Appendix A. 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 8, 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.

Appendix B. Characteristics of the LM_{subj}

Table 2 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).

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

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).

Appendix C. Validation of the LM_{eval}

Table 3 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..

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 3: 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.

Appendix D. Languages and People Dataset

Table 4 present the 19 selected languages along with their characteristics used for the selection process.

Figures 5 and 6 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.

	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 4: The 19 chosen languages with key statistics

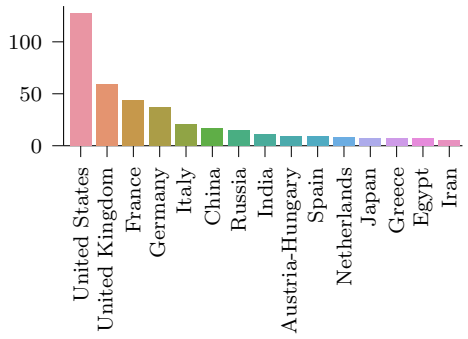

Figure 5: Top 15 citizenship countries

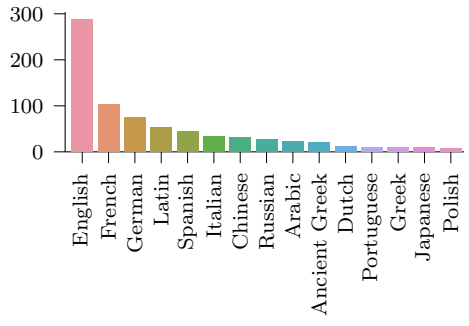

Figure 6: Top 15 languages spoken

Figure 7: Citizenship and language statistics of the entities

Appendix E. Wikipedia FActScore

Figure 8 present the FActScore distribution results when comparing the different Wikipedia summaries in every language to the English Wikipedia one.

We can see that the evaluator is not perfect, since 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 suggest that content in different languages can either contradict or diverge from what is found in English Wikipedia.

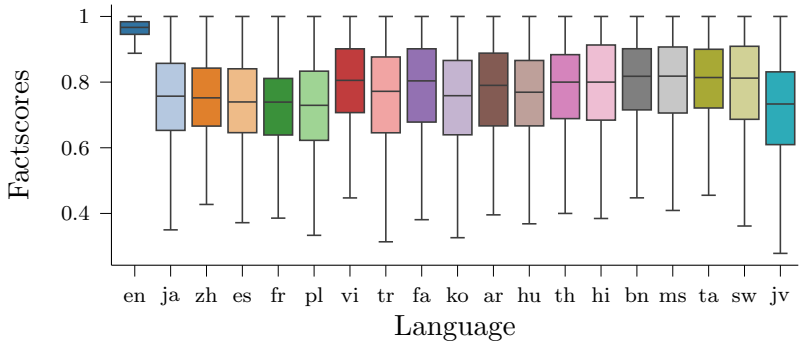

Figure 8: FActScore distribution for Wikipedia pages

Appendix F. Sanity checks

Table 5 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.

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

Appendix G. FActScore results per experiment

Figures 9 and 10 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.

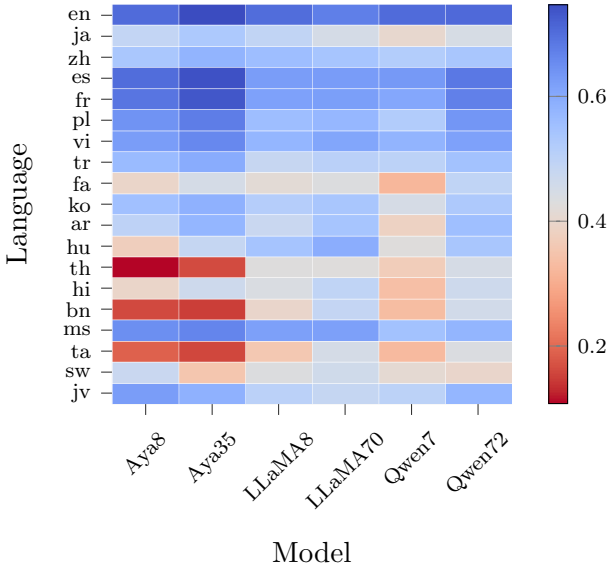

Figure 9: (lang, lang) experiment

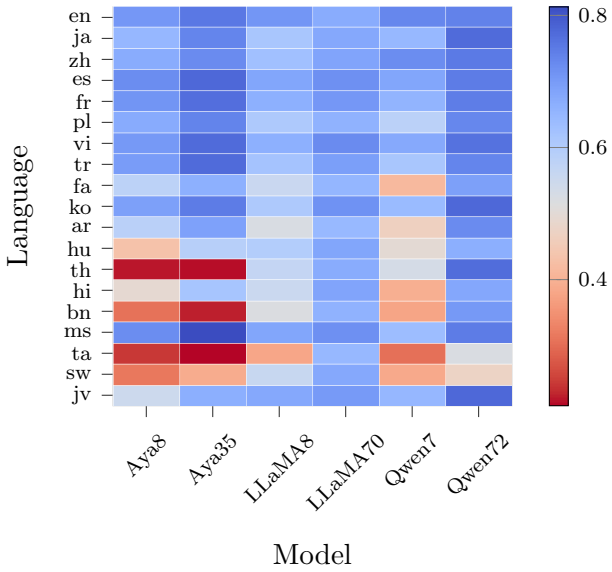

Figure 10: (lang, en) experiment

Appendix H. Standard deviation of FActScore

Table 6 presents FActScore standard deviations across three prompt templates per entity.

Language Category	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 6: Standard deviation across the 3 prompt templates of FActScore by Language Category and Experiment for all models

Appendix I. FActScore results per 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.

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

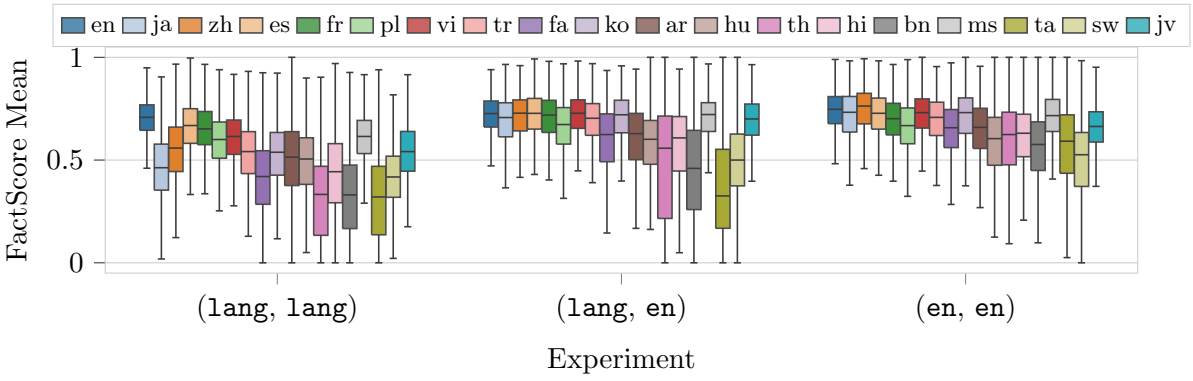

Figure 11: FactScore Mean distribution by Language and Experiment for Aya 8

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

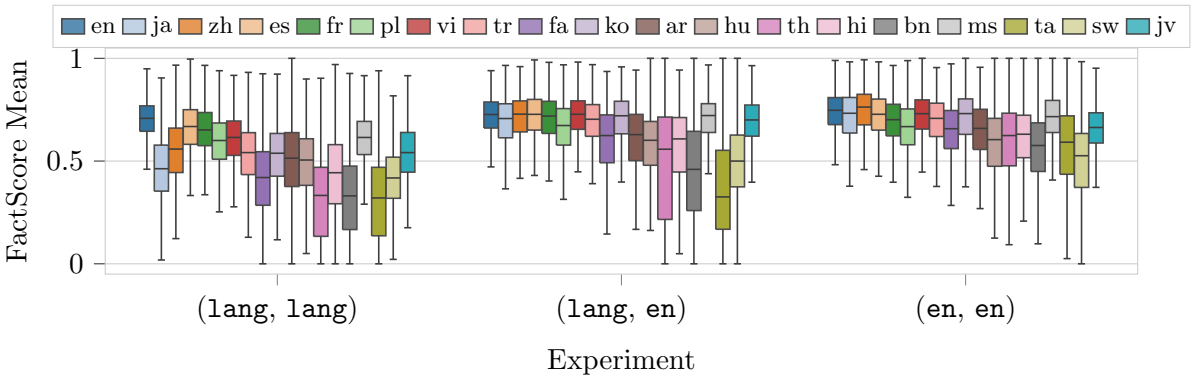

Figure 12: FactScore Mean distribution by Language and Experiment for Aya 35

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

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

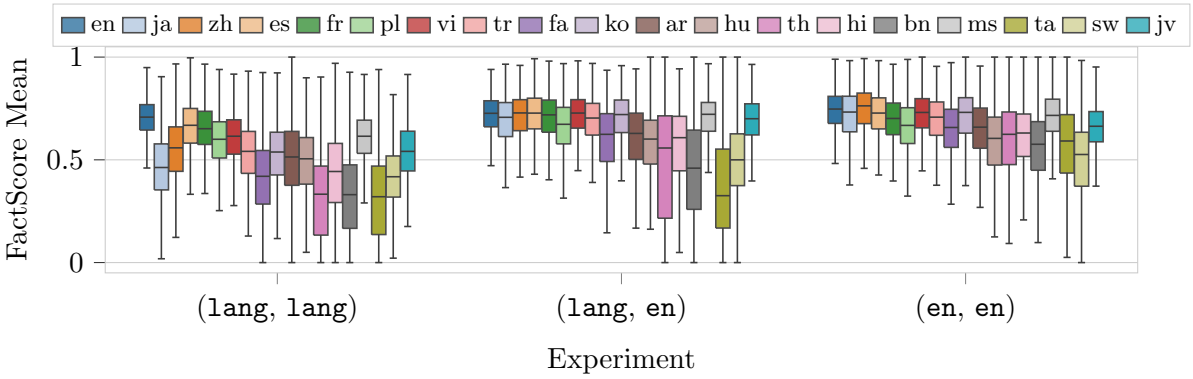

Figure 13: FactScore Mean distribution by Language and Experiment for LLaMA 8

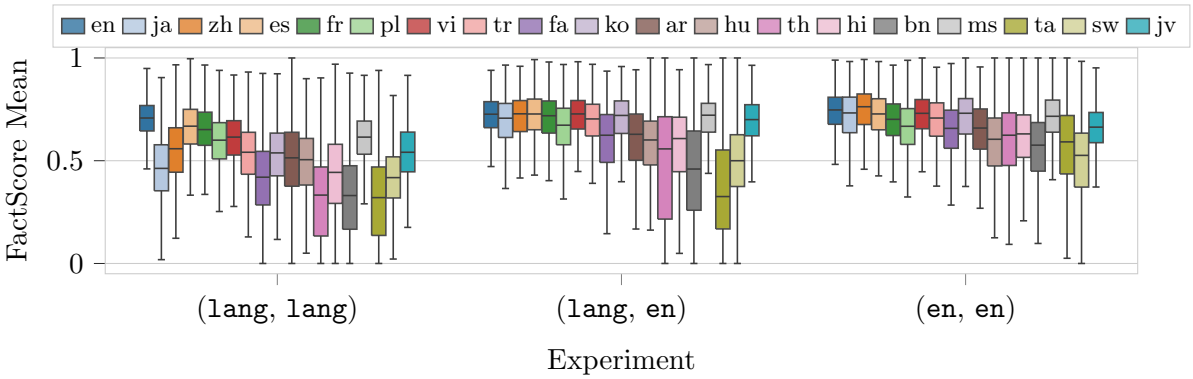

Figure 14: FactScore Mean distribution by Language and Experiment for LLaMA 70

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

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

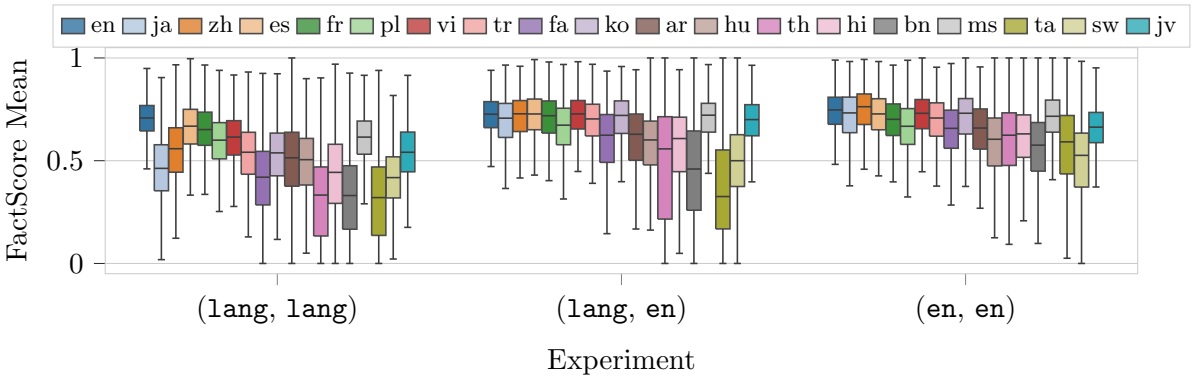

Figure 15: FactScore Mean distribution by Language and Experiment for Qwen 7

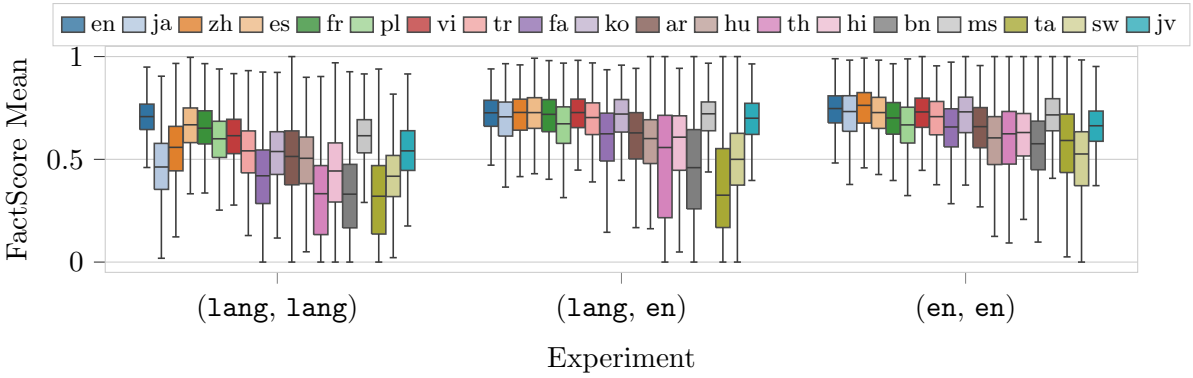

Figure 16: FactScore Mean distribution by Language and Experiment for Qwen 72