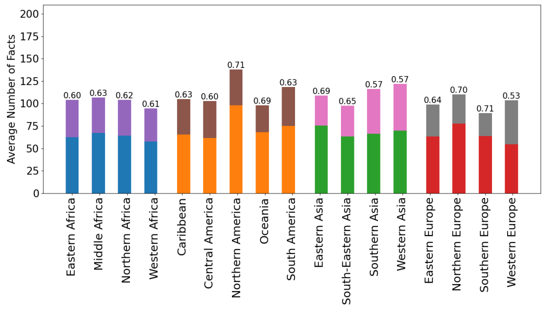
We also note on the bottom row of Figure (9) that the correlation between FActScore evaluated in English and FActScore evaluated in other languages for the texts are very high; and all of them are very closely correlated with human-judgement.

A.3 Additional Results

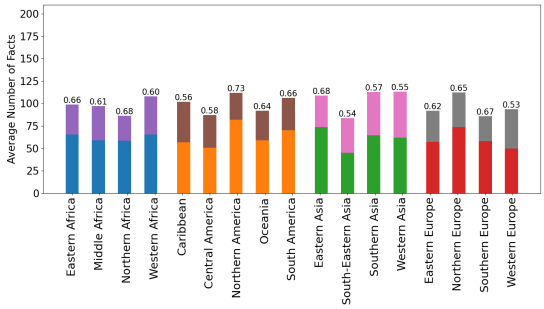
Regional Breakdown (GPT4): Figure 10 shows the regional breakdown for the other languages that were not included in Figure 7.



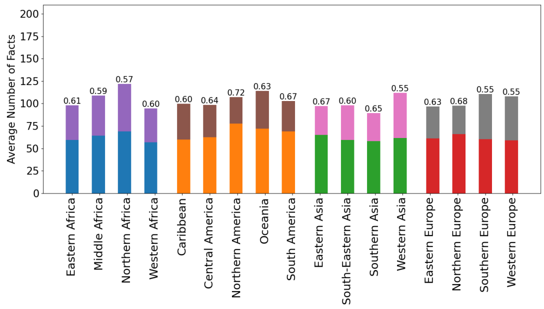
(a) Spanish



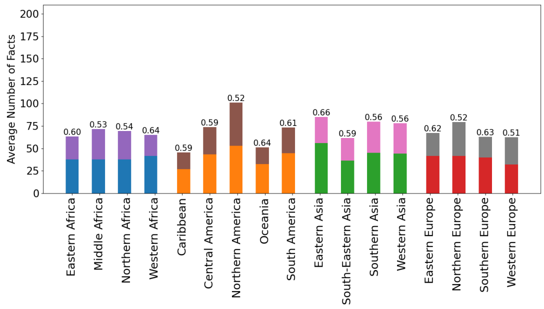
(b) German



(c) Swahili



(d) French



(e) Bengali

Figure 10: Fine-grained Geographic Distribution for Remaining Languages