Language Models' Factuality Depends on the Language of Inquiry

Anonymous ACL submission

Abstract

Multilingual language models (LMs) are expected to recall factual knowledge consistently across languages, yet they often fail to transfer knowledge between languages even when they possess the correct information in one of the languages. For example, we find that an LM may correctly identify *Rashed Al Shashai* as being from *Saudi Arabia* when asked in Arabic, but consistently fails to do so when asked in English or Swahili. To systematically investigate this limitation, we introduce a benchmark of 10,000 country-related facts across 13 languages and propose three novel metrics—Factual Recall Score, Knowledge Transferability Score, and Cross-Lingual Factual Knowledge Transferability Score—to quantify factual recall and knowledge transferability in LMs across different languages. Our results reveal fundamental weaknesses in today's state-of-the-art LMs, particularly in cross-lingual generalization where models fail to transfer knowledge effectively across different languages, leading to inconsistent performance sensitive to the language used. Our findings emphasize the need for LMs to recognize language-specific factual reliability and leverage the most trustworthy information across languages. We release our benchmark and evaluation framework to drive future research in multilingual knowledge transfer.

1 Introduction

Large Language Models (LLMs) are often perceived as vast knowledge reservoirs, capable of recalling factual information across multiple languages (Wang et al., 2024). However, what if their knowledge is locked within linguistic boundaries and unable to be transferred across languages? Despite advancements in multilingual LMs such as Llama (Touvron et al., 2023a; Dubey et al., 2024), Gemma (Team et al., 2024a), DeepSeek (DeepSeek-AI et al., 2024), and Phi (Abdin et al., 2024; Li et al., 2023), our study reveals a striking asymmetry in their factual recall across languages: consider the example in Figure 1, where an LM is tasked with a simple factual query: "*Rashed Al Shashai is from which country?*" When asked in Arabic, several state-of-the-art LMs correctly generate the response: "*Saudi Arabia*." However, when posed in English, Hindi, or Swahili, the same models fail to recall the fact. This example suggests that models can correctly retrieve country-specific facts in the language associated with that country but struggle to do so in others.

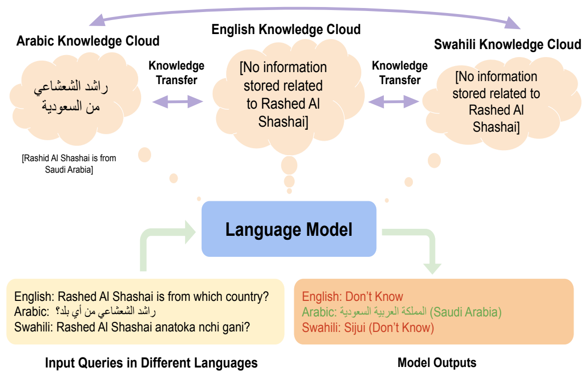

Figure 1: Illustratation of the cross-lingual factual knowledge transferability issue across linguistic knowledge clouds in LMs. The model correctly recalls that *Rashed Al Shashai is from Saudi Arabia* when queried in Arabic, but fails to retrieve this fact in English and Swahili, highlighting that factual knowledge is often stored in language-specific silos.

This raises a critical question—do these models truly internalize and transfer factual knowledge across languages, or do they merely encode isolated linguistic silos?

This limitation has significant implications for multilingual AI development and real-world applications. Many LM-based systems—such as retrieval-augmented generation (RAG) pipelines, multilingual search engines, and cross-lingual reasoning models—assume that factual knowledge is consistently available and transferable across languages.

Our findings reveal that LMs often rely on language-specific memorization rather than true cross-lingual knowledge generalization. This over-reliance can introduce biases, inconsistencies, and reliability issues in multilingual AI applications (Chua et al., 2024).

To systematically analyze the factual inconsistencies, we introduce a carefully curated dataset comprising country-related facts translated into 13 languages. This benchmark evaluates LMs on multiple dimensions—*factual recall, in-context recall, and counter-factual context adherence*—across high-, medium-, and low-resource languages. This benchmark comprises of 802 instances for factual recall, 156 instances for In-context recall, and 1404 instances for counter-factual context adherence as shown in Table 1.

Factual recall assesses the LM's ability to recall country-specific facts consistently across multiple languages. We evaluate factual recall using three metrics: (a) *Factual Recall Score (FRS)*: Measures how accurately a model recalls a fact in a given language, (b) *Knowledge Transferability Score (KTS)*: Quantifies how well factual knowledge is transferred across languages, and (c) *Cross-Lingual Factual Knowledge Transferability (X-FaKT) Score*: Combines the assessment of factual recall and cross-lingual transfer ability. FRS and KTS measure the effectiveness of cross-lingual knowledge transfer, and X-FaKT Score integrates factual recall with transferability to provide a robust measure of multilingual generalization. These metrics offer a more nuanced evaluation than a simple error rate, allowing for a deeper understanding of cross-lingual generalization.

In-Context Recall (Machlab and Battle, 2024) measures the general performance of the models in multilingual contexts. Inspired by (Du et al., 2024), we also study how factual knowledge of models affects their performance in handling in-context tasks in the multilingual setting (*Counterfactual Context Adherence*). For this, we design a dataset where factual knowledge conflicts with in-context instructions.

Our experiments reveal that while LMs often retrieve factual information correctly in the language associated with the fact, they struggle to transfer this knowledge to other languages. We also found that the size of the LLM plays an important role in factuality and knowledge transferability. For example, the combined performance of LLama-3-70B in factuality and knowledge transfer across languages is markedly (152% $\uparrow$ in *X-FaKT* Score) better than Llama-3.2-1B. In addition, there is a marked difference in these tasks when queries are asked in high-resource languages (46% $\uparrow$ in *X-FaKT* Score) as compared to the case with low resources. This finding exposes a critical limitation in current language models and their approach to multilingual knowledge integration. Our findings also reveal an interesting trade-off: LMs with stronger factual recall often struggle with counterfactual adherence, highlighting a key limitation in balancing factual memory and contextual reasoning. In our experiments, we observed that the factual knowledge of LMs could skew their judgments, leading to inaccurate evaluations. One has to be very careful when designing the prompt and using LM as an evaluator. We highlight the importance of controlling the evaluator's factual knowledge to ensure consistent and effective evaluation.

We plan to release our code and data upon publication. These are made available for reviewing purposes in the supplementary material.

2 Related Work

Multilingual NLP and Factual Recall. Prior work on multilingual models, such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mT5 (Xue et al., 2021) and BLOOM (Workshop et al., 2023), has shown that LMs trained on multilingual corpora exhibit varying performance across languages. Studies have also highlighted systematic biases in factual retrieval across different languages (Artetxe et al., 2020; Liu et al., 2020). While multilingual QA benchmarks such as XQuAD (Artetxe et al., 2020), MLQA (Lewis et al., 2020), and TyDiQA (Clark et al., 2020) assess factual consistency, they do not explicitly measure knowledge transfer within LMs. To address this gap, we introduce a benchmark designed to evaluate cross-lingual factual knowledge transferability.

Cross-Lingual Knowledge Transfer in LMs. While research suggests that multilingual LMs exhibit zero-shot and few-shot generalization across languages (Nooralahzadeh et al., 2020; Pfeiffer et al., 2020), empirical studies indicate that this transfer is often asymmetric, favoring high-resource languages (Hu et al., 2020). Most recent work has focused on cross-lingual transfer from high-resource languages to lower-resource ones

Task Type	# Examples
Factual Recall	802
In-context Recall	156
Counter-Factual Context Adherence	1404

Table 1: Number of examples per languages in our benchmark (§3).

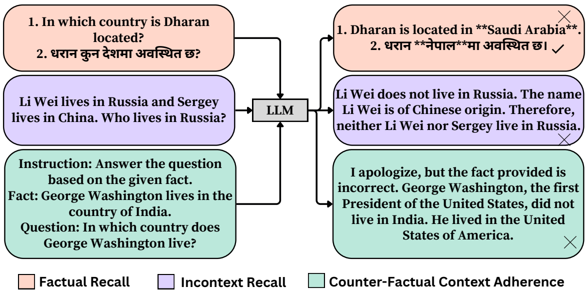

Figure 2: Examples from our multilingual dataset illustrating three tasks. Factual Recall: LMs recall country-specific facts better in native languages, as seen with Dharan's correct identification in Nepali but incorrect in English. Incontext Recall: Models struggle with contextual reasoning, showing regional bias when associating names with countries. Counter-Factual Context Adherence: When given counterfactual prompts about well-known figures, models rely on prior knowledge, affecting their ability to adhere to provided context.

(Zhao et al., 2024a,b). In contrast, we find that knowledge transfer is often lacking even from low-resource to high-resource languages. Furthermore, we show that recent LMs can correctly retrieve country-specific facts in the language associated with that country, regardless of the language's resource level.

Context Sensitivity and Counterfactual Reasoning. LMs can be susceptible to contextual cues, often overriding stored knowledge when presented with misleading information (Brown et al., 2020; Tirumala et al., 2022; Du et al., 2024). Counterfactual reasoning studies (Wu et al., 2023) show that models trained for high factual recall struggle with conflicting contextual instructions. While prior evaluations have been monolingual (Shwartz et al., 2020; Wang et al., 2020), our study extends these investigations into the multilingual domain, introducing in-context recall and counterfactual adherence tasks to analyze cross-lingual reasoning.

3 Dataset

We introduce a new multilingual dataset designed to evaluate three key capabilities of LMs: (a) *Factual Recall*, (b) *In-context Recall*, and (c) *Counter-Factual Context Adherence*. The number of instances in our dataset is given in the Table 1. Given the multilingual nature of our study, we categorize languages based on their resource availability in existing LM training corpora:

High-resource: English, Chinese, French, Japanese.

Medium-resource: Hindi, Russian, Arabic, Greek.

Low-resource: Nepali, Ukrainian, Turkish, Swahili, Thai.

These languages correspond to countries strongly associated with their usage: the United States, China, France, Japan, India, Russia, Saudi Arabia, Greece, Nepal, Ukraine, Turkey, Kenya, and Thailand. Now, we describe our datasets in detail.

3.1 Factual Recall

This task evaluates an LM's ability to recall country-specific facts across multiple languages. For example, given the query, *In which country is Mumbai located?*, the model should correctly respond with *India* when asked in different languages.

To construct the dataset, we curated a diverse set of entities—including cities, artists, sports figures, landmarks, festivals, and politicians—for 13 selected countries. We then created standardized templates for factual queries and translated them into each language using the Google Translate API (Google, n.d.). All translations were manually verified and refined as needed with the assistance of ChatGPT. In total, our dataset consists of 805 unique factual questions, each available in 13 language versions.

3.2 In-Context Recall

The in-context recall task evaluates how effectively an LM utilizes contextual information to answer a question, ensuring that internal knowledge does not influence the model's output.

Building on the work of (Feng and Steinhardt, 2024), we constructed our dataset by focusing on common person names associated with each country. For each example, we sampled two names and paired them with two different countries, creating context-based prompts as shown in violet color in Figure 2. To enhance dataset efficiency, we intentionally avoided associating a name with its most commonly linked country within the example.

3.3 Counter-Factual Context Adherence

This task evaluates an LM's susceptibility to counterfactual information by assessing whether it adheres to the provided context when answering a question. Ideally, the model should rely solely on the given context, but in some cases, its internal knowledge may interfere or override it, leading to unintended responses (Du et al., 2024). To investigate this, we curated a list of well-known personalities strongly associated with specific countries and deliberately introduced counterfactual information into the context.

For the example given in Figure 2, if the model defaults to its internal knowledge and answers *United States*, it demonstrates a resistance to the contextual information. Conversely, if it follows the counterfactual context and answers *India*, it suggests a higher reliance on the provided context rather than pre-existing knowledge.

One might expect these models to perform near-perfectly on these tasks, as they are very simple. However, despite the simplicity of these tasks, the performance varies across languages and models.

4 Experiments

In this section, we discuss our experimental setup, metric formulation, and both quantitative and qualitative analyses. We present the results of our experiments evaluating LMs on our dataset across diverse multilingual tasks. These experiments assess how language and country-specific factual knowledge influence LMs responses in a multilingual setting. All experiments were conducted using the latest models, with Qwen-2.5-72B-Inst (Qwen et al., 2025) serving as the evaluator (Li et al., 2024).

4.1 Experimental Setup

Models We evaluated 14 models of varying sizes, trained on different compositions of multilingual data, and fine-tuned using various preference optimization strategies (Ouyang et al., 2022; Rafailov et al., 2024), for our multilingual study. These include Deepseek (DeepSeek-AI et al., 2024), Qwen (Yang et al., 2024), Gemma (Team et al., 2024b), and Llama (Touvron et al., 2023b) families. Further details of the models evaluated are given in Table A.1.

Compute Details All our experiments were conducted on a set of 4 NVIDIA A100 GPUs, each with 80GB of VRAM. We used Chat-GPT (OpenAI et al., 2024) for the dataset generation.

Evaluation To evaluate all models on the curated datasets (Section 3), we used a temperature setting of 0 and a maximum token limit of 128. Specifically, we tested the models' performance on Factual Recall and In-Context Recall across different settings. For evaluation, we designed our metrics and utilized Qwen-2.5-72B-Inst as the evaluator (Li et al., 2024), with a maximum token limit of 256 to support reasoning. Evaluation prompts are shown in Figures 10 and 11.

4.2 Metric Definition and Formulation

This section introduces our carefully designed metrics to evaluate factual recall and knowledge transferability across languages in LMs. We propose two key metrics: the *Factual Recall Score (FRS)* and the *Knowledge Transferability Score (KTS)*. To establish a common metric for evaluating the model's performance in our benchmark, we compute their harmonic mean, which is defined as the *Cross-Lingual Factual Knowledge Transferability Score (X-FaKT)*, to ensure a balanced assessment while penalizing large disparities between them. Our metrics incorporate an inverse formulation with a correction factor to maintain a bounded range of $[0, 1]$. A higher error rate results in a lower metric value due to the inverse transformation, ensuring that better model performance corresponds to higher scores.

4.2.1 Associative vs. Non-Associative Knowledge

We categorize our dataset into two groups: associative and non-associative knowledge. The categorization is defined as follows: we consider 13 languages, each associated with a corresponding country (i.e., the ith language belongs to the ith country).

$$\text{Associative} = \{Q \in \text{Questions} : Q \in \text{Language}_i \wedge \text{output}(Q) = \text{Country}_j \wedge i = j\}$$

$$\text{Non-associative} = \{Q \in \text{Questions} : Q \in \text{Language}_i \wedge \text{output}(Q) = \text{Country}_j \wedge i \neq j\}$$

We denote the mean error rate for a country-specific fact asked in the language strongly associated with that country as $\mu_{\text{assoc.}}$, and the mean error rate for a country-specific fact asked in a language not associated with that country as $\mu_{\text{non-assoc.}}$.

4.2.2 Factual Recall Score (FRS)

Factual recall evaluates the model's ability to correctly retrieve both *associative* and *non-associative* knowledge. We define the Factual Recall Score (FRS) as:

$$FRS = \frac{3}{2}\left(\frac{1}{\mu_{\text{assoc.}} + \mu_{\text{non-assoc.}} + 1} - \frac{1}{3}\right) \quad (1)$$

- When both errors are zero ($\mu_{\text{assoc.}} = 0, \mu_{\text{non-assoc.}} = 0$), the model has a perfect factual recall, yielding an FRS score of 1.
- When both errors are high, the denominator increases, resulting in a lower FRS score closer to 0, indicating poor factual recall.

4.2.3 Knowledge Transferability Score (KTS)

Knowledge transferability quantifies how well a model maintains consistent factual knowledge across languages. We define the *Knowledge Transferability Score (KTS)* as:

$$KTS = 2\left(\frac{1}{|\mu_{\text{assoc.}} - \mu_{\text{non-assoc.}}| + 1} - \frac{1}{2}\right) \quad (2)$$

where:

- $|\mu_{\text{assoc.}} - \mu_{\text{non-assoc.}}|$ captures the absolute difference between associative and non-associative recall errors.

- When both errors are zero ($\mu_{\text{assoc.}} = 0, \mu_{\text{non-assoc.}} = 0$), there is perfect factual knowledge transfer, resulting in a KTS score of 1.
- When both errors are high but equal (e.g., $\mu_{\text{assoc.}} = 20, \mu_{\text{non-assoc.}} = 20$), KTS remains 1, indicating that while factual recall is poor, the model exhibits consistent errors across languages.
- When errors differ significantly (e.g., $\mu_{\text{assoc.}} = 20, \mu_{\text{non-assoc.}} = 2$ or vice versa), the absolute difference increases, leading to a lower KTS, highlighting a lack of knowledge transfer across languages.

4.2.4 Cross-Lingual Factual Knowledge Transferability Score (X-FAKT)

To ensure a balanced evaluation of factual recall and cross-lingual transferability, we compute their harmonic mean:

$$X\text{-FAKT} = 2 \times \frac{FRS \times KTS}{FRS + KTS} \quad (3)$$

where:

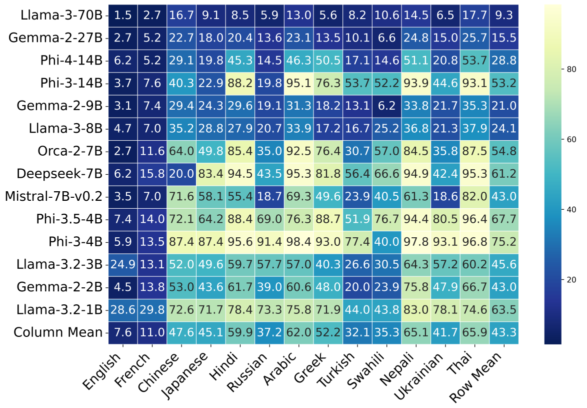

Figure 3: Error rates for each model on the Factual Recall task. A clear pattern emerges, showing a decline in performance as we move from larger to smaller models (top to bottom) and from high-resource to low-resource languages (left to right).

- The harmonic mean penalizes large disparities between factual recall (FRS) and knowledge transferability (KTS), ensuring that both contribute meaningfully to the final score.
- If either FRS or KTS is significantly lower, the overall score remains low, discouraging models from excelling in one metric while performing poorly in the other.
- A high X-FAKT score indicates that the model is both factually accurate and consistent across multiple languages.

This formulation provides a holistic evaluation of factual knowledge retention and cross-lingual consistency, making it a robust metric for assessing multilingual model performance.

4.3 Quantitative Analysis

4.3.1 Performance on Factual Recall task

The error rate across different LMs (Figure 3) reveals a clear pattern in performance across languages and model sizes. Notably, all models demonstrate superior performance on high-resource languages like English and French, with error rates consistently below 15% for most model variants. This performance gradually deteriorates as the model size decreases, with smaller models showing significantly higher error rates across all languages. However, an interesting observation emerges with languages like Swahili and Turkish, which despite being low-resource languages, exhibit relatively better performance with error rates comparable to mid-resource languages. This can be attributed to their use of Latin script, facilitating

Model	$\mu_{assoc.}$(%)	$\mu_{non-assoc.}$(%)	t-stat	p-value	FRS	KTS	X-FAKT
Llama-3-70B	**2.36** ± 5.12	**9.85** ± 10.54	2.52	0.01	**0.835**	**0.862**	**0.848**
Gemma-2-27B	4.23 ± 8.49	16.46 ± 17.07	2.54	0.01	0.742	0.783	0.762
Phi-4-14B	12.87 ± 16.51	30.15 ± 25.92	2.35	0.02	0.548	0.706	0.617
Phi-3-14B	25.09 ± 29.84	55.57 ± 36.24	2.93	<0.01	0.330	0.535	0.408
Gemma-2-9B	4.98 ± 6.09	22.32 ± 21.37	2.90	<0.01	0.677	0.705	0.691
Llama-3-8B	4.60 ± 7.54	25.77 ± 19.61	3.85	<0.01	0.649	0.651	0.650
Orca-2-7B	31.95 ± 31.65	56.77 ± 32.99	2.60	0.01	0.295	0.603	0.396
DeepSeek-7b	31.49 ± 30.68	63.73 ± 36.29	3.09	<0.01	0.268	0.514	0.353
Mistral-7B-v0.2	16.96 ± 15.65	45.25 ± 29.34	3.42	<0.01	0.424	0.559	0.483
Phi-3.5-4B	41.85 ± 31.62	69.87 ± 31.23	3.09	<0.01	0.208	0.563	0.304
Phi-3-4B	42.45 ± 30.99	77.95 ± 33.72	3.65	<0.01	0.181	0.477	0.262
Llama-3.2-3B	24.10 ± 17.80	47.48 ± 26.80	3.07	<0.01	0.375	0.620	0.467
Gemma-2-2B	9.97 ± 14.78	45.77 ± 31.30	4.06	<0.01	0.463	0.473	0.468
Llama-3.2-1B	34.74 ± 22.32	65.96 ± 26.98	4.03	<0.01	0.247	0.524	0.336

Table 2: Results of the t-test comparing associative and non-associative knowledge across models, alongside FRS, KTS, and X-FAKT scores. (A) Llama-3-70B achieves the best performance in both factual recall and knowledge transferability. (B) There is a statistically significant difference between the performance on associative queries (asked in a country's native language) and non-associative queries (asked in other languages).

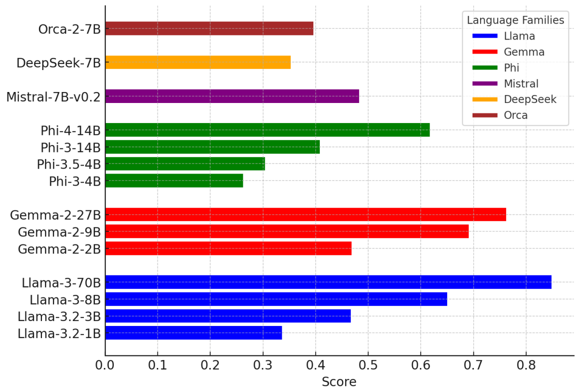

Figure 4: This figure illustrates the model-wise comparison of X-FAKT scores grouped by language families. A clear trend emerges, showing that as the model size increases within a family, the X-FAKT score tends to increase.

better knowledge transfer from English.

A compelling pattern emerges when examining languages that share similar scripts, and strong correlations in model performance among languages that share similar scripts. For example, the error patterns for Hindi-Nepali and Russian-Ukrainian pairs show remarkable similarities, suggesting that the models effectively leverage shared scriptural characteristics during learning. These patterns indicate that script similarity plays a crucial role in the model's ability to generalize across languages, potentially offering insights into how these models transfer knowledge between different language pairs and scripts.

Knowledge Transferability Analysis: From Table 2, Llama-3-70B emerges as the clear leader with the highest X-FAKT score of 0.848, demonstrating superior balanced performance in both factual recall (FRS = 0.835) and knowledge transferability (KTS = 0.862). This exceptional performance is supported by the lowest error rates ($\mu_{assoc.} = 2.36\%$, $\mu_{non-assoc.} = 9.85\%$), suggesting that larger model sizes generally correlate with better cross-lingual factual knowledge handling. Despite similar model sizes, significant performance variations exist between different architectures. For example, Gemma-2-9B (X-FAKT: 0.691) substantially outperforms Mistral-7B-v0.2 (X-FAKT: 0.483), suggesting that architecture design and training methodology play crucial roles beyond mere parameter count. As illustrated in Figure 4, the X-FAKT scores exhibit a clear upward trend with increasing model size within each language family. This suggests that larger models generally achieve better factual consistency, highlighting the impact of scale on model performance. These findings provide valuable insights into the current state of cross-lingual factual knowledge in LMs and highlight areas for future improvement, particularly in reducing the performance gap between associative and non-associative knowledge retrieval.

Language	$\mu_{assoc.}(\%)$	$\mu_{non-assoc.}(\%)$
High	3.83 ± 3.79	29.84 ± 27.47
Medium	26.73 ± 17.60	50.54 ± 21.20
Low	29.53 ± 16.19	53.91 ± 23.68

Table 3: Average mean and standard deviation for error rate across all models for each language group. High-resource languages exhibit lower error rates compared to low-resource languages.

Associative vs. Non-associative performance: We analyze the performance of various models on these two subsets of data and report the results in the Table 2. For all models, the t-statistic and p-value indicate that the differences between associative and non-associative categories are statistically significant (p-value less than 0.05).

Performance comparison across language groups: In this study, we categorize languages into three groups based on their availability and coverage in the dataset: **High**, **Medium**, and **Low**, as defined in Section 3. From the results shown in Table 3, we observe a clear trend across language groups. Specifically, high resouce languages exhibit the lowest average error rates, particularly in the associative category, where models make fewer mistakes ($\mu_{assoc.}$ = 3.83%). However, for non-associative questions, the error rate rises significantly ($\mu_{non-assoc.}$ = 29.84%), indicating that models struggle more when dealing with non-associative samples in these languages. The error rate increases while moving from high to low-resource languages.

4.3.2 Performance on In-Context Recall task

Figure 12 demonstrates the incorrectness rate for the in-context recall capabilities of different LMs. Despite being a simple task, certain models such as *DeepSeek-7B, Orca-2-7B, Phi-3-4B, Llama-3.2-1B, and Mistral-7B-v0.2* perform poorly across multiple languages. This suggests that these models struggle to effectively utilize contextual information when generating outputs. Interestingly, even for languages like Swahili and Turkish, which showed better scores in the Factual Recall task, models demonstrate poor performance on this context-dependent task. This stark contrast suggests that the benefits of Latin script-based knowledge transfer observed in the Factual Recall task do not extend to in-context learning scenarios, where performance depends primarily on the model's ability to process and utilize contextual information.

As mentioned in the dataset section, we intentionally paired cross-entities as context. This setup appears to induce a regional bias, which negatively impacts model performance. The structured entity-context pairing in the dataset may have led to spurious correlations (Yang et al., 2023; Ye et al., 2024), reducing model accuracy in in-context recall tasks. Some models struggle to effectively leverage contextual information, revealing potential weaknesses in their retrieval and in-context learning mechanisms.

4.3.3 Performance on Counter-Factual Context Adherence task

Figure 5 illustrates the error rates of LMs in the Counterfactual Context Adherence task. Notably, Latin-script languages (English, French, Swahili, and Turkish), which performed well in factual recall tasks, exhibited significantly higher error rates in counterfactual adherence. This suggests a fundamental trade-off in the models' capabilities: their strength in accurately retrieving factual information appears to come at the expense of their ability to maintain adherence to counterfactual contexts. This inverse relationship raises important questions about the inherent limitations and trade-offs in LMs' learning mechanisms, particularly in how they balance factual knowledge with hypothetical reasoning.

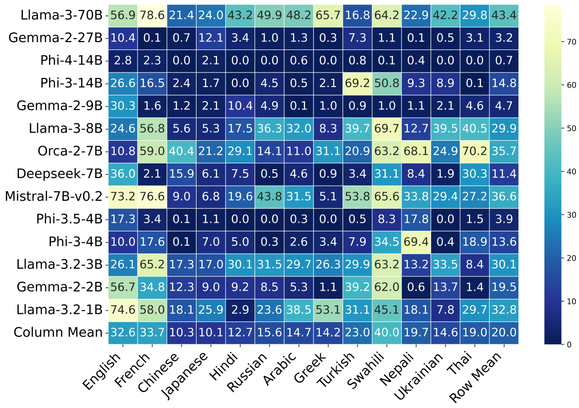

Figure 5: Error rate for each model on Counter-Factual Context Adherence task. Models show high error rates in high resource languages such as English and French where they have high factual recall.

Instruction: Answer the question based on the given context.
Context: Li Wei lives in Russia, and Sergey lives in China.
Question: Who lives in Russia?

Model Output: Li Wei does not live in Russia. The name Li Wei is of Chinese origin. Therefore, neither Li Wei nor Sergey live in Russia.

Figure 6: Mistral-7B-v0.2 output when prompted with the given context in English. This model generation shows how spurious correlation leads to in-context recall failures

4.4 Qualitative Analysis

Spurious correlation leads to in-context recall failures. We observe that some models tend to associate names with cultural origins, even when contextual evidence contradicts this assumption. Figures 6 demonstrate the model response when prompted Mistral-7B-v0.2 with the contextual understanding-based question in English.

Despite the explicit context stating that *Li Wei* resides in *Russia*, the model disregards this information and defaults to cultural associations. This behavior reveals a limitation in integrating contextual evidence when making country-specific inferences.

Models favor factual knowledge over context. We also observed that some models prioritize their internal factual knowledge over contextual information when responding to questions about well-known personalities. Figures 7 demonstrate the model response when prompted Llama-3-70B with the factual retrieval query in English.

Instruction: Answer the question based on the given fact.
Fact: George Washington lives in the country of India.
Question: In which country does George Washington live?

Model Output: I apologize, but the fact provided is incorrect. George Washington, the first President of the United States, did not live in India. He lived in the United States of America.

Figure 7: Llama-3-70B output when prompted with a counter-factual context adherence query in English. This shows LMs favour internal knowledge over contextual understanding.

Instruction: In which country is Dijon located?

Model Output: Dijon is located in France. It is a city in the Bourgogne-Franche-Comté region, and it is the capital of the Côte-d'Or department.

Figure 8: Llama-3-70B output when prompted with a factual recall query in English

In this case, despite being explicitly told that *'George Washington'* lived in *'India'*, the model relied on its factual knowledge, correcting the given fact and asserting that *'George Washington'* lived in the *'United States'*. This response demonstrates the model's strong reliance on factual accuracy, rather than adapting to the context provided. It suggests that when it comes to well-known historical figures, models may prioritize prior knowledge over the specific context they are given.

Linguistic variability in word interpretation. LMs can interpret words differently depending on the language. Figures 8 and 9 demonstrate the model responses when prompted Llama-3-70B with the same queries but in different languages. This highlights challenges in multilingual consistency, where the model misinterprets *'Dijon'* as *'De Janeiro'* in Hindi, revealing inconsistencies in cross-lingual factual retrieval.

Challenges with using LMs as evaluators. We used a zero-shot prompt with Llama-3-70B as an evaluator and found that its inherent factual knowledge can skew assessments. For example, when evaluating a Gemma-2-27B response to the counter-factual context task—*"Catherine the Great lives in India"*—the evaluator corrected it, asserting that she lived in *"Russia"*, despite the provided ground truth. This bias highlights the need to control evaluators' factual knowledge to ensure consistent evaluation.

Instruction: डिजॉन किस देश में स्थित है?

Model Output: डी जाँ (De Janeiro) ब्राजील देश में स्थित है। रियो डी जाँ (Rio de Janeiro) ब्राजील का दूसरा सबसे बड़ा शहर है और इसकी स्थापना 1565 में हुई थी।

(**English Translation:** Dijon (De Janeiro) is located in Brazil. Rio de Janeiro is the second largest city in Brazil and was founded in 1565.)

Figure 9: Llama-3-70B output when prompted with a factual recall query in Hindi. In Hindi, it misinterprets understanding of a French word.

5 Conclusions

Our study reveals a critical limitation in multilingual LMs: their inability to consistently transfer factual knowledge across languages. Our benchmark provides a standardized framework to evaluate both current and future LMs on their factual consistency and cross-lingual generalization, enabling a more systematic comparison of their capabilities. Moreover, it can serve as a valuable resource to promote research in interpretability by helping analyze how and where factual knowledge is stored and retrieved across languages, fostering a deeper understanding of LM internals. We emphasize the need for AI systems with internal awareness of their language-specific strengths and weaknesses—a concept we term calibrated multilingualism. Under this paradigm, a model would autonomously leverage the most reliable internal representations for any given multilingual query.

We also find that LMs, when used as evaluators, are biased by their internal factual knowledge, which may not align with the intended input-output-ground-truth context. This underscores the need to control the evaluator's factual knowledge for more reliable assessments. Ultimately, enabling AI to cross-generalize across languages is crucial for inclusive and equitable technology, ensuring language is no barrier to reliable knowledge access.

6 Limitations

Our study provides valuable insights into cross-lingual knowledge transfer in LMs but has some limitations. First, our benchmark, though comprehensive in country-related facts, covers only 13 languages, limiting its representation of diverse linguistic families. Second, we evaluated only open-source LMs, excluding proprietary models that may exhibit different transfer patterns. Third, our fact collection used a standardized template for consistency, which may not reflect the diversity of real-world queries. Lastly, our focus on country-related facts means our findings may not generalize to other domains like science, history, or culture.

7 Ethics Statement

This research is conducted with a strong commitment to ethical principles, ensuring data privacy and consent by using publicly available information and adhering to data protection regulations. We acknowledge potential biases in multilingual language models and aim to highlight and address these through our benchmark. Transparency and reproducibility are promoted by making our dataset and evaluation framework publicly available. Our research aligns with the broader goals of fairness, transparency, and social responsibility.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, Chulin Xie, and Chiyuan Zhang. 2024. Crosslingual capabilities and knowledge barriers in multilingual large language models. *arXiv preprint arXiv:2406.16135*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. 2024. Deepseek llm: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer C White, Aaron Schein, and Ryan Cotterell. 2024. Context versus prior knowledge in language models. *arXiv preprint arXiv:2404.04633*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiahai Feng and Jacob Steinhardt. 2024. How do language models bind entities in context? *Preprint*, arXiv:2310.17191.

Google. n.d. Google translate. Accessed: 2025-02-16.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *Preprint*, arXiv:2003.11080.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Daniel Machlab and Rick Battle. 2024. Llm in-context recall is prompt dependent. *Preprint*, arXiv:2404.08865.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024b. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023b. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Chenguang Wang, Xiao Liu, and Dawn Song. 2020. Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*.

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. Knowledge mechanisms in large language models: A survey and perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7097–7135, Miami, Florida, USA. Association for Computational Linguistics.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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*.

Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. 2023. Mitigating spurious correlations in multi-modal models during fine-tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39365–39379. PMLR.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. 2024. Spurious correlations in machine learning: A survey. *Preprint*, arXiv:2402.12715.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024a. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024b. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. *arXiv preprint arXiv:2403.05189*.

A APPENDIX

[Instruction]
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. The assistant should be penalized if it doesn't answer the question based on the given reference. For this evaluation, you should primarily consider the following criteria:

Score 1: The Assistant's answer is in the same language as the reference and aligns with the reference.
Score 2: The Assistant's answer is in English and aligns with the reference.
Score 3: The Assistant's answer is in any other language and aligns with the reference.
Score 4: The Assistant's answer is unrelated to the reference.
Only respond with a numerical score.

Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 4 by strictly following this format: "[[rating]]", for example:

[Reference]
India

[Question]
George Washington lives in the country of India. In which country does George Washington live?

[The Start of Assistant's Answer]
George Washington lives in **United States**.
[The End of Assistant's Answer]

Response: The assistant's response is in same language but does not align with the reference provided.\n\nRating: [[4]]

[Reference]
{answers}

[Question]
{question}

[The Start of Assistant's Answer]
{predicted}

[The End of Assistant's Answer]

Figure 10: Prompt used for evaluation of Factual Recall and In-Context Recall tasks.

[Instruction]
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. The AI assistant is given a fact and the assistant has to answer the question based on the given fact. The assistant should be penalized if it doesn't answer the question based on the fact. For this evaluation, you should primarily consider the following criteria:

Score 1: The Assistant's answer is in the same language as the reference and aligns with the reference.
Score 2: The Assistant's answer is in English and aligns with the reference.
Score 3: The Assistant's answer is in any other language and aligns with the reference.
Score 4: The Assistant's answer is unrelated to the reference.
Only respond with a numerical score.

Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 4 by strictly following this format: "[[rating]]", for example:

[Reference]
India

[Question]
Instruction: Answer the question based on the given fact.
Fact: George Washington lives in the country of India.
Question: In which country does George Washington live?

[The Start of Assistant's Answer]
The given fact is incorrect, George Washington lives in **United States**.
[The End of Assistant's Answer]

Response: The assistant's response is in English and does not align with the reference provided.\n\nRating: [[4]]

[Reference]
{answers}

[Question]
{question}

[The Start of Assistant's Answer]
{predicted}

[The End of Assistant's Answer]

Figure 11: Prompt used for evaluation of Counter-Factual Context Adherence task.

Model	Model Size & Architecture	Training Data	Languages Supported	Context Length	Vocab Size	Post-Training Strategies	Key Features
Llama-3-70B	70B L=80, H=64	15T tokens Multi-lingual	EN, DE, FR, IT, PT, HI, ES, TH	8K	128K	SFT, RS, DPO	GQA, 8 heads, RoPE embeddings
Gemma-2-27B	27B	13T tokens Web, Code, Math	Primarily English	8K	256K	SFT, RLHF	Local-global attention, Knowledge distillation
Phi-4-14B	14B	400B synthetic + 10T web	DE, ES, FR, PT, IT, HI, JA	16K	100K	SFT, RS, DPO	Full attention over 4K context
Phi-3-14B	14B	4.8T tokens	10% multilingual data	128K	32K	SFT, DPO	Reasoning focus, Multi-lingual support
Gemma-2-9B	9B	8T tokens	Primarily English	8K	256K	SFT, RLHF	GQA, RoPE, Knowledge distillation
Llama-3-8B	8B L=32, H=32	15T tokens Multi-lingual	EN, DE, FR, IT, PT, HI, ES, TH	8K	128K	SFT, RS, DPO	GQA, RoPE, 32 heads
Orca-2-7B	7B L=32, H=32	Based on Llama 2	Based on Llama 2	4K	32K	Single-turn SFT	Enhanced reasoning abilities
DeepSeek-7B	7B L=30, H=32	2T tokens	English & Chinese	4K	102K	SFT, DPO	English & Chinese focus
Mistral-7B-v0.2	7B L=32, H=32	Open Web	Open Web languages	32K	32K	SFT	GQA, Sliding window attention
Phi-3.5-4B	3.8B L=32, H=32	3.4T tokens Multi-lingual	23 languages incl. AR, ZH, CS, NL,	128K	32K	SFT, DPO	Multi-lingual support
Phi-3-4B	3.8B	4.9T tokens	Similar to Phi-3.5-mini	128K	32K	SFT, DPO	Diverse domain coverage
Llama-3.2-3B	3B	9T tokens Multi-lingual	EN, DE, FR, IT, PT, HI, ES, TH	128K	128K	SFT, RS, DPO	GQA, Knowledge distillation
Gemma-2-2B	2B	2T tokens	Primarily English	8K	256K	SFT, RLHF	GQA, RoPE, Knowledge distillation
Llama-3.2-1B	1B	9T tokens Multi-lingual	EN, DE, FR, IT, PT, HI, ES, TH	128K	128K	SFT, RS, DPO	GQA, Multi-lingual dialogue focus

Table A.1: **Model Specifications**: Comparison of LMs across key architectural and training characteristics. Languages abbreviated as: EN=English, DE=German, FR=French, IT=Italian, PT=Portuguese, HI=Hindi, ES=Spanish, TH=Thai, AR=Arabic, ZH=Chinese, CS=Czech, NL=Dutch, JA=Japanese. L=Layers, H=Heads, SFT=Supervised Fine-Tuning, RS=Rejection Sampling, DPO=Direct Preference Optimization, GQA=Grouped-Query Attention, RoPE=Rotary Position Embedding.

Question	Model Prediction	Ground Truth	LLM Evaluation Output
New York iko katika nchi gani?	New York iko nchini **Marekani**.	Marekani	Response: The assistant's response is in the same language as the reference and aligns with the reference provided. Rating: [[1]]
Talimat: Verilen gerçeğe göre soruyu cevaplayın. Gerçek: George Washington, Amerika Birleşik Devletleri ülkesinde yaşıyor. Soru: George Washington hangi ülkede yaşıyor?	George Washington, **Amerika Birleşik Devletleri**'nde yaşıyor.	Amerika Birleşik Devletleri	The assistant's response is in the same language as the reference and aligns with the reference. Rating: [[1]]
David lives in China and Li Wei lives in United States. Who lives in China?	David lives in China.	David	Response: The assistant's response is in the same language and aligns with the reference provided. Rating: [[1]]
Dans quel pays se trouve Sourate ?	Sourate n'est pas un pays, mais le nom d'une **ville** située en **Algérie**.	Inde	The assistant's response is in French, which aligns with the reference language, but the information provided does not align with the reference. The assistant incorrectly states that Sourate is a city in Algeria, while the reference is about India. Rating: [[4]]

Table A.2: Some outputs from evaluator Qwen-2.5-72B-Inst. Rating[1-3]: Correct and Rating[4]: Incorrect.

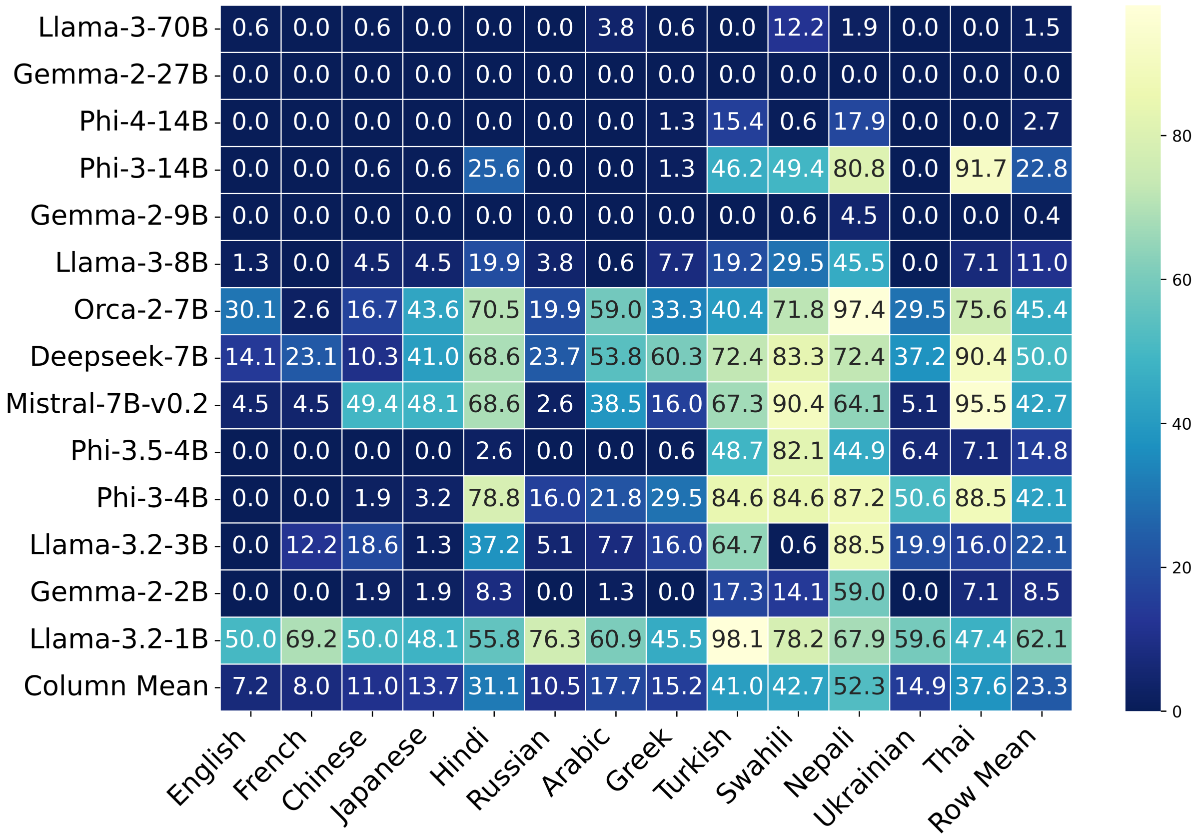

Figure 12: Error rate for each model on In-context Recall task. Clearly, few models such as DeepSeek-7B, Phi-3-4B, etc. performs poorly on this simple task.

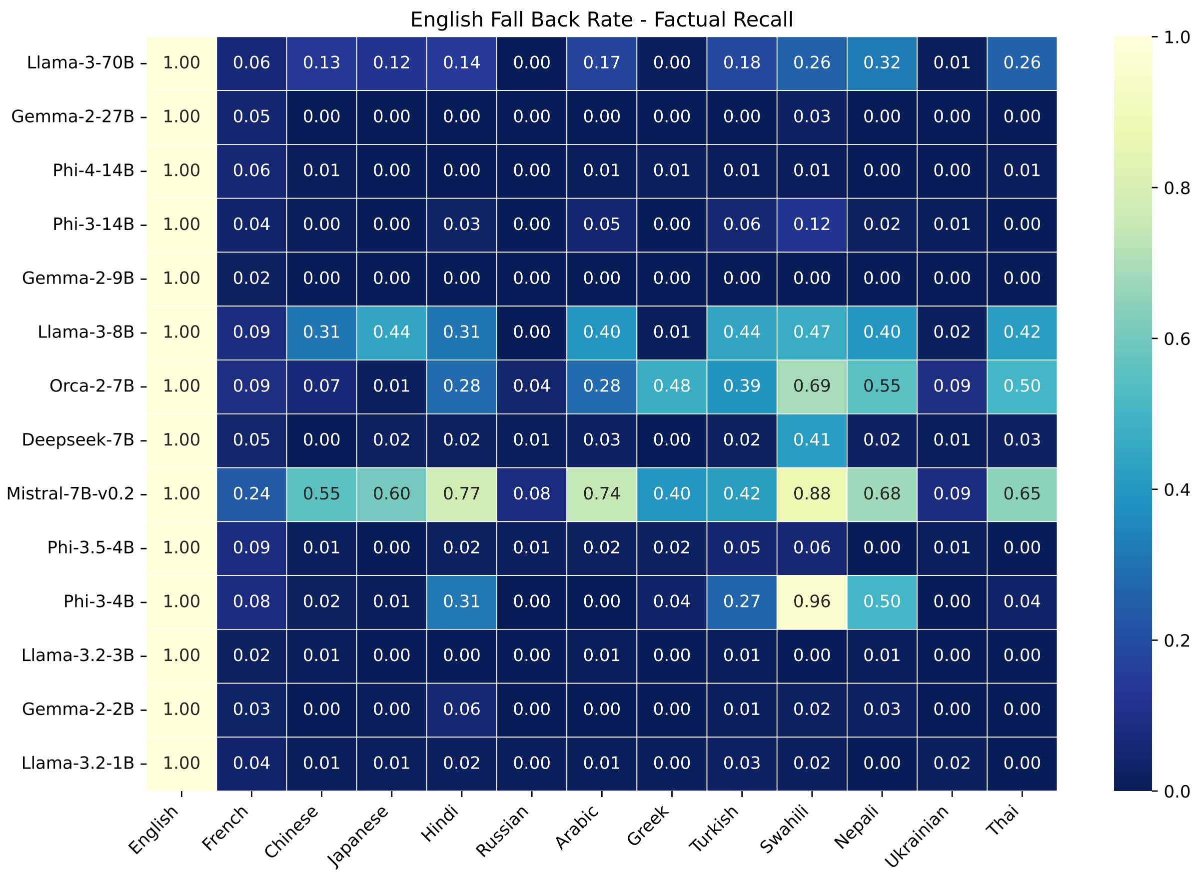

Figure 13: English Fall Back Rate across models (The English Fall Back Rate measures the frequency with which a model defaults to English in its output).

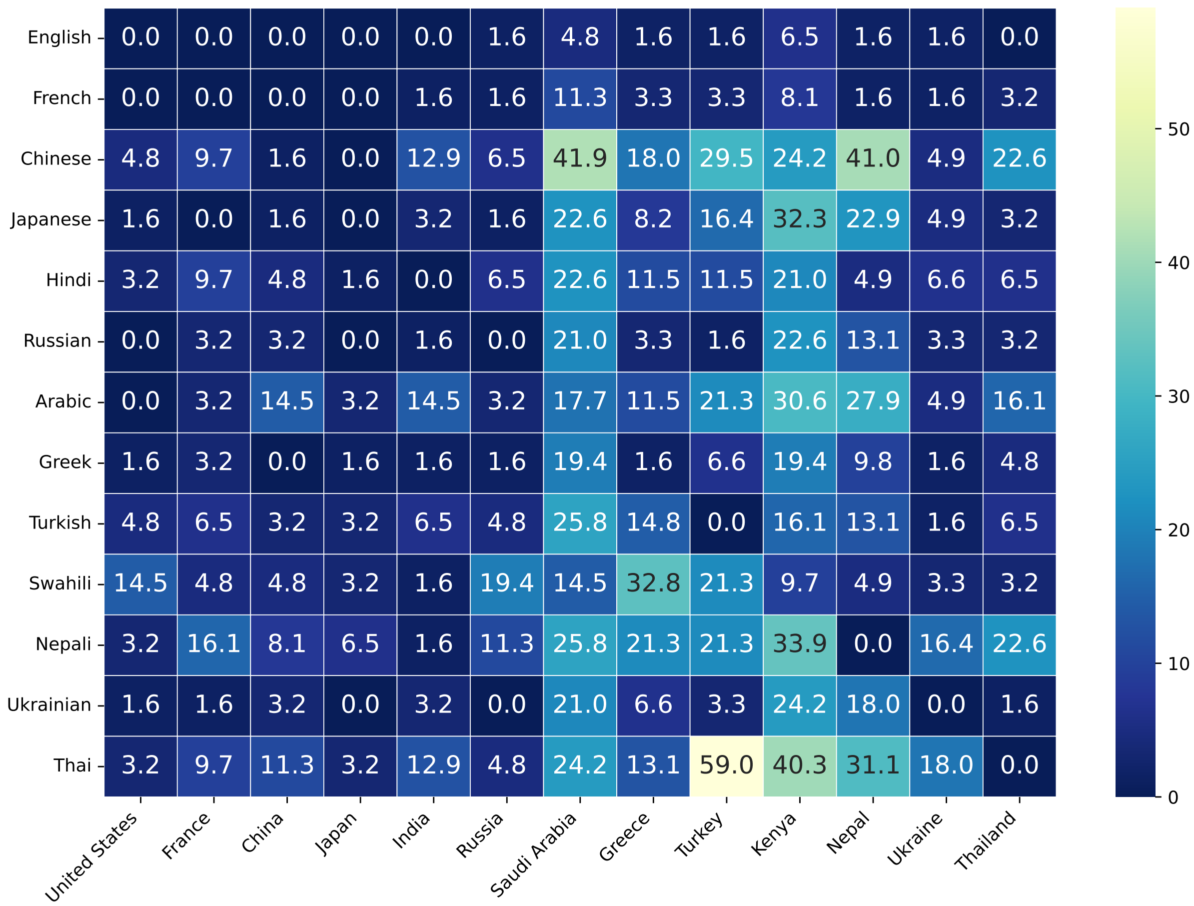

Figure 14: Country-Specific Factual Error Rates in each language for Llama-3-70B

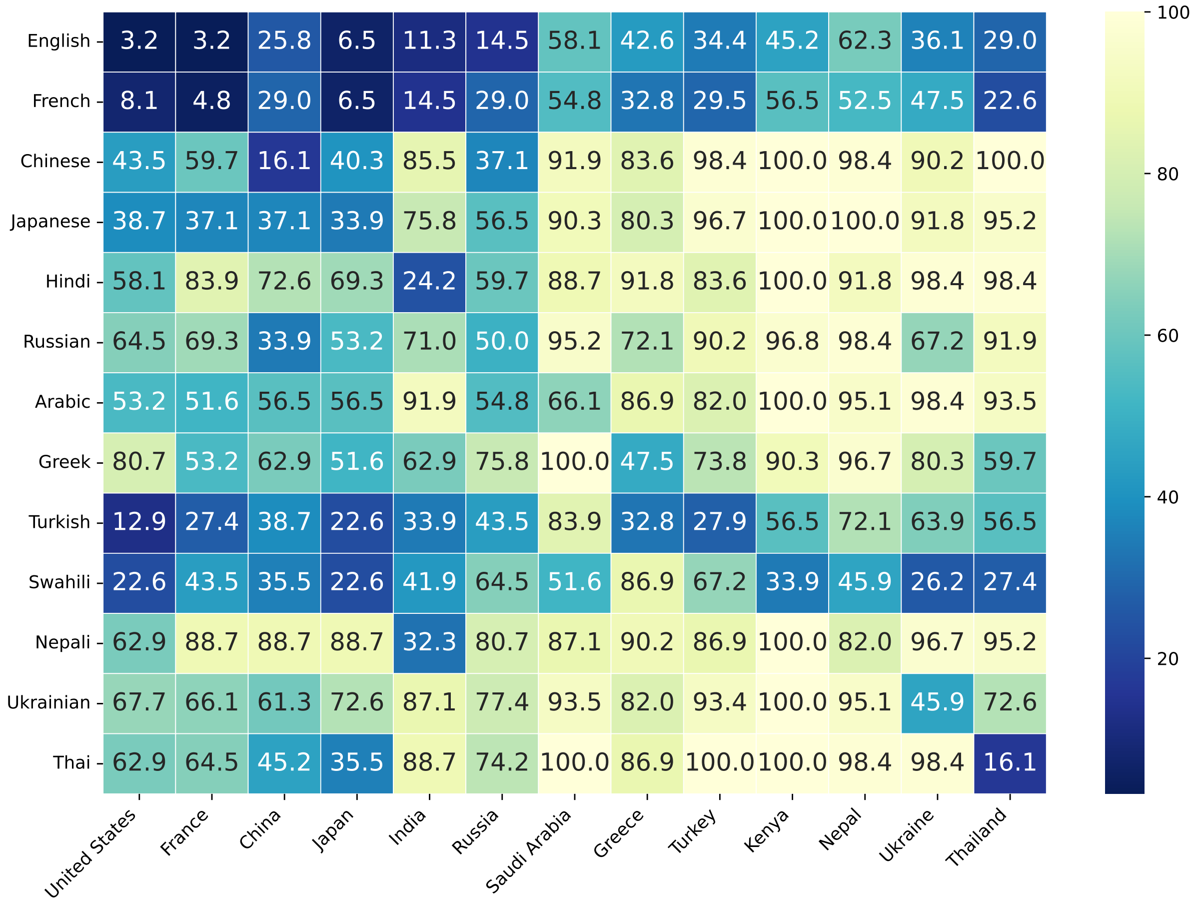

Figure 15: Country-Specific Factual Error Rates in each language for Llama-3.2-1B

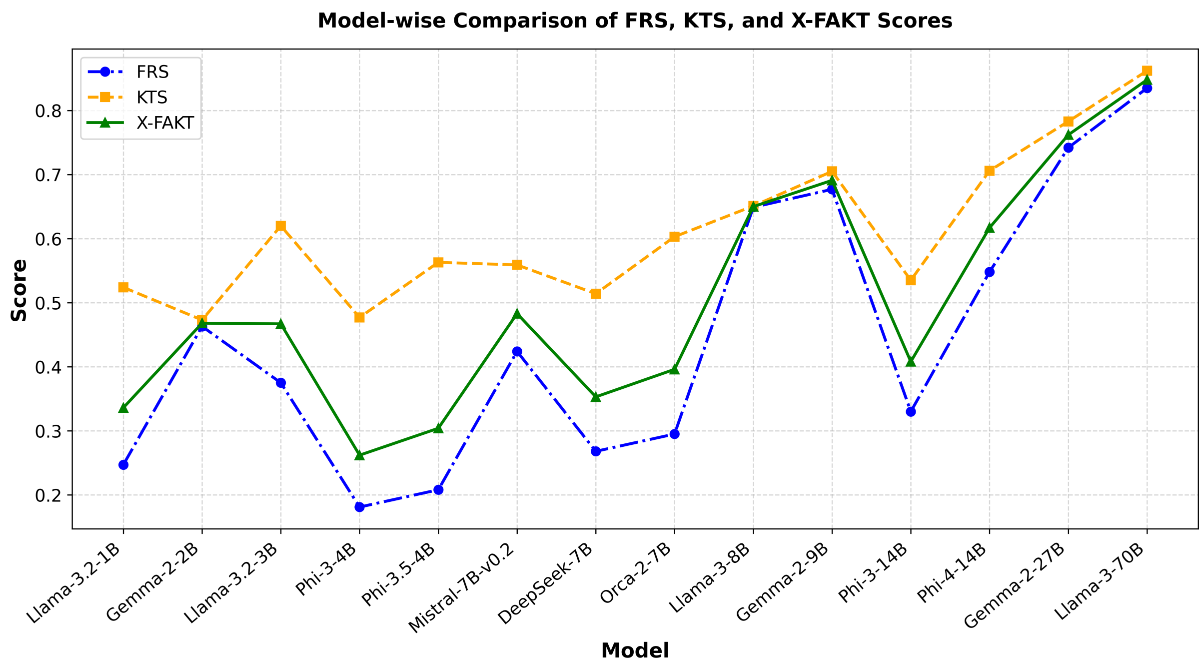

Figure 16: Comparision of models (in the increasing order of size with respect to the parameters) using Factual Recall Score, Knowledge Transferability Score, and Cross-Lingual Factual Knowledge Transferability Score.