Beneath the Surface of Consistency: Exploring Cross-lingual Knowledge Representation Sharing in LLMs

Anonymous ACL submission

Abstract

The veracity of a factoid is largely independent of the language it is written in. However, language models are inconsistent in their ability to answer the same factual question across languages. This raises questions about how LLMs represent a given fact across languages. We explore multilingual factual knowledge through two aspects: the model's ability to answer a query consistently across languages, and the ability to "store" answers in a shared representation for several languages. We propose a methodology to measure the extent of representation sharing across languages by repurposing knowledge editing methods. We examine LLMs with various multilingual configurations using a new multilingual dataset. We reveal that high consistency does not necessarily imply shared representation, particularly for languages with different scripts. Moreover, we find that script similarity is a dominant factor in representation sharing. Finally, we observe that if LLMs could fully share knowledge across languages, their accuracy in their best-performing language could benefit an increase of up to 150% on average. These findings highlight the need for improved multilingual knowledge representation in LLMs and suggest a path for the development of more robust and consistent multilingual LLMs.

1 Introduction

Pretrained large language models (LLMs) have demonstrated a remarkable capacity to encode and retrieve factual knowledge (Petroni et al., 2019; Chang et al., 2024) across diverse languages (Kassner et al., 2021; Jiang et al., 2020). However, substantial variation in model performance across languages with a strong bias toward high-resource languages (Kassner et al., 2021; Jiang et al., 2020; Fierro and Søgaard, 2022; Jiang et al., 2022; Qi et al., 2023a) highlights the issue of cross-lingual knowledge (in-)consistency.

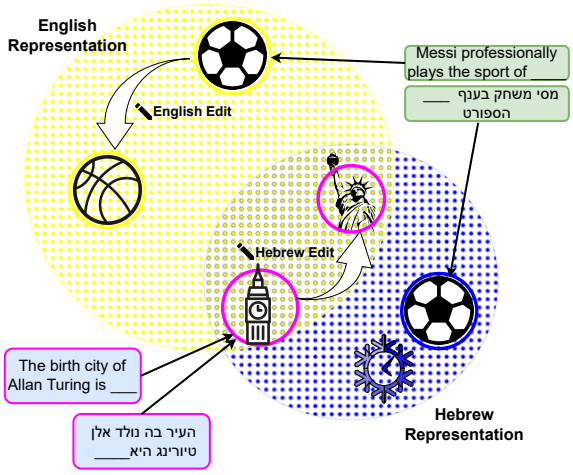

Figure 1: Illustration of our method for distinguishing between cross-lingual consistency and representations sharing in a pairwise language setting. The sports (green) question demonstrates mere cross-lingual answer consistency, while the query about Allan Turing's birthplace (blue) exemplifies a shared underlying representation. Edits to the shared representation propagate across both languages, unlike the consistent-only fact. This method exposes the crucial difference between surface-level answer consistency and genuine cross-lingual knowledge sharing.

This inconsistency raises questions about how LLMs represent factual knowledge in different languages. On one end of the range of possibilities, models may store a set of distinct knowledge copies for each language. On the other end, models may store a single, shared representation of the factual knowledge and "decode" it into surface forms in different languages. Thus, a shared representation manifests in consistency across languages, but consistent behavior can also occur without it.

While consistency can be readily measured through agreement on identical queries across languages, measuring the extent to which knowledge representation is shared across languages requires more than just evaluating black box outputs. To

quantify representation sharing, we propose editing factual knowledge in one language and examining the effects on other languages. For this purpose, we employ three knowledge editing techniques: two locate-and-edit methods, ROME (Meng et al., 2022a) and MEMIT (Meng et al., 2022b), and a finetuning-based method (Gangadhar and Stratos, 2024). These methods are designed to surgically modify the relevant components of the model responsible for storing factual knowledge and only them, as illustrated in Fig 1.

As a test bed for our experiments, we compiled CLIKE, a multilingual "fill-in-the-blank" factual knowledge probing dataset with 35k samples, designed for evaluation and editing across paraphrases in 13 languages with 7 scripts. We experiment with diverse 7B-parameter LLMs supporting different sets of languages in different setups: monolingual, bilingual, multilingual, and language-extended models. Our analysis reveals a significant disparity in factual knowledge retrieval across languages. We assessed that on average these models answer correctly in at least one language 150% more facts than their best-performing language and triple the number of facts averaged across all 13 languages.

We assess for the first time the extent of cross-lingual knowledge representation sharing. We find that languages within the same script family exhibit the highest degree of representation sharing. This trend is consistent across all the models we studied, regardless of their level of multilingualism. Moreover, we find that high agreement in answers across certain language pairs does not entail shared internal representations, especially with language pairs that do not share the same script. This mismatch is particularly evident when comparing lower-resource languages to any language with a different script, which shows high consistency but limited representation sharing.

We expect that shedding light on these mechanisms will support the development of better multilingual models, with more efficient representation of factual knowledge. This will in turn lead to a more balanced knowledge across different languages, ultimately enhancing LLM performance across languages.

2 Methodology

In our analysis, we would like to measure two main aspects:

1. **Cross-lingual Knowledge Consistency (CKC).** The extent to which a model shows consistency in answering factual questions when asked in different languages.

2. **Cross-lingual Knowledge Representation Sharing (CKR).** The degree to which the model uses a common inner representation for the same fact across different languages.

2.1 Measuring CKC

For simplicity, we say an LLM knows a fact in a specific language, modeled as a question and answer pair, if it can correctly answer it through a query written in this language. We start by defining a model's Knowledge Base (KB) for a specific language, as a set of facts an LLM 'knows' in that language. Formally, for a given LLM M and a dataset $D = \{f_i\}_{i \in [N]}$ of facts. Where $f_i^l := (q_i^l, a_i^l)$ is a question-answer pair written in the language $l \in L$:

$$KB_l := \{f_i \in D \mid M\left(q_i^l\right) = a_i^l\}.$$

To capture the pairwise relationship of knowing a fact in language l_1 to know it in l_2, we define $C_{(l_1,l_2)}$ as the conditional probability

$$P(f_i \in KB_{l_2} \mid f_i \in KB_{l_1}) = \frac{|KB_{l_1} \cap KB_{l_2}|}{|KB_{l_1}|}.$$

We continue by defining the Number of Consistent Languages (NCL) a fact f known in as:

$$NCL(f) = |\{l \in L : f \in KB_l\}|$$

With this aggregation, we can then compute the overall CKC of a model, as the average number of languages in which the LLM knows a fact:

$$\mathbb{E}[NCL] = \frac{1}{N} \sum_{f \in D} NCL(f)$$

2.2 Measuring CKR

Measuring the extent of shared knowledge representation across languages in LLMs cannot be done by merely evaluating model outputs. The same correct answer to a factual query across multiple languages could be generated from distinct, language-specific representations within the model, rather than a unified, language-agnostic abstraction. This requires a more sophisticated approach.

To measure this, we use an editing method E that modifies the model's parameters to provide a

wrong answer for a query in a given language. We then examine the impact of such a change on the same fact query in other languages. Let M_i^l denote the updated model applying E to the fact f_i^l in the language l to the target answer t_i^l. The model's KB for a specific language l' after the modification in language l is defined as the collection of facts for which the incorrect target answer, edited in language l, also propagates to language l', which can be formally expressed as:

$$KB_{l'}^{l} := \{f_i \in KB_{l_1} : M_i^{l_1}\left(q_i^{l_2}\right) = t_i^{l_2}\}.$$

We can then estimate the amount of pairwise CKR between a language l_1 to l_2 by defining $SR_{(l_1,l_2)}$ as the conditional probability

$$P(f \in KB_{l_2}^{l_1} \mid f \in KB_{l_1}^{l_1}) = \frac{\left|KB_{l_1}^{l_2} \cap KB_{l_1}^{l_1}\right|}{\left|KB_{l_1}^{l_1}\right|}.$$

We further define the Number of Transferred Languages (NTL) for a given fact edited in the language l as

$$NTL(f^l) = \left|\{l' : f \in KB_{l'}^{l}\}\right|$$

With this aggregation, we can then compute overall CKR, as the average number of languages in which the LLM represents a fact as

$$\mathbb{E}[NTL] = \frac{1}{|L|}\sum_{l \in L}\left(\frac{1}{|KB_l|}\sum_{f \in KB_l} NTL(f^l)\right).$$

3 Experimental Setup

3.1 Data

Dataset. We present CLIKE (Cross-LIngual Knowledge Editing), a dataset for evaluating and editing factual knowledge of pretrained LMs across languages and paraphrased expressions. CLIKE contains approximately 35k facts spanning 13 languages: English (en), French (fr), Italian (it), Spanish (es), Russian (ru), Ukrainian (uk), Bulgarian (bg), Hindi (hi), Bengali (bn), Chinese (zh), Japanese (ja), Hebrew (he), and Arabic (ar). Each fact is modeled as a language-independent (subject, relation, object) triplet and each relation has 3 paraphrased natural language templates for every language. Each template forms a sentence that conveys a fact and ends with the object, which we omit and expect the model to fill.

For example, the triplet $(Bach, BirthCity, Leipzig)$ will be converted to 'Bach was born in the city of' and 'The birth city of Bach is', and 'The birthplace of Bach is the city of' expecting the pretrained LM to complete the prompt with 'Leipzig' correctly using its initial pretraining task without altering the model with a finetuning intervention.

Fact Collection. Following a similar approach to Petroni et al. (2019); Sciavolino et al. (2021); Kassner et al. (2021); Wei et al. (2024), fact triplets were collected from Wikidata Query Service. We manually crafted and published 14 SPARQL relation queries. Each query extracts wikidata entries for subjects and objects satisfying the query relation with their labels in all available languages. We then filtered all triplets with labels containing less than 8 of the examined languages to balance the languages in the dataset. Appendix B in the appendix includes the languages and relation distributions.

Dataset Construction. We used "Gemini Advanced" and "Claude Opus" to generate the templates of each relation in all languages. For each relation, we generated 3 paraphrases adjusted to grammar rules such as the gender of the subject. The prompts for these templates were executed on the models' official websites (Gemini, Claude). Subsequently, professional translators or native speakers refined the templates and sampled generated fill-in-the-blank queries across all languages, following instructions detailed in Appendix A. For the knowledge editing task, we generated false but plausible objects for each fact by randomly sampling from other facts within the same relation category. This approach provided consistent incorrect alternatives across all languages for each query.

3.2 Models

We examine a range of LLMs with 7B parameters and decoder-only architectures. We focus on base pretrained language models to capture the knowledge acquired during the pretraining process, prior to any finetuning. BLOOM-7B (Scao et al., 2022) serves as our **multilingual** model. Qwen-7B (Bai et al., 2023) represents a **bilingual** Chinese-English model with a low tokenization compression rate multilingual vocabulary. We include two **monolingual** English models: Llama-2-7B (Touvron et al., 2023) and Mistral-7B-v0.1 (Jiang et al., 2023). Additionally, we examine two **language-extended models**, Chinese-llama-2-7B, and Hebrew-Mistral-

7B, based on Llama-2-7B and Mistral-7B-v0.1 with additional pretraining in English and their expanded language (EL) and an expand EL tokenizer vocabulary. These models represent a diverse set of multilingual configurations, enabling a extensive analysis of cross-lingual knowledge representation.

3.3 Knowledge Editing Methods

We employ three knowledge editing methods: Finetuning (FT) (Gangadhar and Stratos, 2024), ROME (Meng et al., 2022a), and MEMIT (Meng et al., 2022b). The ROME and MEMIT editing methods leverage causal mediation analysis (Vig et al., 2020a,b) to identify the LM layer that has causally contributed to factual knowledge recall, suggesting the middle MLP layers act as key-value associative memory. ROME then computes a closed-form rank-one update to the layer's weights, inserting a new fact while minimizing disruption to existing knowledge stored in the weights. Similarly, MEMIT identifies a range of MLP layers that jointly contribute to the model's factual associations. Then it iteratively updates the weights of each MLP layer, distributing the changes across the MLP layers.

Both ROME and MEMIT use interpretability techniques to precisely locate and surgically modify the relevant components of the model responsible for storing factual knowledge. This approach allows for direct control over the model's memorized information while preserving its overall capabilities, providing a framework for isolating changes in actual knowledge without altering other components.

Finetuning, our baseline approach, involves updating the weights of all middle layers in the model without the MLP restrictions imposed by ROME and MEMIT. For each fact to be edited, we finetuned the model on a single example consisting of the edition prompt paired with its new target answer. It incorporates new factual knowledge that resembles standard language model training practices.

We use the EasyEdit code library (Wang et al., 2023b) to perform all language model knowledge edits. Default parameters are employed for all models except BLOOM. Since BLOOM lacks a pre-existing implementation, we optimized and published custom hyperparameters for the editing methods.

3.4 Metrics and Evaluation

We employed the Exact Match (EM) metric to evaluate all answers to queries across our experiments. To provide context for the pretrained LLM, we used 3-demonstrations fewshot concatenated facts for both evaluation and editing tasks, maintaining the same examples and order. All answer generation was performed using greedy decoding to ensure deterministic outputs.

Model performance and CKC in a given language were assessed as follows. The overall accuracy for a language was computed as the percentage of facts correctly answered in at least one paraphrased form. $C(l_1, l_2)$, was measured by computing the mean score in language l_2 across all paraphrases for facts known in language l_1. Similarly, within-language consistency, $C(l, l)$, was computed using the same approach, evaluating the model's consistency across paraphrases within a single language.

For the knowledge editing experiments, we randomly selected 500 known facts in each language to modify. We assessed the effectiveness of these edits using three standard metrics: *Reliability*, *Generalization*, and *Locality*. Reliability measures the accuracy of the model on the edited prompt itself. Generalization, denoted with $SR(l1, l2)$, evaluates the mean score across all paraphrases of the edited fact in all languages, including the language in which the edit was made (l1). This quantifies how well the edit transfers across both languages and paraphrases variations of the fact. For Locality testing, we randomly sampled known facts for each language and evaluated the model's mean accuracy to answer these unrelated queries correctly. This ensured that the edits did not negatively impact other knowledge in different languages.

4 Results

Before presenting our main findings, we first validate the methodology's performance. We found a strong correlation (0.87) between the results obtained from different knowledge editing methods. This consistency across various editing techniques suggests that our findings are robust and not reliant on a specific method. Given this consistency, we primarily present results using MEMIT, with other methods' results available in Appendix C. Additionally, all knowledge editing methods maintained high locality scores (averaging above 70%), indicating that edits were specific and preserved

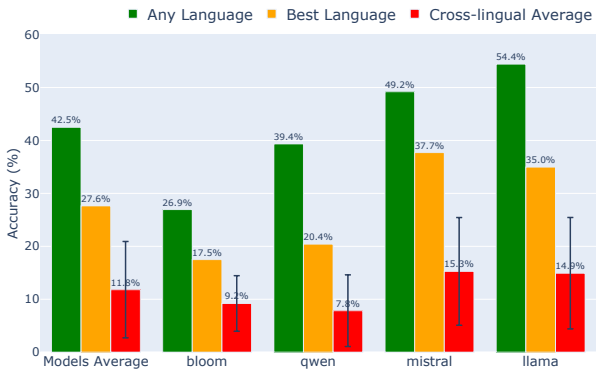

Figure 2: Cross-lingual performance variability: the accuracy of factual knowledge retrieval across different languages for several LLMs supporting different language sets. 'Any Language' (green) – facts known in at least one language, 'Best Language' (orange) – accuracy in the best-performing language, and 'Cross-lingual Average' – mean accuracy across all 13 languages in the CLIKE dataset, with error bars indicating standard deviation.

broader model knowledge. Furthermore, all models exhibit some variation in performance across paraphrases even within the same language, aligning with findings from Mizrahi et al. (2024) and further justifying our approach of assessing knowledge using multiple paraphrases.

4.1 The Issue of Knowledge Variability

Large language models (LLMs) exhibit significant variability in their factual knowledge retrieval across different languages, as illustrated in Fig. 2. Our analysis of four 7B-parameter LLMs reveals a striking disparity: while models demonstrate knowledge of 42.5% of the facts on average in at least one language, their best-performing language achieves only 27.6% accuracy, and their average performance across all 13 languages in the CLIKE dataset is merely 11.8%.

If models could share knowledge across all languages, the best-performing language could potentially increase its accuracy by up to 53%. Moreover, models could then potentially more than triple their current cross-lingual average accuracy. This observation motivates our subsequent investigations into CKR, as we seek to understand and potentially leverage these untapped reservoirs of knowledge.

4.2 Consistency Does Not Imply Representation Sharing

We decouple CKC and CKR between languages, examining both general measures across languages (Fig. 3) and pairwise language relationships addressing their specific identities (Fig. 4). Our analysis reveals that high CKC does not necessarily imply high CKR, and in some cases, we observe inverse patterns.

At the general level, we observe for all models that $\mathbb{E}[NCL]$ is consistently higher than $\mathbb{E}[NTL]$, indicating that models tend to exhibit CKC across more languages than they share representations between. Interestingly, while $\mathbb{E}[NCL]$ values show considerable variation across models, $\mathbb{E}[NTL]$ values are more uniform. The persistent gap between $\mathbb{E}[NCL]$ and $\mathbb{E}[NTL]$ across all models highlights that consistent answers do not necessarily translate to shared internal representations. Moreover, we find that models with a lower proportion of facts known in only one language ($NCL = 1$) tend to have a higher proportion of facts represented in only one language ($NTL = 1$) as shown in Fig. 3.

At the pairwise language level, we find differences between CKC and CKR patterns. For instance, most models (except Qwen) exhibit a high degree of CKC among low-resource languages with different scripts (Chinese, Japanese, Hebrew, Arabic). However, when examining CKR, we find limited evidence of shared encoding between these languages. Conversely, we observe a higher degree of shared representation among Cyrillic languages compared to the shared representation between Cyrillic and Latin languages. This is despite the fact that CKC scores show an opposite trend, with higher CKC between Cyrillic and Latin languages than among Cyrillic languages themselves.

4.3 The Key Role of the Language Script

Our analysis provides quantifiable measures of CKR in LLMs. Following previous work (Qi et al., 2023a; Beniwal et al., 2024), our study highlights the importance of the script of a language for multilingual knowledge. We observe that the pairwise SR measure is relatively consistent across models, despite their varying language support.

We find that languages within the same script family exhibit the highest degree of CKR across all models. As shown in Fig. 4, we observe a script-based grouping in both CKC and CKR likely highlighting a tokenization induced bias (Singh et al., 2019). Notably, we observe strong CKR between languages with Latin scripts (English, French, Italian, Spanish) and between languages with Cyrillic scripts (Russian, Ukrainian, Bulgarian). For Devanagari script languages (Hindi, Bengali) we ob-

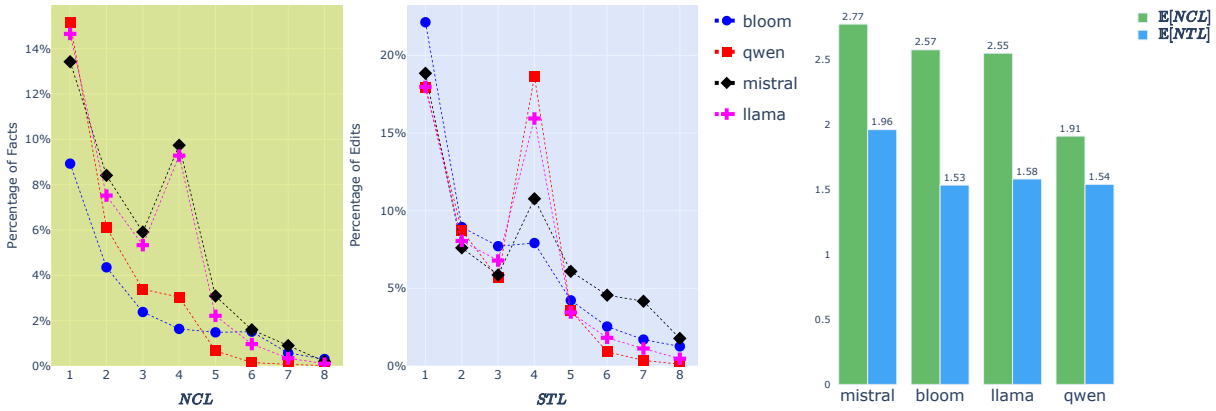

Figure 3: Distribution and Expectation of CKC and CKR. Left: Percentage of facts known (NCL) or represented (NTL) across multiple languages for different models. Right: Expected number of languages per fact (E[NCL]) and expected number of languages sharing representation per edited fact (E[NTL]) for each model, illustrating the relationship between knowledge CKC and CKR.

serve relatively high CKR for models that perform well on these languages (Bloom, Mistral).

While most CKR occurs among languages that use the same script, there is still some knowledge transfer between languages with different scripts. This cross-script transfer is particularly evident between Cyrillic and Latin script languages across various models. Additionally, in specific cases, such as with the BLOOM model, we observe a moderate degree of CKR between seemingly unrelated language pairs, e.g., 28% from Italian to Hindi.

Notably, these relations between language scripts are sometimes asymmetrical. For example, knowledge in Cyrillic script languages implies a higher probability (approximately 40-60%) of knowing the same facts in Latin script languages. However, the reverse relation is weaker, with only about 10-20% probability of Cyrillic knowledge given Latin script knowledge. A similar asymmetrical relation appears across models suggesting a stronger transfer of knowledge from Cyrillic to Latin. We hypothesize that the dominance of Latin script languages, especially English, in the training data leads to more robust fact representations in Latin scripts, facilitating easier transfer from Cyrillic to Latin than vice versa.

4.4 Impact of Model Design Languages

How does a model's designed language support affect its CKR and CKC patterns? Although the patterns of CKR are relatively similar across models supporting different language sets, our analysis reveals some nuanced differences.

The multilingual BLOOM model demonstrates the highest pairwise average of language pairwise CKC (36%) and CKR (8.4%) across different script pairs. As shown in Fig. 4, BLOOM exhibits notable transfer between seemingly unrelated language pairs. These cross-script patterns validate BLOOM's design as a multilingual model, emphasizing its cross-lingual relationships rather than its overall low accuracy performance.

We find that the bilingual English-Chinese Qwen model is showing relatively high overall accuracy in Chinese. However, this Chinese knowledge remains largely distinct from English both in terms of language pairwise CKC and CKR Fig. 4. This pattern validates Qwen's design as a bilingual model, emphasizing its language-specific capabilities rather than cross-script knowledge sharing. Surprisingly, when examining the global shared representation, Qwen exhibits a higher number of 4-lingual representations sharing (NTL = 4) from uniquely represented facts. Although Qwen lacks cross-script knowledge sharing, it developed some degree of multilingual representation, particularly within script families.

Monolingual English models (Mistral, LLaMA) exhibit a unique pattern. We discover an anomalous peak in the results for facts known and represented in exactly four languages (Fig. 3), corresponding primarily to the four Latin script languages in our dataset. This highlights the strong association between script similarity and knowledge sharing even in ostensibly monolingual models. Surprisingly, Mistral demonstrates the highest $\mathbb{E}[NCL]$, $\mathbb{E}[NTL]$ and average of language pairwise CKC (54.7%) and CKR (37.6%) within script families, despite being designed as a monolingual

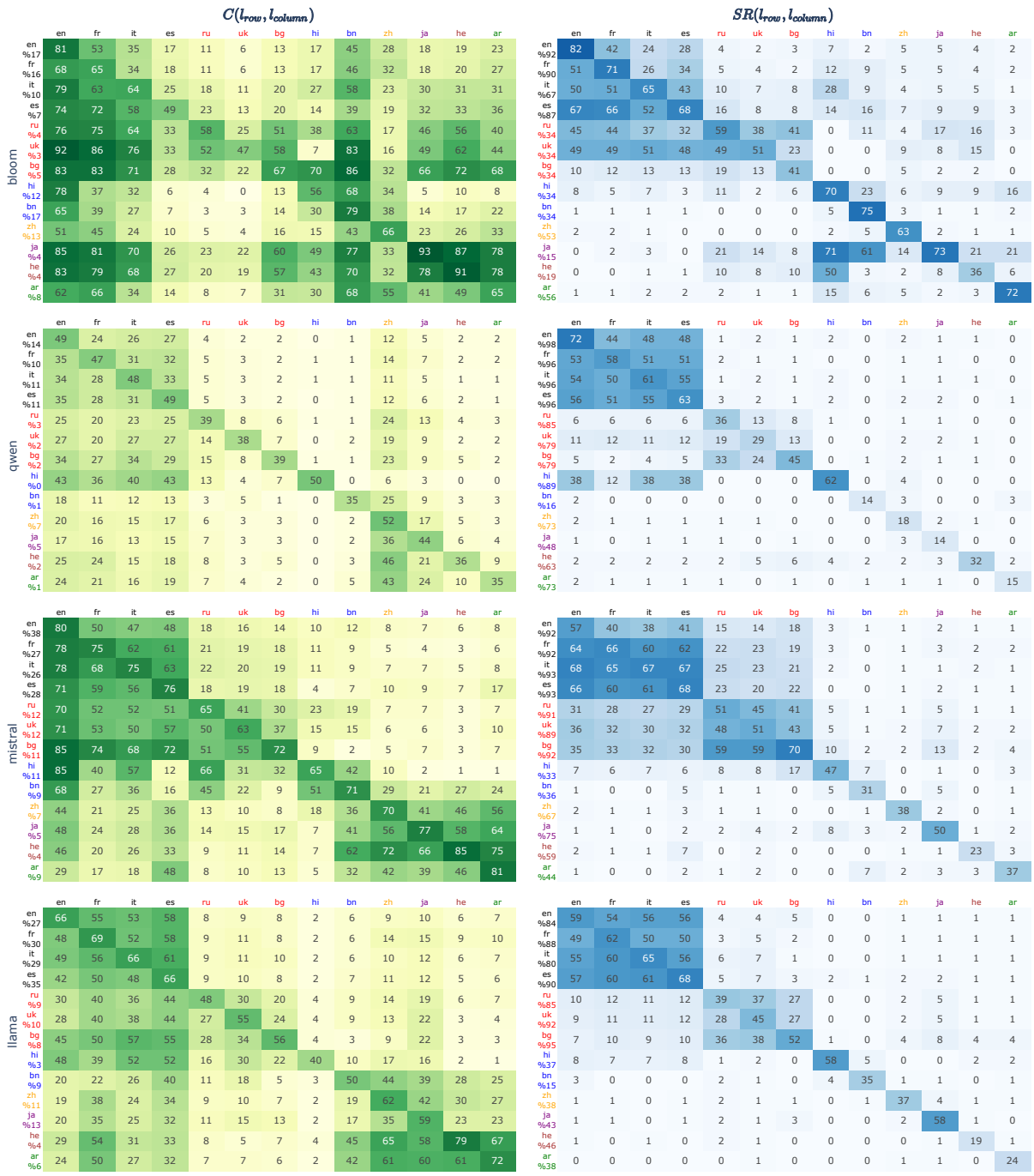

Figure 4: The pairwise relationships of factual knowledge across languages for four language models, with all scores reported using Exact Match. $C(l_1, l_2)$ shows the percentage of facts known in the row language which were also retrieved in the column language. $SR(l_1, l_2)$ indicates the percentage of successfully edited facts in the row language which generalized to the column language using MEMIT. Under each language abbreviation is the overall accuracy for initial knowledge retrieval and the edition-reliability score for C and SR measures respectively. Languages are color-coded by script family.

English model. This result highlights how Mistral's strong English foundation naturally extends to other Latin script languages, underscoring the impact of script similarity on cross-lingual knowledge representation even in monolingual models.

4.5 Language Extended LMs

How does additional pretraining on both English and an extended language (EL) impact cross-lingual CKC and shared representation in initially monolingual models? Analysis of chinese-llama-2-7b and he-mistral-7b reveals a similar trade-off:

	Model	EL Acc/Rel	En Acc/Rel	EL → en	en → EL
C	zh EL	10 (142%↑)	5 (29%↓)	12 (80%↓)	22 (440%↑)
C	he EL	18 (600%↑)	10 (32%↓)	13 (37%↓)	18 (900%↑)
SR	zh EL	96 (252%↑)	81 (96%↓)	2 (200%↑)	4 (400%↑)
SR	he EL	90 (152%↑)	83 (90%↓)	10 (500%↑)	6 (600%↑)

Table 1: CKC and CKR in extended LLMs compared to their base models. EL: Extended Language, Acc: Accuracy, and Rel: Reliability.

while gaining substantial knowledge in EL, models sacrifice much of their original English expertise. These extensions reshape cross-lingual knowledge distribution but fall short of fully bridging the gap between disparate writing systems.

As shown in Table 1, both models exhibit increased accuracy in the extended language (EL) coupled with decreased English accuracy. CKC measures paint a nuanced picture: models acquire extensive new knowledge in EL, largely unknown in English, yet this new EL knowledge covers more of English knowledge. Shared representation metrics underscore this asymmetry. Despite increased bidirectional knowledge transfer between English and EL, transfer remains stubbornly low. This suggests that even with targeted pretraining, models struggle to forge robust representations sharing across linguistically distant languages.

For analysis of how different relation types affect CKR, see Appendix D.

5 Related Work

Cross-lingual Knowledge Consistency. While monolingual knowledge consistency has been studied often in LMs (Elazar et al., 2021; Mizrahi et al., 2024), limited work has been done on cross-lingual knowledge consistency. Qi et al. (2023b) proposed a cross-lingual consistency metric named RankC to measure similarity across multiple candidate answers, whether correct or incorrect. Our focus on correct answers allows a simpler assessment without being limited to pairwise language comparisons.

Cross-lingual Knowledge Representation Sharing. Previous studies explored this angle through different approaches. Some works studied parameter sharing across languages by analyzing neuron activation/deactivation when evaluating knowledge in different languages (Libovický et al., 2020; Zhao et al., 2024b; Chen et al., 2024; Tang et al., 2024; Kojima et al., 2024). Enhancing language-independent neurons resulted in better multilingual abilities in a specific language without compromising others. Other works investigated the knowledge related to the training data and identified the language source of the acquired data (Choenni et al., 2023; Zhao et al., 2024a), providing evidence that knowledge from training data in one language can benefit the model in other languages. Another line of work analyzed how inputs in different languages affect the activation patterns, showing that semantically equivalent content in different languages tends to produce similar activation patterns (Singh et al., 2019; Libovický et al., 2020; Chang et al., 2022).

These works pointed to a connection between knowledge in different languages. However, they do not yield an assessment of the amount of shared knowledge. While passive analysis can take as far as measuring the similarity between languages, active modification tools can also suggest a clear causal relation between the knowledge representation in different languages.

Multilingual Knowledge Editing. Previous work on multilingual knowledge editing (Si et al., 2024; Xu et al., 2022; Wei et al., 2024; Wang et al., 2023a) primarily focused on comparing and improving editing methods' performance in multilingual settings. Our approach is different. We use these editing tools as analytical tools to understand representation sharing across languages and across models with different multilingual configurations.

6 Conclusion

This work investigated the relationship between cross-lingual knowledge consistency and representation sharing in LLMs. Our findings reveal that high consistency across languages does not necessarily imply shared internal representations, particularly for languages with different scripts. We introduced a novel methodology and dataset for quantifying these phenomena, providing a more nuanced understanding of how LLMs represent and retrieve factual knowledge. The significant disparity we observed in factual knowledge retrieval across languages, coupled with the potential for substantial performance improvements if knowledge could be fully shared, underscores the importance of developing more effective multilingual knowledge representations. We expect that our insights will guide the development of more efficient and equitable multilingual models, ultimately enhancing their performance across all languages.

Limitations

Our main limitation lies in the constraints imposed by our chosen editing methods and their focus on specific model components. By primarily targeting middle layers associated with factual knowledge storage, our analysis may have overlooked important cross-lingual interactions occurring elsewhere in the model architecture. Our reliance on specific editing techniques (ROME, MEMIT, and Finetuning) may not capture the full spectrum of knowledge representation and modification within the model. There might be multiple pathways to change the output in a specific language, potentially exhibiting different cross-lingual generalization patterns than those we observed.

Our analysis focused exclusively on decoder-only language models with 7B parameters, limiting the generalizability of our findings across different architectures and sizes. Similarly, while our CLIKE dataset covers a diverse range of languages and relations, it may not fully represent the breadth of factual knowledge or linguistic phenomena. These constraints in both model selection and dataset composition could influence the observed patterns of cross-lingual representation.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

Himanshu Beniwal, Mayank Singh, et al. 2024. Cross-lingual editing in multilingual language models. *arXiv preprint arXiv:2401.10521*.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. 2024. How do large language models acquire factual knowledge during pretraining? *arXiv preprint arXiv:2406.11813*.

Tyler A Chang, Zhuowen Tu, and Benjamin K Bergen. 2022. The geometry of multilingual language model representations. *arXiv preprint arXiv:2205.10964*.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17817–17825.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. How do languages influence each other? studying cross-lingual data sharing during llm fine-tuning. *arXiv preprint arXiv:2305.13286*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard H. Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. *arXiv preprint arXiv:2203.11552*.

Govind Gangadhar and Karl Stratos. 2024. Model editing by pure fine-tuning. *Preprint*, arXiv:2402.11078.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Xiaoze Jiang, Yaobo Liang, Weizhu Chen, and Nan Duan. 2022. Xlm-k: Improving cross-lingual language model pre-training with multilingual knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10840–10848.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-factr: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual lama: Investigating knowledge in multilingual pretrained language models. *arXiv preprint arXiv:2102.00894*.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. *arXiv preprint arXiv:2404.02431*.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. *arXiv preprint arXiv:2004.05160*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. In *Neural Information Processing Systems*.

Kevin Meng, Arnab Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *ArXiv*, abs/2210.07229.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? In *Conference on Empirical Methods in Natural Language Processing*.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023a. Cross-lingual consistency of factual knowledge in multilingual language models. *arXiv preprint arXiv:2310.10378*.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023b. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, 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 Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine 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 vSavsko, 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 L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. Longpre, Somaieh Nikpoor, S. 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, 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 V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Y 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 Franccois Lavall'ee, 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, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, 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, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenvek Kasner, Zdeněk Kasner, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ayoade Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatim Tahirah Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Lívia 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, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, R. 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 Pratap Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yu Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nianwen Si, Hao Zhang, and Weiqiang Zhang. 2024. Mpn: Leveraging multilingual patch neuron for cross-lingual model editing. *ArXiv*, abs/2401.03190.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. Bert is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020a. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020b. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yu Cao, and Jiarong Xu. 2023a. Cross-lingual knowledge editing in large language models. *ArXiv*, abs/2309.08952.

Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Guozhou Zheng, et al. 2023b. Easyedit: An easy-to-use knowledge editing framework for large language models. *arXiv preprint arXiv:2308.07269*.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2024. Mlake: Multilingual knowledge editing benchmark for large language models.

Yang Xu, Yutai Hou, and Wanxiang Che. 2022. Language anisotropic cross-lingual model editing. *ArXiv*, abs/2205.12677.

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

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024b. How do large

language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

A Native Speaker Instruction

Dear <Native Speaker>,

We are reaching out to you for assistance in an important project that aims to improve the ability of Artificial Intelligence (AI) to understand and generate text in your native language. Your skills and knowledge as a native speaker are crucial to the success of this project.
Our research team has created a collection of fill-in-the-blank sentences and templates in multiple languages, including yours. These sentences will be used to evaluate the knowledge and understanding of AI language models. To ensure the accuracy and effectiveness of our collection, we need your help in verifying the grammatical correctness of the sentences and templates we have created.
Attached, you will find a list of approximately 60 simple sentences and sentence templates and templates that cover various relationships between subjects and objects in your native language. The task should take no more than 15-20 minutes to complete. Your task is to review each sentence and template and determine whether they are grammatically correct. If you find any grammatical errors, please provide a corrected version of the template. Additionally, if you wish, you may provide an optional explanation in English of what was wrong with the original template.

Example:

Relation: *Birth City*, Subject: *Wolfgang Amadeus Mozart*, and Object: *Salzburg*
Original template:
"[subj] birthplace the city [obj]" -> "Wolfgang Amadeus Mozart birthplace the city Salzburg"
Fixed template:
"[subj]'s birthplace is the city of [obj]" -> "Wolfgang Amadeus Mozart's birthplace is the city of Salzburg"
Explanation:
The original template is missing the verb "is" and the preposition "of" to form a grammatically correct sentence.

When fixing the templates, please keep in mind the following guidelines:

1. Be explicit about the relationship to avoid ambiguity. For example, given the information (Bach, Birth Year, 1685) and the template "[subj] born in [obj]", the AI might complete the prompt "Bach was born in" with the object "Leipzig" (his birth city) or "31 March" (his birth date) rather than the year "1685". Therefore, a good template should contain words that explicitly describe the relationship. The template "Bach was born in the year [obj]" will likely output "1685" since the word "year" appeared in the sentence.

2. For each relationship, we have provided three different prompt paraphrases that are supposed to be different from one another.

3. The subject must always appear before the object.

4. The last word of the prompt template should be the object.

Please note that the sentences do not necessarily need to sound natural or be well-written. Our primary focus is on ensuring that they are grammatically correct. However, if you have suggestions on how to make the sentences sound more natural without changing the core structure, feel free to include them in your feedback.
Your contribution to this project is valuable and will help us create a reliable collection of sentences for advancing AI's ability to understand and generate text in your language. If you have any questions or concerns about the project or your role in it, please don't hesitate to reach out to us.
Thank you for your participation in this project. We look forward to receiving your feedback.
Best regards,

B CLIKE Dataset Key statistics.

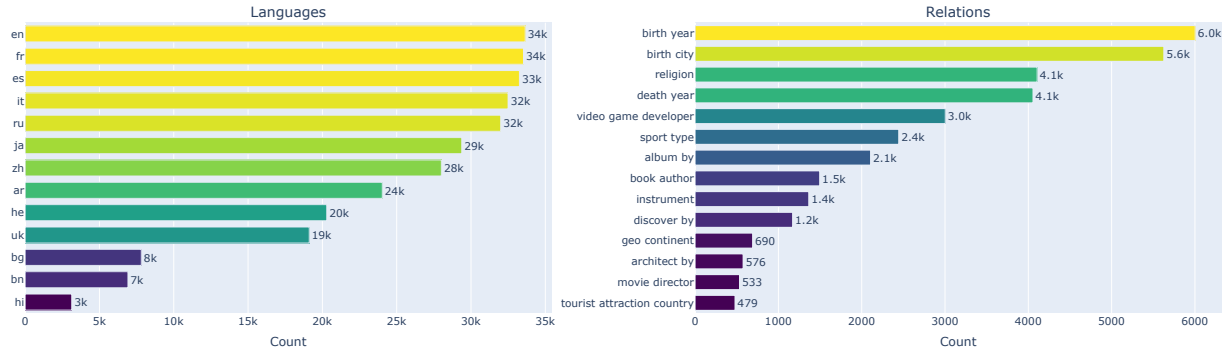

Figure 5: Left: histogram of number of examples for each language, right: histogram of number of examples for each relation type.


```
{'id': '11',
  'subj': {'label': {'it': 'Manto Mavrogenous', 'ar': 'مانتو مافروجينوس', 'ru': 'Манто Мавроген ус', 'bn': 'মান্টো মাভরোগেনি'
           'qid': 'Q269970',
           'origin': 'it',
           'gender': 'f'},
  'rel': {'label': 'birth_city',
           'qid': 'P19'},
  'obj_true': {'qid': 'Q546',
               'label': {'it': 'Trieste', 'ar': 'تريستي', 'ru': 'Триест', 'bn': 'ত্রিয়েস্তে', 'uk': 'Трієст', 'he': 'טריאסטה',
  'prompt': {'it': 'Manto Mavrogenous è nata nella site di [obj]', 'ar': '[obj] ولدت مانتو مافروجينوس في مدينة', 'ru': '
  'paraphrase_prompts': {'it': ['La città natale di Manto Mavrogenous è [obj]', 'Il luogo di nascita di Manto Mavrogenou
  'target_true': {'en': 'Beijing', 'he': "בייג'ינג", 'fr': 'Pékin', 'ar': 'بكين', 'ru': 'Пекин', 'it': 'Pechino', 'es':
}
```


Figure 6: A Dataset Sample Example

C ROME and FT Methods Performance Comparison

$SR(l_{row}, l_{column}) - ROME$

bloom	en	fr	it	es	ru	uk	bg	hi	bn	zh	ja	he	ar
en %78	44	25	19	21	4	5	2	6	2	6	1	1	3
fr %72	38	37	26	30	6	4	4	11	6	5	2	3	4
it %51	40	33	46	28	12	11	2	7	2	5	0	2	2
es %74	52	47	46	48	14	10	7	7	6	6	3	2	6
ru %47	39	35	38	27	54	35	15	9	0	4	2	9	1
uk %24	45	40	61	48	51	71	14	0	0	10	6	17	2
bg %8	15	13	29	22	21	26	23	0	0	0	0	2	0
hi %61	9	6	7	4	1	0	2	38	13	4	1	2	3
bn %47	5	4	3	2	1	2	2	5	32	3	2	2	3
zh %45	2	3	2	1	0	0	0	2	3	44	1	0	1
ja %11	0	1	1	1	1	0	0	0	0	3	32	1	0
he %6	0	0	1	0	2	0	0	0	0	0	3	55	1
ar %27	1	3	4	2	1	0	0	11	2	3	0	0	41

qwen	en	fr	it	es	ru	uk	bg	hi	bn	zh	ja	he	ar
en %97	81	51	55	56	3	3	3	2	0	3	2	1	0
fr %95	64	69	61	62	4	3	2	0	0	1	1	1	0
it %96	64	62	79	68	5	4	2	3	0	2	1	1	1
es %96	65	60	67	76	4	3	3	2	0	2	2	0	0
ru %90	9	9	10	9	48	21	18	1	0	2	1	0	0
uk %83	14	16	17	17	25	37	20	0	0	5	4	2	1
bg %93	6	5	4	6	43	31	55	0	0	7	9	0	0
hi %89	42	21	46	29	0	0	0	88	0	0	0	0	0
bn %11	0	0	4	0	0	0	0	0	25	0	0	0	0
zh %93	1	1	1	2	1	1	0	0	0	24	2	0	0
ja %85	1	0	0	0	1	0	0	0	0	2	29	1	0
he %96	2	1	1	1	3	4	4	0	0	0	3	32	2
ar %99	1	1	0	1	0	0	1	0	1	1	0	0	20

mistral	en	fr	it	es	ru	uk	bg	hi	bn	zh	ja	he	ar
en %90	66	42	41	44	10	8	8	0	1	1	2	1	2
fr %91	70	69	64	64	16	18	13	1	0	2	2	2	2
it %90	71	67	70	68	16	15	11	5	0	2	2	2	1
es %90	70	63	64	71	16	15	17	0	0	1	2	1	2
ru %86	26	20	19	22	52	45	40	3	1	1	3	1	1
uk %87	27	23	20	22	43	48	41	4	0	2	4	2	1
bg %88	21	20	18	18	53	55	69	5	0	3	7	2	3
hi %30	6	3	5	5	4	4	11	35	4	0	0	1	1
bn %32	2	1	1	7	0	1	0	5	21	1	3	0	3
zh %60	3	2	2	4	1	1	0	0	1	35	3	1	2
ja %69	1	0	0	2	3	3	5	0	1	2	43	1	2
he %48	1	1	1	7	1	3	0	0	0	1	2	21	5
ar %33	2	0	0	1	3	3	0	17	6	4	4	5	34

llama	en	fr	it	es	ru	uk	bg	hi	bn	zh	ja	he	ar
en %77	62	54	56	55	3	4	3	0	0	2	2	1	0
fr %82	50	64	52	53	3	4	3	1	1	1	1	1	1
it %80	60	61	68	63	5	7	3	0	0	1	2	1	1
es %87	59	60	61	69	3	4	3	2	1	2	1	1	1
ru %91	7	8	7	8	36	34	20	0	1	2	4	1	1
uk %94	6	9	8	10	29	41	25	1	0	3	5	1	1
bg %95	3	6	4	6	32	33	49	0	0	4	9	4	4
hi %42	15	6	13	12	0	0	0	67	2	0	3	0	1
bn %19	3	0	0	0	3	0	0	4	20	0	1	0	1
zh %46	1	1	0	1	2	1	2	0	0	38	4	0	1
ja %51	1	1	0	1	1	1	2	0	1	2	63	1	1
he %45	2	1	1	1	2	1	2	0	0	0	0	16	1
ar %40	0	0	0	0	0	0	0	0	0	1	0	1	23

$SR(l_{row}, l_{column}) - Finetuning$

bloom	en	fr	it	es	ru	uk	bg	hi	bn	zh	ja	he	ar
en %99	6	4	3	2	1	0	0	0	0	1	1	1	1
fr %99	5	7	7	5	2	0	0	5	1	2	1	2	1
it %99	4	4	5	2	1	1	0	0	0	1	0	1	0
es %98	11	12	13	10	4	2	1	1	3	2	2	2	0
ru %97	4	4	4	1	5	5	0	0	0	1	1	1	0
uk %96	2	2	3	1	2	4	0	0	0	2	0	1	0
bg %87	0	0	0	1	0	1	2	0	0	2	0	1	0
hi %98	1	1	1	0	0	0	0	2	1	2	1	2	0
bn %97	1	1	1	1	1	1	0	0	2	2	1	2	0
zh %99	1	1	0	0	0	0	0	0	1	1	0	1	0
ja %87	0	0	0	1	0	0	0	0	0	1	0	0	0
he %81	0	0	0	0	0	0	0	0	0	1	0	1	0
ar %98	0	0	0	0	0	0	0	1	0	1	0	0	2

qwen	en	fr	it	es	ru	uk	bg	hi	bn	zh	ja	he	ar
en %98	27	17	19	19	1	1	1	0	0	1	1	0	0
fr %97	19	20	19	17	1	0	1	1	0	0	1	0	0
it %97	18	16	20	19	1	1	1	1	0	0	1	1	0
es %98	20	19	20	23	1	0	1	0	0	1	1	0	1
ru %92	1	1	1	1	13	4	3	1	0	1	1	0	0
uk %89	4	4	4	4	4	10	6	0	0	1	2	0	0
bg %96	0	0	0	1	5	6	13	0	0	0	0	0	0
hi %89	12	0	8	12	0	0	0	19	0	0	0	0	0
bn %38	2	1	1	1	0	0	0	0	1	0	1	0	1
zh %97	0	0	1	1	0	1	1	0	0	3	1	0	0
ja %93	0	0	0	0	0	0	0	0	0	0	2	0	0
he %96	2	2	0	1	2	2	3	0	0	0	2	3	1
ar %99	1	1	1	1	1	0	1	0	1	0	1	0	3

mistral	en	fr	it	es	ru	uk	bg	hi	bn	zh	ja	he	ar
en %98	30	15	16	16	4	3	2	1	1	1	1	1	1
fr %98	23	25	22	22	6	7	6	3	0	2	2	2	2
it %99	18	17	18	17	5	4	3	1	1	2	2	1	1
es %98	25	20	20	23	6	4	5	1	1	1	1	1	1
ru %94	10	8	9	10	25	23	17	3	2	2	2	2	1
uk %94	14	12	12	13	25	23	19	1	1	3	5	2	2
bg %99	15	15	15	15	33	28	38	6	2	4	6	4	4
hi %33	1	1	3	2	2	7	3	16	3	2	1	1	2
bn %39	1	0	0	7	0	1	0	5	15	0	4	0	1
zh %94	1	1	1	3	0	1	1	0	1	13	3	1	1
ja %84	1	1	1	1	2	1	1	0	0	2	21	2	2
he %64	1	1	0	6	1	1	2	0	2	1	2	9	1
ar %47	1	0	0	2	3	3	2	0	4	1	4	3	21

llama	en	fr	it	es	ru	uk	bg	hi	bn	zh	ja	he	ar
en %98	22	18	19	20	1	1	1	0	0	1	2	1	1
fr %95	17	18	18	18	1	1	1	0	0	1	1	1	1
it %97	22	22	23	22	2	2	2	1	0	1	1	1	1
es %95	19	19	20	21	1	1	1	0	1	1	1	1	1
ru %95	2	2	2	2	15	10	8	0	0	1	1	1	1
uk %96	1	2	2	2	7	12	4	0	0	2	1	1	1
bg %98	1	2	2	2	10	9	19	0	0	2	3	4	4
hi %55	1	0	0	0	1	2	2	28	0	0	2	0	1
bn %22	2	1	0	0	3	1	0	3	4	1	1	0	0
zh %75	0	0	0	0	0	0	1	1	0	9	1	0	0
ja %75	0	0	0	0	0	0	0	0	1	0	8	1	1
he %60	1	0	0	0	1	0	0	0	0	0	0	4	0
ar %54	0	0	0	0	1	0	1	0	0	1	0	0	6

D On CKR Features

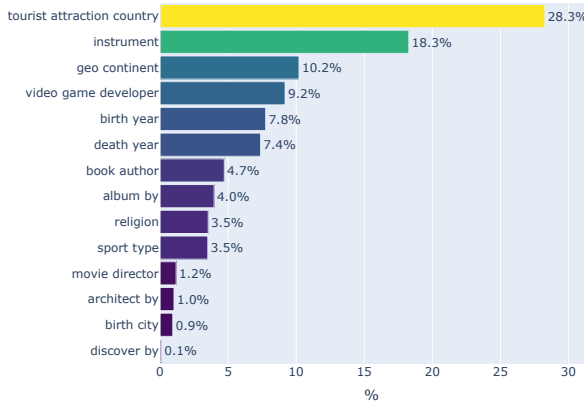

Figure 7: Distribution of shared representation facts across different relation types.

Figure 7 illustrates that relations with fewer possible categories generally exhibit higher CKR. This trend is evident for relations such as countries, instruments, continents, and company developers, with sports type and religion as notable exceptions. Numerical relations like birth year and death year also demonstrate strong transfer capabilities. In contrast, relations involving names (e.g., book authors, movie directors, discoverers) and those with numerous categories (e.g., cities) show lower transfer rates.