Multilingual Hallucination Gaps in Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used as alternatives to traditional search engines given their capacity to generate text that resembles human language. However, this shift is concerning, as LLMs often generate hallucinations—misleading or false information that appears highly credible. In this study, we explore the phenomenon of hallucinations across multiple languages in free-form text generation, focusing on what we call *multilingual hallucination gaps*. These gaps reflect differences in the frequency of hallucinated answers depending on the prompt and language used. To quantify such hallucinations, we used the FACTSCORE metric and extended its framework to a multilingual setting. We conducted experiments using LLMs from the LLaMA, Qwen, and Aya families, generating biographies in 19 languages and comparing the results to Wikipedia pages. Our results reveal variations in hallucination rates, especially between high- and low-resource languages, raising important questions about LLM multilingual performance and the challenges in evaluating hallucinations in multilingual free-form text generation.

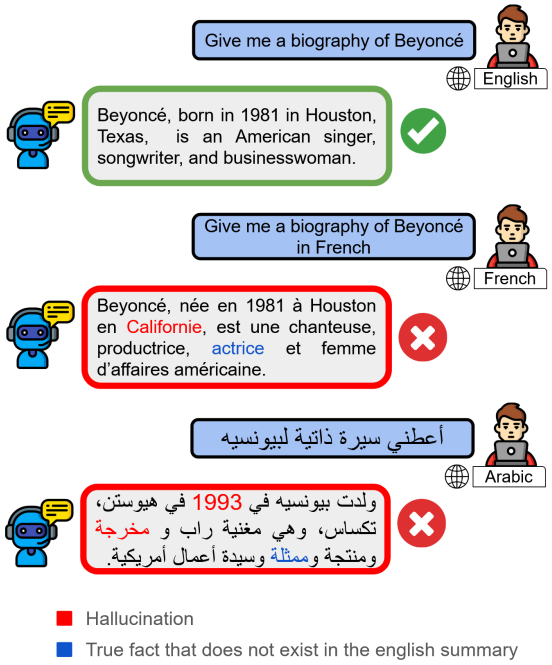


Figure 1: Example of Factual Hallucinations Gaps between languages

1 Introduction

Since the public release of ChatGPT, large language models (LLMs) have gained popularity. They are increasingly being integrated into or even replacing traditional search engines, such as the LLaMA model for Meta mobile applications or Gemma for Google. This trend shows an increasing reliance on LLMs as sources of knowledge, due to their ability to generate human-like text. However, such use is concerning as LLMs tend to produce hallucinations.

A hallucination occurs when a LLM generates *false* content (Rawte et al., 2023) with respect to a specific *reference*. Based on the reference type, hallucinations can be classified as follows (Zhang et al., 2023c): input-conflicting, where the generated content contradicts the user's input; context-conflicting, where it contradicts earlier outputs from the model; and fact-conflicting, where it contradicts established external knowledge. This work focuses exclusively on fact-conflicting hallucinations.

Understanding and tackling the issue of hallucinations in LLMs bring unique challenges. For example, detecting hallucinations is inherently difficult as they often appear highly credible. The wide range of tasks that LLMs are applied to also adds to the complexity, making it harder to comprehensively evaluate and mitigate hallucinations across different applications (Zhang et al., 2023c).

Besides these well-known and investigated issues, hallucinations are also not produced in the same way depending on the prompt fed to the model. We introduce the concept of *multilingual hallucination gaps*, which refers to variations in the proportion of hallucinated outputs generated in response to prompts in different languages.

Measuring these gaps can reveal that prompts in certain languages are more likely to induce hallucinations than others, which can significantly impact the reliability and trustworthiness of LLMs, especially in low-resource languages.

Previous work (Hong et al., 2024; Lin et al., 2022) focused on measuring hallucinations through benchmarks that require human annotations, which can be costly and hard to scale for multilingual LLMs. These benchmarks are also not suited for a free-form text generation setup. As a result, an automated evaluation pipeline becomes highly desirable. Statistical measures like ROUGE fail to capture semantic variations (Sellam et al., 2020) while NLI-based approaches transfer poorly to these tasks (Falke et al., 2019). Among LLM-based methods, Chen et al., 2024 proposed an eigenvalue metric to measure self-consistent hallucination. However, measuring this metric is costly and might not be suitable for the evaluation of free-form generated text across multiple languages.

Consequently, we explore FACTSCORE (Min et al., 2023), a different LLM-based method to evaluate hallucinations. In particular, the FACTSCORE metric uses an LLM to fact-check outputs of other LLMs against a knowledge source. As the FACTSCORE was only developed and tested on English text, we extend the methodology to encompass different languages by comparing toknowledge sources in various languages and leveraging translation.

We evaluate LLMs from the LLaMA, Qwen and Aya families. We prompt them to generate biographies in 19 different languages and we then compute the FACTSCORE metric for each answer by comparing it to an external knowledge source, Wikipedia for this project. The computation is done through three different experimental setups. We finally analyzed the results with respect to the target language, to the experimental setup and to the LLM used for text generation. Our results show gaps in the FACTSCORE metric distribution across the prompt languages, particularly between high, medium, and low-resource languages.
Our main contributions are:

1. Extending the FACTSCORE framework to a multilingual setting to quantify hallucination gaps across languages, with a focus on the disparities between high-resource and low-resource languages;

2. Evaluating a range of open-source and multilingual models to investigate improvements associated with different architectures and model sizes;

3. Assessing the robustness of the FACTSCORE framework across knowledge sources, prompt languages, and prompt templates.

2 Related work

Evaluating hallucinations

Previous research has concentrated on evaluating, explaining, and mitigating hallucinations in language models (Ji et al., 2023; Zhang et al., 2023c). All these efforts have been focused on detecting hallucinations in English-generated text exclusively.

There are several human-annotated benchmarks available for this purpose, including those compiled in the unified benchmark on HuggingFace by Hong et al. (2024). Since these benchmarks rely on human annotation, they usually focus on short answers and are time-consuming to create, making them ill-suited for evaluating multilingual hallucinations in free-form text generation.

Automatic metrics for measuring hallucinations encompass statistical and model-based ones (Ji et al., 2023), many of which draw inspiration from summarization evaluation. Statistical metrics like ROUGE can only handle lexical information and fail to deal with syntactic or semantic variations (Sellam et al., 2020). NLI-based approaches are robust to lexical variability, but NLI models transfer poorly to abstractive summarization (Falke et al., 2019) and struggle to locate specific errors in generated content. Faithfulness Classification metrics (Liu et al., 2022) address this issue, but they rely heavily on English-annotated datasets.

Among LLM-based methods, Chen et al. (2024) proposed an eigenvalue-based metric for detecting self-consistent hallucinations. However, this approach is not well-suited for free-form text generation, where repeated prompts can produce different, yet correct, responses, leading to lower scores despite valid outputs. More relevant to long-form text generation are the methods proposed by Min et al. (2023) and Farquhar et al. (2024), both of which

decompose answers into atomic facts. Min et al. (2023) employs a LLM to fact-check these facts against a knowledge source, while Farquhar et al. (2024) uses semantic entropy probabilities. In this study, we adopt the FACTSCORE (Min et al., 2023) approach, computationally less expensive.

Multilingual LLM

Several studies have focused on evaluating language generation models within a multilingual framework. Some of these datasets include M3Exam (Zhang et al., 2023a) for performance on human exam and Flores-101 (Goyal et al., 2022) for translation abilities. For the performance of the LLMs we used on these datasets, refer to Annex A. We can note that these evaluation metrics are still not consistently disclosed in technical reports or widely-recognized benchmarks. Despite covering a broad range of applications, these datasets do not cover hallucinations. However, they do provide evidence that LLMs exhibit different performance across different languages, which serves as motivation for our work.

An additional open research question concerns how multilingual abilities in these models are acquired (Zhang et al., 2023b; Wendler et al., 2024), as some models demonstrate proficiency in languages that are not officially supported. This observation motivated our decision to test models across a wide range of languages, even those not explicitly supported.

Hallucination metrics for multilingual generation

Kang et al. (2024) examine automatic hallucination detection metrics across different languages, including ROUGE, Named Entity Overlap, and the NLI-based SUMMAC score. Their findings show that these metrics do not correlate. Previous studies have suggested that these metrics may not be reliable for assessing hallucinations (Ji et al., 2023), which motivates our investigation into LLM-based metrics, specifically the FACTSCORE metric, to evaluate hallucinations across languages on a range of open-source models.

The most related work to ours is Shafayat et al. (2024), as the authors also study how to extend the FACTSCORE metric to a multilingual context. However, their methodology revolves around prompting in the original language and then translating generated content before assessing factuality. We broaden our investigation by adding an experiment that prompts in English while requesting answers in another language, as well as another experiment that directly compares generations to the original language Wikipedia page. Further, we investigate the reliability of this choice of knowledge source. We also explore a wider range of languages, a different set of entities beyond politicians, as well as multilingual open-source models instead of ChatGPT. Additionally, our work critically examines the robustness of the metric itself and identifies areas for improving its reliability.

3 Measuring factuality

To evaluate factual hallucinations in multilingual free-form text generation, we use the FACTSCORE metric (Min et al., 2023). This metric is particularly suited for our goal because it offers an intuitive, automated evaluation pipeline that can be easily adapted to different languages. By breaking down responses into atomic facts, FACTSCORE not only provides a more precise measure of factuality but also provides two key pieces of information: the factuality rate and the number of facts in the response.

Let's suppose we have a response $\mathcal{R}$ generated by an LLM, hereinafter referred to as LM_{SUBJ}. The FACTSCORE metric for this response $\mathcal{R}$ then consists of the following steps:

1. Decompose $\mathcal{R}$ into a set of atomic facts $\mathcal{A}(\mathcal{R})$. An atomic fact is a short sentence conveying a single piece of information. This is achieved by prompting an LLM, hereinafter referred to as LM_{EVAL}, to "Please breakdown the following sentence into independent facts" after showing it some decomposition examples.

2. Compare each fact $a \in \mathcal{A}(\mathcal{R})$ with an external knowledge source $\mathcal{C}$. To do so, we retrieve the proper passage from the source $\mathcal{C}$. We then construct a prompt by concatenating the retrieved passage, the given atomic fact and "True or False?". We then feed this prompt to an LLM. We use the same LLM that was used to decompose atomic facts, the LM_{EVAL}. The answer gives us:

$$\text{Supported}(a, \mathcal{C}) = \mathbb{1}\{a \text{ supported by } \mathcal{C}\}$$

3. (Optional) Add a length penalty p depending on a hyperparameter γ if our response $\mathcal{R}$ does

not contain enough facts:

$$p = \exp\left(\frac{1-\gamma}{|\mathcal{A}|}\right) \quad \text{if } |\mathcal{A}(\mathcal{R})| \leq \gamma$$

For instance, without applying this penalty, a response containing only one correct fact would receive a 100% FACTSCORE score, while a response with hundreds of facts, 99 of which are accurate, would get 99%. We set the default parameter to $\gamma = 10$, meaning responses with fewer than 10 facts are subject to a penalty.

4. Compute the FACTSCORE $\mathrm{F}(\mathcal{R}, \mathcal{C})$:

$$\mathrm{F}(\mathcal{R}, \mathcal{C}) = \frac{p}{|\mathcal{A}(\mathcal{R})|} \sum_{a \in \mathcal{A}(\mathcal{R})} \text{Supported}(a, \mathcal{C}) \tag{1}$$

4 Methodology

In this section, we explain and discuss the experimental settings, i.e. the choices of $\mathrm{LM}_{\mathrm{EVAL}}$, $\mathrm{LM}_{\mathrm{SUBJ}}$, content to be generated and knowledge source. We then detail the experimental process.

4.1 Experimental settings

Experiments are built following the methodology of the FACTSCORE paper (Min et al., 2023). We prompt a $\mathrm{LM}_{\mathrm{SUBJ}}$ to generate content in different languages and compute the FACTSCORE metric [1] with an $\mathrm{LM}_{\mathrm{EVAL}}$ for each answer, by comparing it to an external knowledge source, specifically Wikipedia for this project.

Choice of the $\mathrm{LM}_{\mathrm{SUBJ}}$

Our objective is to evaluate open-source models with strong multilingual capabilities and developed in various countries. We also want to include for each model at least two different sizes to assess the impact of model size on the FACTSCORE. The final choice was set on LLaMA-3 (8B and 70B parameters), Aya-23 (8B and 35B parameters) and Qwen7 (7B and 72B parameters). Refer to Annex A for more details on the LLMs chosen. In the rest of the paper, we will refer to these models without specifying their versions.

Choice of the $\mathrm{LM}_{\mathrm{EVAL}}$

To reduce bias in the evaluation process, we opt to use a different $\mathrm{LM}_{\mathrm{EVAL}}$ than the $\mathrm{LM}_{\mathrm{SUBJ}}$ models selected for the study. We use Mistral-7B-Instruct-v0.3 as the $\mathrm{LM}_{\mathrm{EVAL}}$ and first compute the error rate and $\mathrm{F1}_{micro}$ metrics (as detailed in Min et al., 2023) by computing FACTSCORE with Mistral on human annotated data. In Min et al., several methods are employed:

- No-context LM: Prompt "<atomic-fact> True or False?";
- Retrieve→LM: Retrieve passages from a knowledge source, concatenate these with the atomic fact and "True or False?" and prompt the concatenated result;
- Nonparametric Probability (NP): Mask each token in the atomic fact, calculate its likelihood with a nonparametric masked LM, average probabilities, and make a prediction based on thresholding;
- Retrieve→LM + NP: Assign "Supported" only if both Retrieve→LM and NP assign "Supported".

Based on their findings, we limit our experiments to the methods involving retrieval, comparing both Retrieve→Mistral and Retrieve→Mistral+NP. The results, presented in Table 4 in Annex B, show that Mistral performs competitively compared to the models used in Min et al., validating our choice of Mistral as $\mathrm{LM}_{\mathrm{EVAL}}$. However, unlike with the model Inst-LLaMA, adding NP did not enhance performance. We will thus use the Retrieve→Mistral method for all subsequent experiments.

Prompts

The FACTSCORE metric can be applied to any task, provided an appropriate knowledge source is available. We choose biographies because their factuality is easier to assess, as they generally include verifiable details such as birth dates and significant events, and they cover a wide range of nationalities. Besides, Wikipedia offers a multilingual knowledge source for this task, with biographies available in multiple languages.

We now have to choose the people whose biographies we will ask for and in which languages. For language selection, our goal is to cover a range of languages that includes high-resource, medium-resource, and low-resource languages. We define languages categories, i.e. high, medium and low, based on their data ratios from the Common Crawl corpus[1], drawing inspiration from Lai et al.,

[1] commoncrawl.github.io/statistics/languages

2023. For example, Spanish and Chinese are high-resource languages, Persian and Hindi fall under medium-resource, and Tamil and Swahili are examples of low-resource languages.

We also want to include as many language families as possible, while keeping the world's most widely spoken languages. Since we will be using Wikipedia as the knowledge source, it is also important to ensure that all selected languages have sufficient Wikipedia coverage, which we measure by the number of Wikipedia pages available in each language. Table 5 in Annex C provides details on the 19 languages chosen, including the key statistics used to perform the selection. We do not take into account the languages supported by the LM_{SUBJ} in this selection, as some models demonstrate proficiency in languages that are not officially supported (Zhang et al., 2023b; Wendler et al., 2024). We then proceed to select a set of notable figures with Wikipedia pages available in all these languages, resulting in a list of 485 individuals. This selected set of entities is interestingly biased. Figures 6a and 6b in Annex C illustrate their top 15 countries of citizenship and languages spoken, respectively. The data distribution is largely skewed towards the American citizenship and the English language.

Wikipedia as a knowledge source

We retrieve Wikipedia summaries in every language for the 485 notable figures to serve as the knowledge source. Recognizing that Wikipedia content quality can vary across languages, we start by comparing these summaries to each other. For a given language, we translate the Wikipedia pages with GPT-4o-mini and we compute the FACTSCORE comparing to the English Wikipedia. We choose the English Wikipedia because the FACTSCORE indicates whether an atomic fact is supported by the knowledge source, rather than giving information on its presence in the knowledge source. Our assumption is that English Wikipedia is more comprehensive than other language versions. We also compute FACTSCORE for the English Wikipedia pages, i.e. comparing them against themselves. Results are presented in Figure 2.

We can directly see that the evaluator is not perfect. Indeed, comparing the English Wikipedia to itself does not always yield a 100% FACTSCORE, even if it typically falls within a high range of 90-100%. For the other languages, cross-checking with English yields much lower FACTSCORE. This

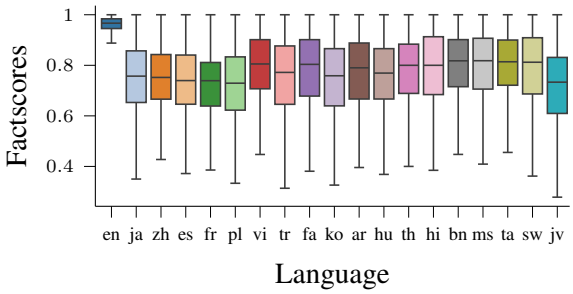


Figure 2: FACTSCORE distribution for Wikipedia pages

suggests that content in different languages can either contradict or diverge from what is found in English Wikipedia. These observations are important for the rest of our analysis. We will compare a generated biography with both its original language Wikipedia version and the English one. This may give us insight on how the LLM captures and represents knowledge in a multilingual setting, depending on whether the FACTSCORE is different when using these two different knowledge sources. It is also important to consider both for a more precise analysis.

4.2 Experiments

In this section, we outline the experimental process, detailing the steps from data generation to FACTSCORE evaluation, with intermediate sanity checks.

Data generation

To ensure robustness in measuring the hallucination gaps, we use three established prompt templates for generating biographies from the literature: "Tell me a biography of {}", "Give me a biography of {}" and "Please give me a biography of {}".
We use two prompting methods:

- `lang`-prompt: Translate the template in the target language `lang` and use the translated prompt;
- `en`-prompt: Use the English template but add "in {`lang`}" at the end (e.g., "Tell me a biography of {} in French").

We use these methods for our six models LM_{SUBJ}, resulting in a total of 12 text generation setups. Each setup produces $485 \times 19 \times 3$ generated responses.

Sanity checks

Once the biographies are generated, we carry out some sanity checks to ensure we have good enough

generated answers to compute FACTSCORE with. For each answer, we verify that it is in the correct target language with the module `py3langid`. We set a threshold of 20 distinct words to remove outputs with same words repeated infinitely. We do not compute FACTSCORE for generated answers that do not pass sanity checks.

FACTSCORE evaluation

We perform three experiments depending on the knowledge source and prompting method used. We will refer to these experiments as (prompt language, Wikipedia language):

1. (`lang`, `lang`): Compare the response produced with a `lang`-prompt to the `lang` Wikipedia page;
2. (`lang`, `en`): Translate the response produced with a `lang`-prompt to English and compare to the English Wikipedia page;
3. (`en`, `en`): Translate the response produced with an `en`-prompt to English and compare to the English Wikipedia page.

Recall that we generated three answers for each notable figure and language. We thus compute the FACTSCORE score for each answer and then average these three scores to obtain a single final score per entity. We also report the standard deviation. This process is repeated across all three experiments: (`lang`, `lang`), (`lang`, `en`) and (`en`, `en`).

Translation

All translation steps, including prompt translations, generations, and Wikipedia pages, were performed with GPT-4.

5 Results

5.1 Sanity checks

The percentages of generated answers that passed sanity checks for each LM_{SUBJ} are shown in Table 6 in Annex D. These initial results already demonstrate whether the LM_{SUBJ} is able to generate multilingual text, including in languages stated as not supported. Figure 3 shows the percentage of generated responses produced in the correct target language, for each LM_{SUBJ} and prompt setting. As expected, the models generally perform better in generating text in high-resource languages compared to low- and very low-resource ones. For very low-resource languages like Javanese (jv) and Malay (ms), the models rarely generate text in the correct target language.

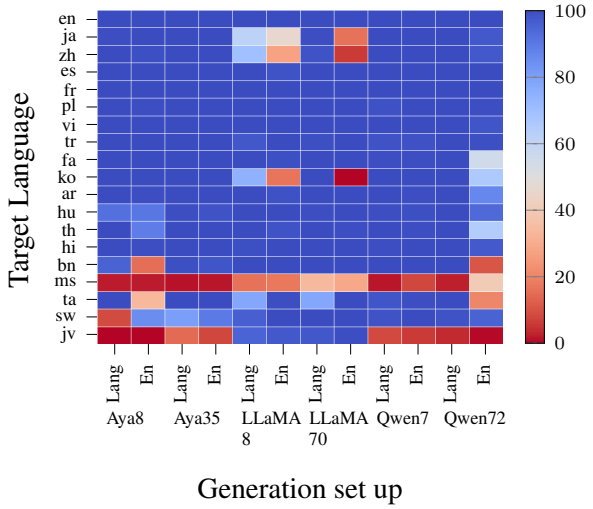


Figure 3: Percentage of correct language produced per target language and generation set up

The best performance overall is achieved with Qwen7 both in English (`en`) and original language (`lang`) prompting. We can note that for all models prompting in English tend to decrease the percentages of generated responses in the correct target language. Interestingly, for the LLaMA and Qwen families of models, increasing the number of parameters does not always lead to better performance, especially when prompted in English.

We also observe poorer performance in the Japanese, Chinese, and Korean languages for the LLaMA models compared to others. When examining generated answers that failed the sanity checks, we notice that the LLaMA models often produced Romaji (i.e., Japanese writing in Roman characters) instead of Kanji (i.e., Japanese writing using Chinese characters).

5.2 FACTSCORE

Hallucination rates differ across languages. Table 1 presents the average results across all models and entities, grouped by language category. We observe multilingual hallucination gaps, with both factuality and the number of facts decreasing as the language resource level decline. Due to high standard deviations within these categories, we examine the FACTSCORE distribution at a finer level, for each language.

Figure 4 shows these distributions across all models and entities, broken down by language and experiment. See Annex F for the same results per

Language Category	FACTSCORE (%)			# of Facts		
	(en, en)	(lang, en)	(lang, lang)	(en, en)	(lang, en)	(lang, lang)
Very-High	73.7 (± 10.1)	71.8 (± 9.9)	70.3 (± 9.7)	79	82	103
High	70.2 (± 12.6)	69.3 (± 13.4)	58.5 (± 16.4)	68	73	65
Medium	64.7 (± 16.0)	61.3 (± 19.5)	47.8 (± 19.4)	54	59	49
Low	56.9 (± 18.9)	47.6 (± 23.4)	44.4 (± 20.4)	38	34	53

Table 1: Mean FACTSCORE (± STD) and Mean number of facts by Language Category and Experiment for all models

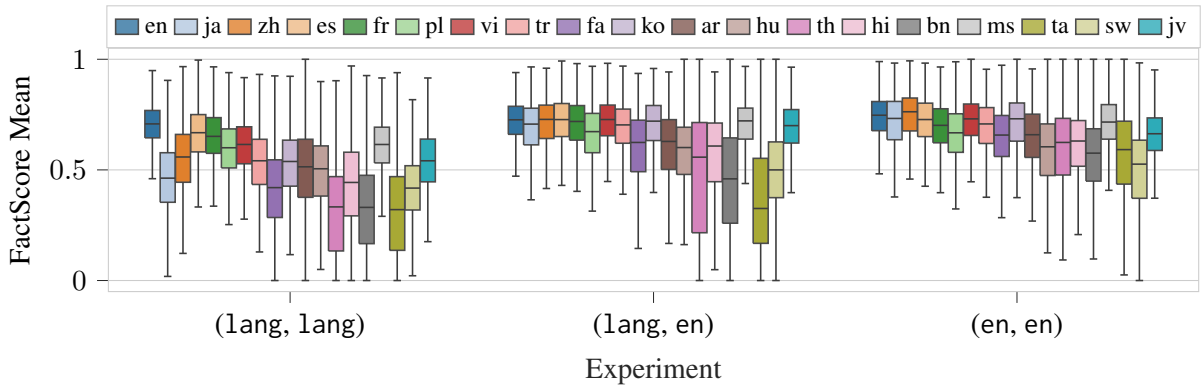


Figure 4: FactScore Mean distribution by Language and Experiment for all models

LM_{SUBJ}. Aside from Malay (ms) and Javanese (jv) in all experiments, and Japanese (ja) and Chinese (zh) languages in the (lang, lang) experiment, we can observe the same trend as FACTSCORE distributions are more spread out and shifted toward lower values as the language resource level decline. It should be noted that after filtering out unsane answers that were in the incorrect language, we have less data points for the Malay and Javanese languages (see Figure 3).

On the contrary, only for the (lang, lang) experiment, Japanese and Chinese show distributions that are more spread out and shifted toward lower values compared to other high-resource languages. This raises the question of why different experimental setups influence the results.

Different pipelines show different results. Table 1 and Figure 4 highlight differences in results across the (lang, lang), (lang, en), and (en, en) experiments, showing that the choice of knowledge source and prompt language can influence the FACTSCORE outcomes.

Regarding the prompt language, we gain insights into the performance of the LM_{SUBJ}. As shown in Figure 3, models respond differently to prompts given in English versus the target language lang. When comparing (lang, en) and (en, en) — where the knowledge source remains the same but the prompt languages differ — we observe the highest results with the (en, en) setting. This raises concerns, as we would want models to respond accurately to prompts in original languages.

The impact of the knowledge source presents more significant challenges, as it directly affects the evaluator. In comparing (lang, lang) and (lang, en), where the prompt language remains the same but the knowledge source differs, we see a decrease in FACTSCORE for the (lang, lang) experiment, and more dispersed distributions. This trend may be attributed to the quality of Wikipedia pages in their original languages, as highlighted in Figure 2.

The performance of (lang, lang) compared to (lang, en) and (en, en) is also influenced by the respective multilingual capabilities of the LM_{EVAL} and the translator (GPT-4). The (lang, lang) experiment is the only one that involves prompting the LM_{EVAL} in languages other than English. We noticed that for most languages, while breaking down a generation into atomic facts, the LM_{EVAL} also translated these facts into English even when not instructed to do so. Addressing this behavior

Language Category	STD of FACTSCORE (%)		
	(en, en)	(lang, en)	(lang, lang)
Very-High	4.9	5.1	4.8
High	6.2	6.6	7.0
Medium	7.5	8.0	8.6
Low	8.8	10.2	9.3

Table 2: Standard deviation across the 3 prompt templates of FACTSCORE by Language Category and Experiment for all models

might improve results in the (lang, lang) setting. The (lang, en) approach seems to be the most suitable option, as it allows us to maintain prompts in their original languages.

FACTSCORE's robustness depends on the language. Table 2 present standard deviation of FACTSCORE when computed across the three prompt templates for each entity. We then take the average of these standard deviations by language category and experiment for all models and entities. The results show that as the language resource level decrease, the FACTSCORE standard deviation increases. This suggests that the FACTSCORE metric becomes less consistent when measured across the three different prompt templates, reflecting greater variability in the generated answers for low-resource languages.

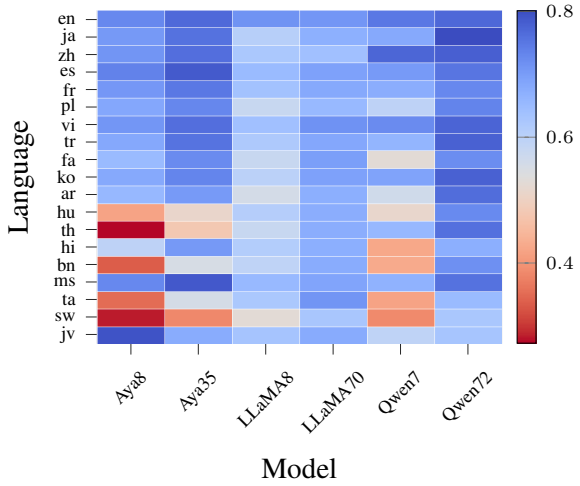


Figure 5: FACTSCORE per language and per model for the (en, en) experiment

The LM_{SUBJ} show different behaviors across languages. Figure 5 illustrates the mean FACTSCORE per model and per language for the (en, en) experiment. For results from the two other experiments, see Annex E. The model Qwen72 performs the best overall. This aligns with the models' multilingual performance on other benchmarks, where Qwen72 also ranks highest (see Table 3).

The models Aya8, Aya35 and Qwen7 show the largest discrepancies across languages, with poor results on low-resource languages. The LLaMA models, while showing more consistency across all languages, generally perform worse compared to other models, even in English. For the Aya family, the lower FACTSCORE scores align with the unsupported languages (see Annex A). For LLaMA and Qwen, we do not observe such behaviour. The Qwen models, that are primarily trained on both English and Chinese, exhibit the best performance in Chinese as expected.

Within each model family, a higher number of parameters leads to higher FACTSCORE.

6 Conclusion

Our research shows multilingual hallucination gaps in LLMs. In higher-resource languages, which have more extensive training data, these models show greater factual accuracy, whereas in low-resource languages, they tend to hallucinate more. This raises important concerns about the equitable performance of LLMs across different linguistic groups and the broader implications for fairness in AI technologies.

Our findings also indicate that model size and architecture influence these gaps. Larger models generally perform better but still show hallucinations in low-resource languages. Even models with strong multilingual capabilities hallucinate in such languages, and struggle to generalize effectively to unsupported languages, suggesting that simply increasing model size or expanding training data is not a complete solution.

Finally, our work raises concerns about adapting the FACTSCORE metric for multilingual contexts. Scores on the FActScore metric vary based on the chosen knowledge source, and translations are generated by the LM_{EVAL} within the pipeline. Future research could build on the metric developed by Farquhar et al. (2024), which does not rely on an external knowledge source but still uses an LM_{EVAL} that may show biased performance across languages. Additionally, a thorough comparison of LLM-based methods for free-form text generation could provide insights beyond human annotations.

Limitations

The FACTSCORE metric, while useful for automatically assessing factuality in generated text, has several limitations that need to be addressed. One major issue is its robustness, especially in a multilingual setting. For instance, calculating a FACTSCORE using the generated text itself as the knowledge source does not always yield a perfect 100% score, as shown on Figure 2, highlighting potential inconsistencies. Computing FACTSCORE can be quite resource-intensive, which can limit its widespread use. Besides, we can only verify *intrinsic* hallucinations (Ji et al., 2023) with this metric, i.e. when the generated content directly contradicts the reference. Future work could extend the metric to *extrinsic* hallucinations, when the generated output cannot be verified with the source reference (i.e., it is neither supported nor contradicted by the reference), to provide more insights. This would be useful for languages in which the Wikipedia coverage may be weaker than the English one.

The use of Wikipedia as a knowledge source also presents limitations. Some Wikipedia entries may not be fully accurate, and certain facts could be ambiguous. Wikipedia's coverage also varies significantly across languages, which can impact the effectiveness of the FACTSCORE metric in the (`lang`, `lang`) setting. Despite these issues, Wikipedia remains one of the most comprehensive public multilingual knowledge sources available.

Another limitation comes from potential biases in the FACTSCORE metric evaluation, as the computation is done by another LLM, the LM_{EVAL}. This is especially evident in a multilingual setup, as for the (`lang`, `lang`) experiment the evaluation relies on the performance of the LM_{EVAL} across languages. We assume equal performance across languages, which is not accurate in practice. For the other experiments, we rely on GPT4's performance in translation tasks which can also add variability. To address this, creating a human benchmark for assessing multilingual hallucination gaps could offer a more reliable and unbiased evaluation.

Finally, we only focus on biographies of a specific group of individuals. While we cover a diverse set of people, future work could explore how these gaps evolve when the LM_{SUBJ} are confronted with other tasks, for instance other types of articles on Wikipedia (e.g., scientific topics) or text about historical events whose knowledge source can be a collection of articles. However, for consistency with our experimental settings, these tasks would need to have multilingual knowledge sources for evaluation and less room for subjectivity.

Code availability

The code and data are accessible at the anonymized GitHub repository: `https://anonymous.4open.science/r/Multilingual_Hallucination_Gaps-7155`.

References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, and Pasquale Minervini. 2024. The hallucinations leaderboard–an open effort to measure hallucinations in large language models. *arXiv preprint arXiv:2404.05904*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea

Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-fact: Assessing multilingual llms' multi-regional knowledge using factscore. *Preprint*, arXiv:2402.18045.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023a. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023b. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023c. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

A Characteristics of the LM_{SUBJ}

Table 3 presents the models chosen as LM_{SUBJ}, as well as their characteristics and their multilingual performance on different benchmarks, as reported in LLaMA-3 (Dubey et al., 2024), Qwen2 (Yang et al., 2024) and Aya-23 (Aryabumi et al., 2024) technical reports.

LlaMA officially supports 8 languages (Dubey et al., 2024): English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai, although the underlying foundation model has been trained on a broader collection of languages.

The Aya models (Aryabumi et al., 2024) support only a limited set of languages that intersect with ours: English (en), Japanese (ja), Chinese (zh), Spanish (es), French (fr), Polish (pl), Vietnamese (vi), Turkish (tr), Persian (fa), Korean (ko), Arabic (ar), and Hindi (hi).

The Qwen models (Yang et al., 2024) are the ones covering the most languages of our dataset, with the exception of Hungarian (hu), Tamil (ta), Swahili (sw) and Javanese (jv).

	Aya-23		LLaMA-3		Qwen2	
# Parameters	8B	35B	8B	70B	7B	72B
Architecture	Dense		Dense		Dense	
Developer	Cohere		Meta-AI		Alibaba	
Origin	Canada		USA		China	
Exam[1]	48	58	52	70	60	**78**
Understanding[2]	-	-	69	80	72	**81**
Mathematics[3]	37	47	36	67	57	**87**
Translation[4]	37	**40**	32	38	32	38

[1] mMMLU [2] BELEBELE, XCOPA, XWinograd, XStoryCloze, PAWS-X [3] MGSM [4] Flores-101

Table 3: Characteristics and Multilingual Performance of the Large Language Models chosen as LM_{SUBJ}. The **bold** values indicate the best performance for each multilingual benchmark.

B Validation of the LM_{EVAL}

Table 4 presents validation results for the FACTSCORE estimated by Mistral compared to human annotated scores. We also include results of the two best models of Min et al..

C Languages and People Dataset

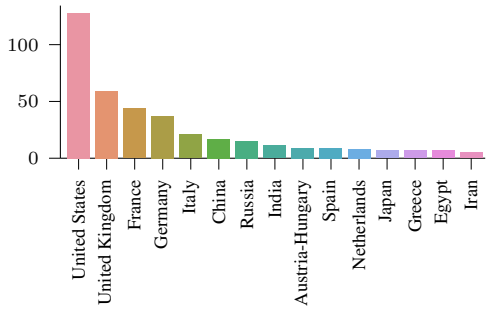
Table 5 present the 19 selected languages along with their characteristics used for the selection process.

Figures 6a and 6b illustrate the top 15 countries of citizenship and languages of the entities. It is important to note that a figure may have multiple citizenships or speak multiple languages. We can observe that the data distribution is largely skewed towards the American citizenship and the English language.

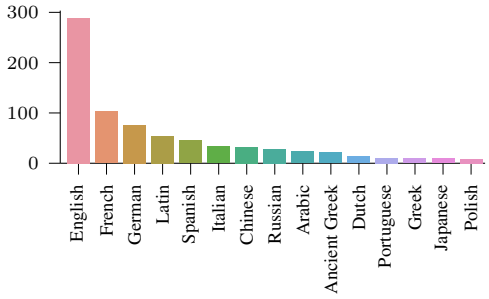
D Sanity checks

Table 6 show the percentages of generated answers that passed sanity checks for each LM_{SUBJ}. Overall, the percentages are similar across across all LM_{SUBJ}.

For the LLaMA models, we initially encountered lower percentages of sane answers (22%) for the `lang` prompt. When prompted in another language than English, these models failed to understand that they had to respond in that language and not in English. To address this issue, we translated the entire English prompt, with the added directive "in {`lang`}". For example, for the French language, the prompt was adjusted to "Donne-moi une biographie de {} en Français." instead of only "Donne-moi une biographie de {}.". This adjustment significantly improved the percentage of sane answers, raising it from 22% to 88%. No additional output regeneration was needed for the other models, as they already produced satisfactory percentages of sane responses.



(a) Top 15 citizenship countries



(b) Top 15 languages spoken

Figure 6: Citizenship and language statistics of the entities

	en-prompt	lang-prompt
LLaMA-3 8B	83.68	88.50
LLaMA-3 70B	81.55	94.96
Qwen 7B	89.76	89.08
Qwen 72B	74.48	89.44
Aya 8B	70.79	80.42
Aya 35B	84.52	83.85

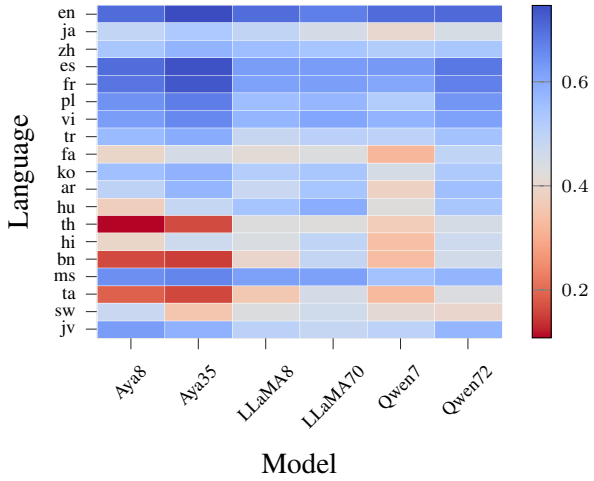
Table 6: Percentages of generated answers kept after sanity checks per generation setup

E FACTSCORE results per experiment

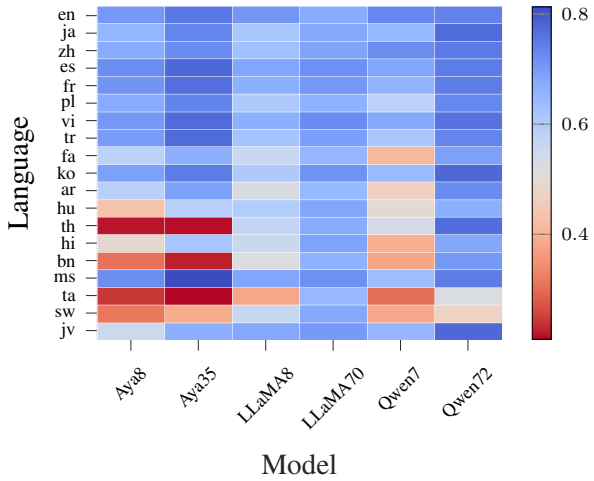
Figures 7a and 7b present the mean FACTSCORE per model and per language for the (`lang`, `lang`) and (`lang`, `en`) experiments. We observe the same trends across LM_{SUBJ} as for the (`en`, `en`) experiment.

$\mathbf{LM_{EVAL}}$	SUBJ: InstGPT		SUBJ: ChatGPT		SUBJ: PPLAI	
	ER	F1	ER	F1	ER	F1
Always Not-supported	0.42	71.4	0.58	58.3	0.80	30.9
Retrieve→ChatGPT	0.14	**86.2**	0.18	68.5	**0.09**	54.9
Retrieve→Inst-LLaMA+NP	0.22	73.3	0.29	60.2	0.36	39.6
Retrieve→Mistral	**0.09**	85.4	**0.11**	73.5	0.11	**58.4**
Retrieve→Mistral+NP	0.11	84.8	0.12	**74.0**	0.17	56.3

Table 4: Results on Error Rate (ER) and $F1_{micro}$ (F1) for the FACTSCORE estimated by Mistral compared to human annotated scores. We also include results of the 2 best models of Min et al.. The **bold** values indicate the best performance for each metric.



(a) (lang, lang) experiment



(b) (lang, en) experiment

Figure 7: FACTSCORE per languages and per model

F FACTSCORE results per $\mathbf{LM_{SUBJ}}$

We present in this section the FACTSCORE results for every LM_{SUBJ} instead of the average over all models. For every LM_{SUBJ} we present both the table of FACTSCORE and number of facts averaged across language categories and the boxplot figures of distribution for each language.

	Family	Branch	CC Ratio	Worldwide Speakers (in millions)	Wikipedia pages (in thousands)
		Very High Resource Language			
English (en)	Indo-European	Germanic	46.45	1,456	6,832
		High Resource Languages			
Japanese (ja)	Japonic	-	5.09	123	1,419
Chinese (zh)	Sino-Tibetan	Sinitic	4.17	1,138	1,423
Spanish (es)	Indo-European	Romance	4.55	559	1,957
French (fr)	Indo-European	Romance	4.64	310	2,616
Polish (pl)	Indo-European	Balto-Slavic	1.76	41	1,620
		Medium Resource Languages			
Vietnamese (vi)	Austroasiatic	Vietic	0.99	86	1,294
Turkish (tr)	Turkic	Oghuz	0.99	90	608
Persian (fa)	Indo-European	Iranian	0.67	79	1,004
Korean (ko)	Koreanic	-	0.65	82	672
Arabic (ar)	Afro-Asiatic	Semitic	0.59	274	1,235
Hungarian (hu)	Uralic	Hungarian	0.56	17	543
Thai (th)	Kra–Dai	Zhuang–Tai	0.41	61	165
Hindi (hi)	Indo-European	Indo-Aryan	0.18	610	162
		Low and Very-Low Resource Languages			
Bengali (bn)	Indo-European	Indo-Aryan	0.10	273	154
Malay (ms)	Austronesian	Malay	0.07	290	377
Tamil (ta)	Dravidian	Southern	0.04	87	166
Swahili (sw)	Niger-Congo	Bantu	0.008	72	80
Javanese (jv)	Austronesian	Malayo-Polynesian	0.002	68	73

Table 5: The 19 chosen languages with key statistics

Language Category	FACTSCORE (%)			# of Facts		
	(en, en)	(lang, en)	(lang, lang)	(en, en)	(lang, en)	(lang, lang)
Very-High	72.6 (± 12.8)	70.2 (± 12.5)	70.1 (± 12.6)	67	75	82
High	70.6 (± 13.9)	68.3 (± 15.4)	60.8 (± 19.4)	68	80	54
Medium	59.6 (± 20.7)	55.0 (± 23.4)	43.7 (± 23.9)	51	58	36
Low	32.9 (± 23.7)	28.4 (± 17.3)	41.7 (± 26.0)	26	19	36

Table 7: Mean FACTSCORE (± STD) and Mean number of facts by Language Category and Experiment for Aya 8

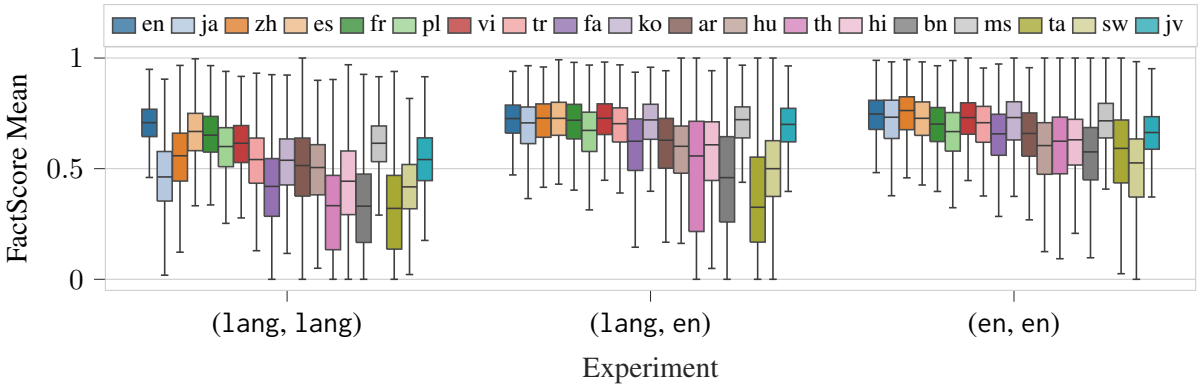


Figure 8: FactScore Mean distribution by Language and Experiment for Aya 8

Language Category	FACTSCORE (%)			# of Facts		
	(en, en)	(lang, en)	(lang, lang)	(en, en)	(lang, en)	(lang, lang)
Very-High	76.4 (± 9.8)	75.1 (± 9.6)	74.7 (± 9.6)	83	80	89
High	75.4 (± 11.1)	74.4 (± 12.6)	65.0 (± 17.7)	78	80	59
Medium	67.0 (± 17.1)	63.2 (± 23.4)	49.5 (± 23.2)	58	60	40
Low	52.2 (± 22.6)	28.3 (± 21.7)	38.0 (± 26.2)	26	21	37

Table 8: Mean FACTSCORE (± STD) and Mean number of facts by Language Category and Experiment for Aya 35

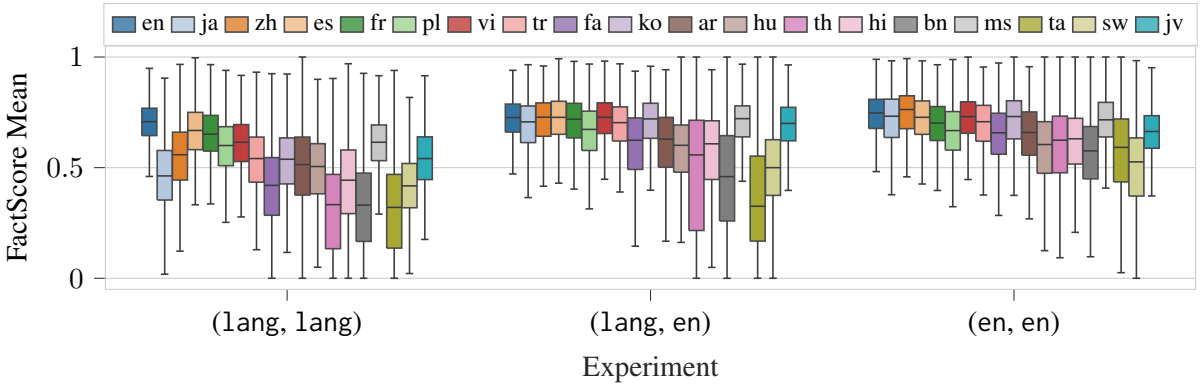


Figure 9: FactScore Mean distribution by Language and Experiment for Aya 35

Language Category	FACTSCORE (%)			# of Facts		
	(en, en)	(lang, en)	(lang, lang)	(en, en)	(lang, en)	(lang, lang)
Very-High	71.2 (± 9.0)	70.5 (± 8.5)	69.9 (± 8.1)	76	82	107
High	61.8 (± 11.9)	63.7 (± 13.8)	56.7 (± 13.7)	60	66	73
Medium	59.8 (± 11.6)	58.4 (± 16.2)	48.1 (± 16.5)	55	65	57
Low	59.7 (± 12.8)	54.3 (± 18.8)	46.0 (± 16.6)	39	43	56

Table 9: Mean FACTSCORE (± STD) and Mean number of facts by Language Category and Experiment for LLaMA 8

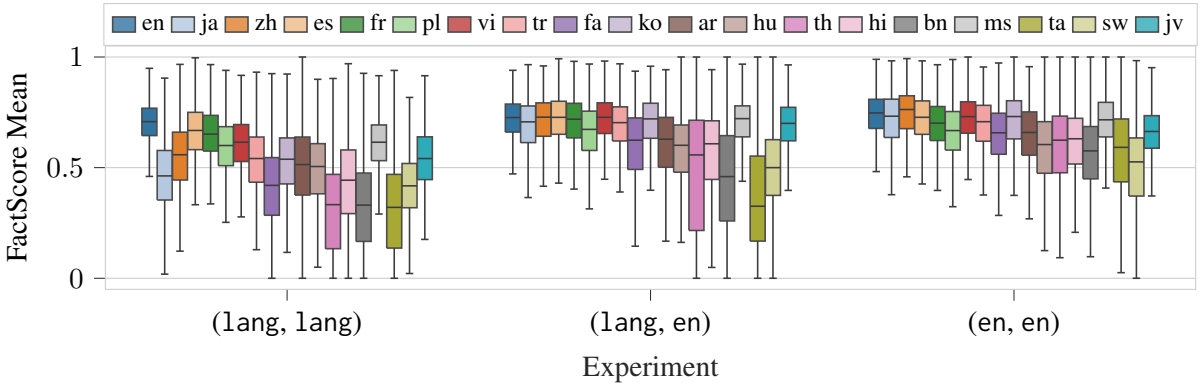


Figure 10: FactScore Mean distribution by Language and Experiment for LLaMA 8

Language Category	FACTSCORE (%)			# of Facts		
	(en, en)	(lang, en)	(lang, lang)	(en, en)	(lang, en)	(lang, lang)
Very-High	70.7 (± 8.7)	66.7 (± 8.1)	67.1 (± 7.9)	81	92	120
High	67.1 (± 10.7)	68.6 (± 11.5)	56.0 (± 13.4)	66	69	74
Medium	68.2 (± 9.8)	68.1 (± 11.5)	51.5 (± 15.4)	64	68	63
Low	67.3 (± 11.1)	67.4 (± 15.6)	49.8 (± 14.4)	45	48	66

Table 10: Mean FACTSCORE (± STD) and Mean number of facts by Language Category and Experiment for LLaMA 70

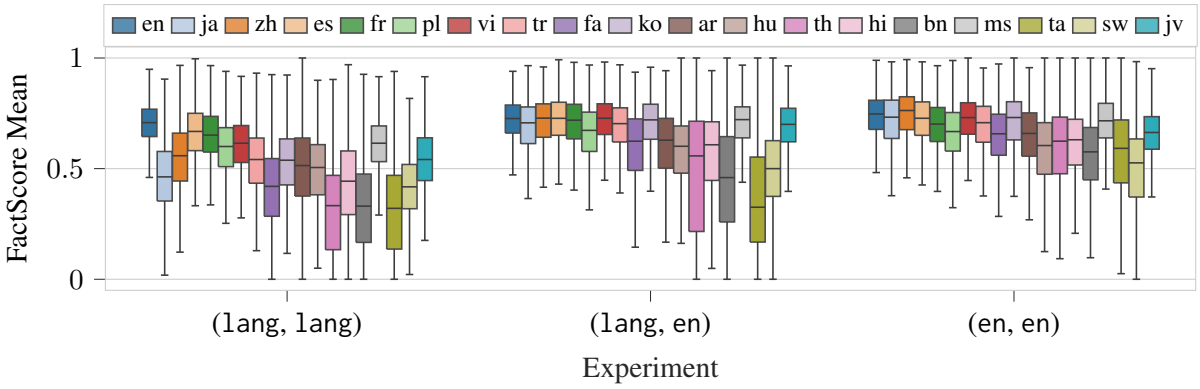


Figure 11: FactScore Mean distribution by Language and Experiment for LLaMA 70

Language Category	FACTSCORE (%)			# of Facts		
	(en, en)	(lang, en)	(lang, lang)	(en, en)	(lang, en)	(lang, lang)
Very-High	74.7 (± 9.2)	73.0 (± 9.2)	70.1 (± 9.4)	81	83	108
High	68.3 (± 11.7)	65.6 (± 12.2)	53.3 (± 15.0)	66	70	65
Medium	59.3 (± 15.2)	52.8 (± 17.5)	41.9 (± 17.5)	50	51	48
Low	43.3 (± 15.3)	37.8 (± 17.3)	42.2 (± 16.7)	36	32	62

Table 11: Mean FACTSCORE (± STD) and Mean number of facts by Language Category and Experiment for Qwen 7

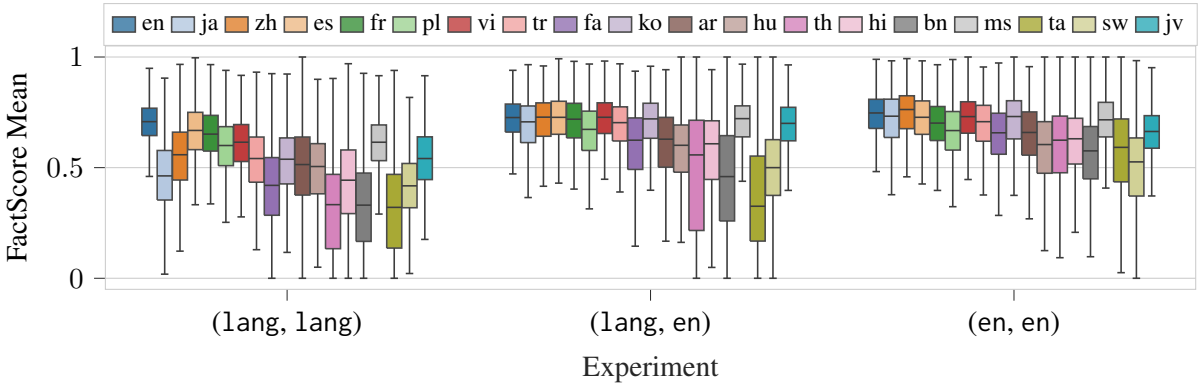


Figure 12: FactScore Mean distribution by Language and Experiment for Qwen 7

Language Category	FACTSCORE (%)			# of Facts		
	(en, en)	(lang, en)	(lang, lang)	(en, en)	(lang, en)	(lang, lang)
Very-High	76.6 (± 8.8)	73.5 (± 8.3)	70.2 (± 8.5)	84	86	109
High	75.8 (± 9.9)	74.7 (± 9.6)	59.3 (± 15.5)	66	71	62
Medium	74.4 (± 11.7)	72.3 (± 11.3)	52.2 (± 15.8)	48	52	49
Low	67.5 (± 13.8)	57.4 (± 19.7)	48.7 (± 15.8)	43	29	62

Table 12: Mean FACTSCORE (± STD) and Mean number of facts by Language Category and Experiment for Qwen 72

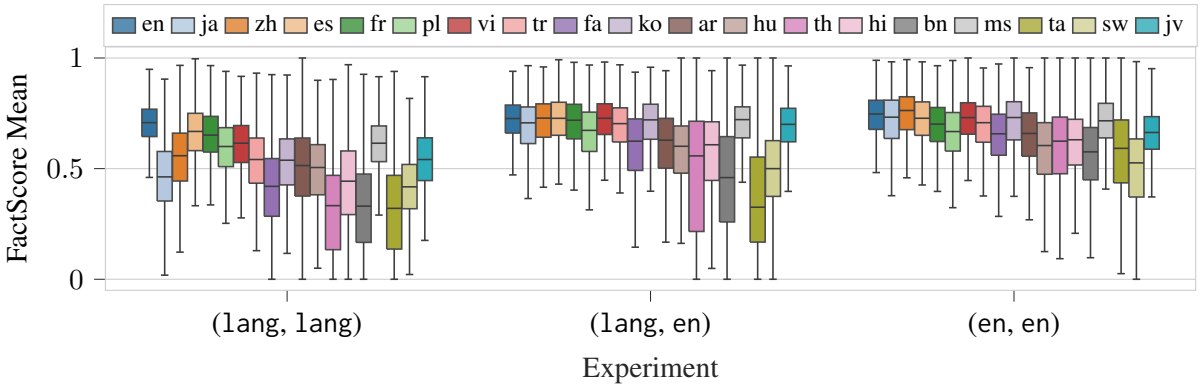


Figure 13: FactScore Mean distribution by Language and Experiment for Qwen 72