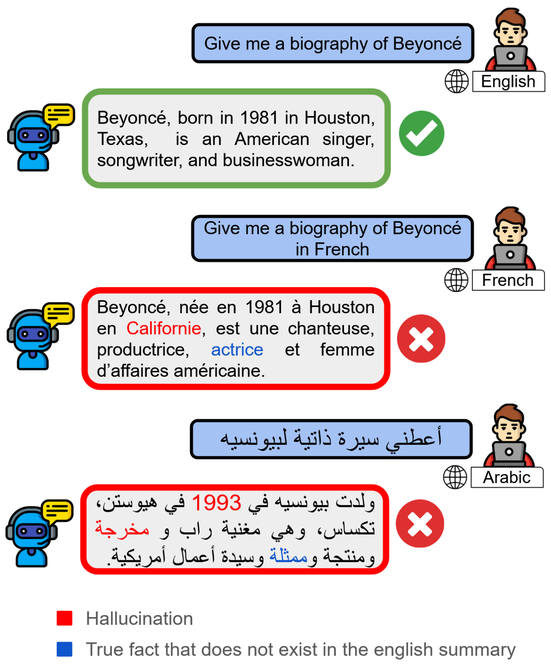
Multilingual Hallucination Gaps in Large Language Models

Cléa Chataigner [1 2] Afaf Taïk [1 3] Golnoosh Farnadi [1 2 3]

[1]Mila – Quebec AI Institute [2]McGill University [3]University of Montreal

Context and Contributions



1. Extending the FActScore framework to a multilingual setting to quantify hallucination gaps across 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.

FActScore metric [1]

1. **Data generation:** Generate a response $\mathcal{R}$ with a LM_{subj}
2. **Decomposition:** Split $\mathcal{R}$ into atomic facts with a LM_{eval}
3. **Fact-checking:** Compare each fact to an external knowledge source with a LM_{eval}
4. **Final score:**

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

Experimental setup

Task and Dataset

- Biographies for verifiable details
- 485 notable figures, 19 languages

Data generation

- LM_{subj}: LLaMA-3 (8B and 70B), Aya-23 (8B and 35B) and Qwen2 (7B and 72B)
- 3 prompt templates semantically equivalent
- 2 prompting methods: `lang`-prompt (directly in the target language) and `en`-prompt (in English, asking for an answer in the target language)

Sanity Checks

- Check the language with the module `py3langid`
- Threshold of 20 distinct words

FActScore evaluation

- (`lang`, `lang`) Compare the response produced with a `lang`-prompt to the `lang` Wikipedia page
- (`lang`, `en`) Translate the response produced with a `lang`-prompt to English and compare to the English Wikipedia page
- (`en`, `en`) Translate the response produced with an `en`-prompt to English and compare to the English Wikipedia page

Preliminary results

Wikipedia as a knowledge source

- Compare Wikipedia pages to the English ones using FActScore

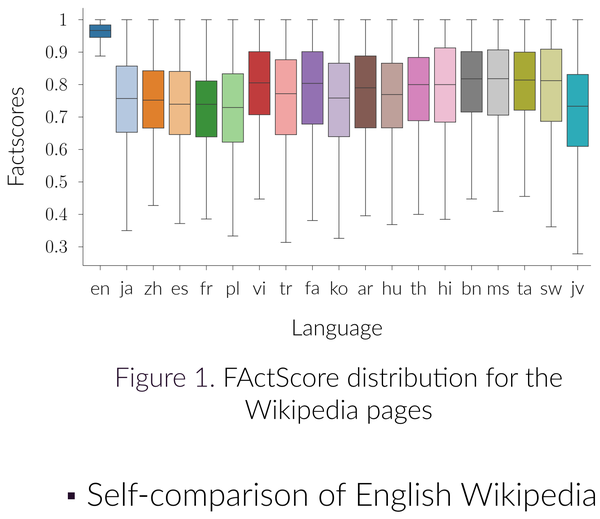


Figure 1. FActScore distribution for the Wikipedia pages

- Self-comparison of English Wikipedia doesn't yield 100% FActScore
- Wikipedia multilingual content can contradict or diverge from the English one

Language sanity check

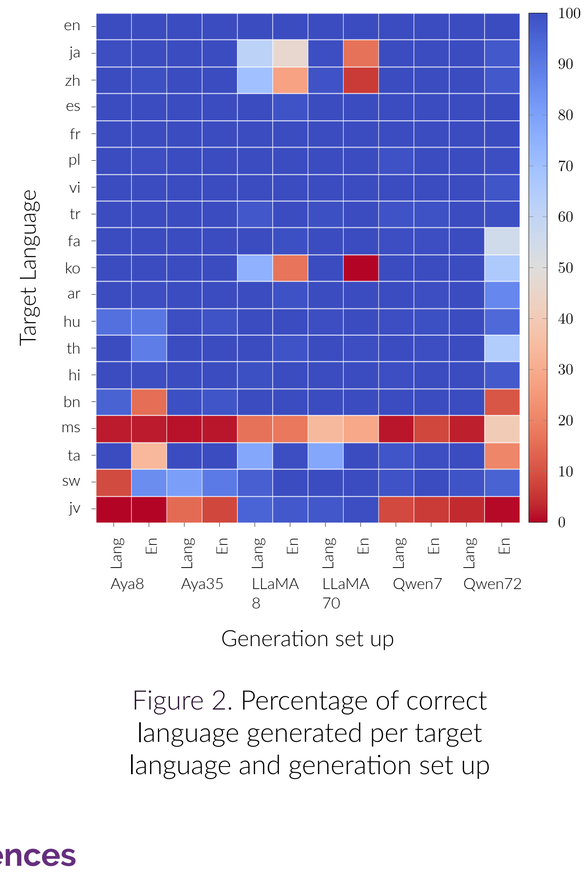


Figure 2. Percentage of correct language generated per target language and generation set up

FActScore results

- Hallucination rates differ across languages and different experimental pipelines show different results

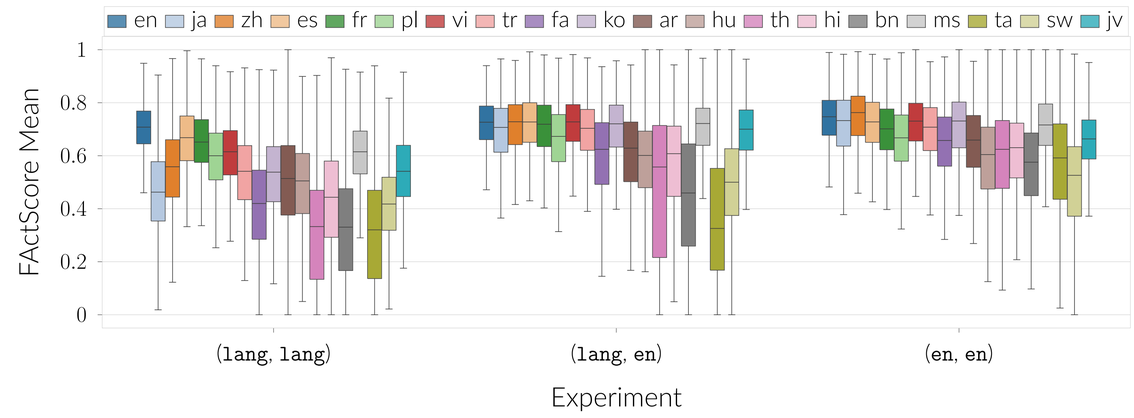


Figure 3. FActScore Mean distribution by Language and Experiment for all models

- The LM_{subj} show different behaviors across languages

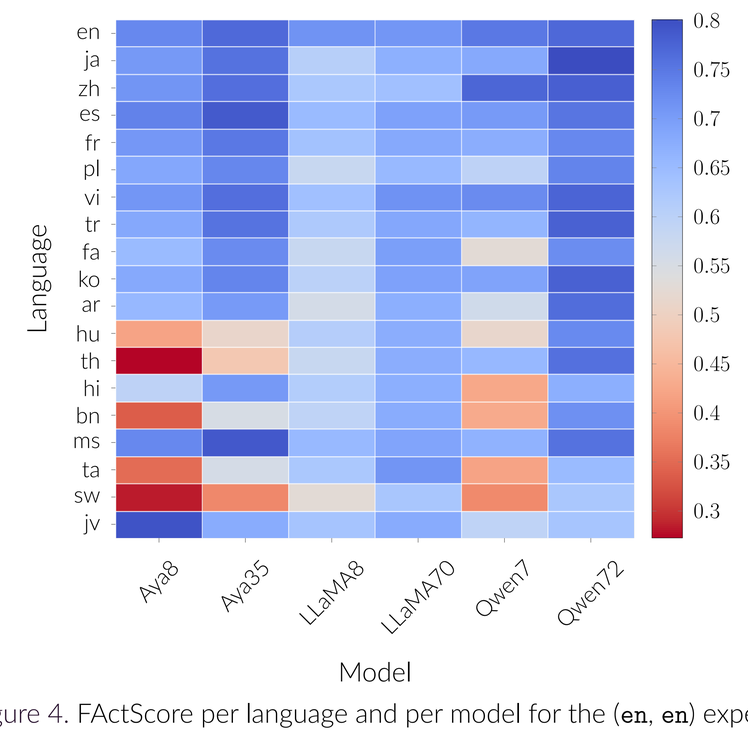


Figure 4. FActScore per language and per model for the (`en`, `en`) experiment

- FActScore's robustness depends on the language

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 1. Standard deviation across the 3 prompt templates of FActScore by Language Category and Experiment for all models

Limitations

- **Robustness**: Comparing a text to itself sometimes fails to yield a perfect 100% FActScore (see Figure 1)
- **Resource-Intensive Computation**
- **Intrinsic vs. Extrinsic Hallucinations**: Extrinsic hallucinations (unverifiable with available sources) cannot be captured
- **Wikipedia as Knowledge Source**: Inconsistent coverage across languages; ambiguous; potential inaccuracies
- **Use of LLMs as evaluators**: Performance relies on LM_{eval} abilities across languages and GPT-4 translation quality

Future work

- Further optimize FActScore computations
- Expand beyond biographies to study multilingual hallucination gaps in other tasks provided there exists a multilingual knowledge source
- Add a human benchmark for more insights, even though we chose to have a fully automated pipeline for wide language coverage
- Investigate alternative automated uncertainty metrics without relying on a knowledge source
- Conduct a comprehensive comparison of LLM-based methods for free-form text generation

References

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

Acknowledgement

Funding support for project activities has been partially provided by Canada CIFAR AI Chair, Facebook award, MEI award and FRQNT award. We also express our gratitude to Compute Canada and to Mila (mila.quebec) for their support in providing facilities for our evaluations.