Lost in Multilinguality: Dissecting Cross-lingual Factual Inconsistency in Transformer Language Models

Mingyang Wang[1,2,3] **Heike Adel**[4] **Lukas Lange**[1]
Yihong Liu[2,3] **Ercong Nie**[2,3] **Jannik Strötgen**[5] **Hinrich Schütze**[2,3]
[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]LMU Munich, Germany [3]Munich Center for Machine Learning (MCML)
[4]Hochschule der Medien, Stuttgart, Germany
[5]Karlsruhe University of Applied Sciences, Germany
mingyang.wang2@de.bosch.com

Abstract

Multilingual language models (MLMs) store factual knowledge across languages but often struggle to provide consistent responses to semantically equivalent prompts in different languages. While previous studies point out this cross-lingual inconsistency issue, the underlying causes remain unexplored. In this work, we use mechanistic interpretability methods to investigate cross-lingual inconsistencies in MLMs. We find that MLMs encode knowledge in a language-independent concept space through most layers, and only transition to language-specific spaces in the final layers. Failures during the language transition often result in incorrect predictions in the target language, even when the answers are correct in other languages. To mitigate this inconsistency issue, we propose a linear shortcut method that bypasses computations in the final layers, enhancing both prediction accuracy and cross-lingual consistency. Our findings shed light on the internal mechanisms of MLMs and provide a lightweight, effective strategy for producing more consistent factual outputs.

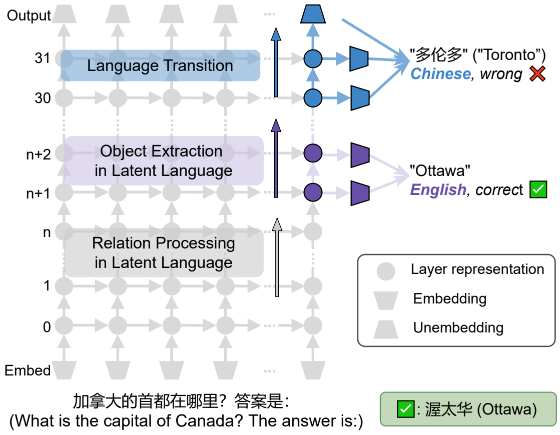


Figure 1: Illustration of language transition failure in LLaMA2 when answering the question: “加拿大的首都在哪里?答案是：” (“What is the capital of Canada? The answer is:”). In intermediate layers, the model processes information in its latent language, i.e., a concept space independent of the input language.[1] While it correctly identifies “Ottawa” in English during the concept-space object extraction, the final output “多伦多” (“Toronto”) is incorrect after transitioning to Chinese. This indicates the model’s failure to adapt knowledge from the concept space to the target language, leading to cross-lingual inconsistency.

1 Introduction

Multilingual language models (MLMs) have shown remarkable capabilities in storing and retrieving factual knowledge across languages (Jiang et al., 2020; Kassner et al., 2021). However, they often exhibit inconsistencies when responding to semantically equivalent prompts in different languages. For instance, an MLM might correctly predict the capital of Canada when asked in English but fail to do so when queried in another language, e.g., Chinese. This phenomenon is known as *cross-lingual factual inconsistency* (Qi et al., 2023). It raises questions about how effectively MLMs transfer knowledge across languages, and shows limitations in their robustness and fairness.

Understanding the root causes of such inconsistencies is crucial, yet research in this area remains limited. While prior studies have explored the inner workings of MLMs (Wendler et al., 2024; Dumas et al., 2024; Fierro et al., 2024), they mainly focus on scenarios where models make correct predictions, leaving the reasons behind inconsistent predictions unexplored. Furthermore, while Qi et al. (2023) identify frequent cross-language inconsistencies in MLMs, they do not investigate the underlying causes behind them.

In this work, we address this research gap by analyzing cross-lingual factual inconsistency through the lens of mechanistic interpretability (Olah, 2022; Nanda et al., 2023), which aims at

[1] This concept space in LLaMA2, as seen through the Logit Lens (Nostalgebraist, 2020), exhibits a bias towards English, reflecting its English-centric nature (Wendler et al., 2024).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5075–5094
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

reverse-engineering and, thereby, understanding language models. We trace information flows within MLMs to identify where inconsistencies arise on two complementary scenarios: (1) cases where models produce correct predictions consistent with English and (2) cases where models predicts correctly in English but generates incorrect answers in other languages.[2] This comparison aims at uncovering the causes of both success and failure in multilingual factual recall.

Our analysis reveals that MLMs process factual knowledge in a concept space largely independent of the input language through most layers, and transition to language-specific spaces in the final layers. However, even when the correct prediction is encoded in this concept space, the model can fail the language transition, leading to incorrect predictions in the target language (see Figure 1). This highlights the critical role of the language transition mechanism for cross-lingual consistency.

Overall, our contributions are as follows:

(i) **Dataset Construction (§3)**: We introduce KLAR, an enhanced **K**now**L**edge probing dataset for **A**uto-**R**egressive models, covering 17 languages and 20 relation types. It provides a robust framework for multilingual knowledge probing, which we use to evaluate the cross-lingual consistency of two state-of-the-art MLMs (§4).

(ii) **Mechanistic Analysis (§5)**: We conduct the first interpretability-driven study of cross-lingual factual inconsistency, revealing how MLMs encode and process factual knowledge across layers.

(iii) **Failure Mode Identification (§6)**: In a detailed layer-wise analysis, we identify the language transition mechanism as main failure point that leads to cross-lingual inconsistency.

(iv) **Approach (§7)**: We propose a shortcut method that bypasses the model's final-layer computations, enhancing both prediction accuracy and cross-lingual consistency in MLMs.[3]

2 Related Work

Mechanistic Interpretability (MI) aims to understand LLMs by decomposing their computations into smaller, interpretable components. It has gained significant attention for studying factual knowledge recall in LLMs (Meng et al., 2022; Dai et al., 2022; Geva et al., 2023; Yu et al., 2023; Lv et al., 2024; Wang et al., 2024; Liu et al., 2025).

Following Olah et al. (2020) and Rai et al. (2024), MI research is categorized into the study of *features*, which capture human-interpretable properties in model representations or components like neurons and attention heads (Elhage et al., 2022; Gurnee et al., 2023), and the study of *circuits*, which refer to subgraphs of the model's computation graph responsible for implementing specific behaviors (Wang et al., 2023; Elhage et al., 2021).

In this work, we focus on representation-level feature-based interpretability analysis to interpret the behavior of multilingual language models in the knowledge probing task. Specifically, we use Logit Lens (Nostalgebraist, 2020) to project latent state representations of LMs into the vocabulary space, enabling the analysis of intermediate representations and tracking how information evolves across layers.

Interpreting Multilingual Language Models. Recent studies have explored the internal workings of MLMs. Wendler et al. (2024) examine the latent language of LLaMA2 models using controlled translation, completion, and cloze tasks, finding that LLaMA2 internally relies on English as a pivot language. Building on this setup, Dumas et al. (2024) investigate the disentanglement of language and concept representations, demonstrating that LLaMA2 processes language and concept information independently. Fierro et al. (2024) analyze knowledge probing tasks to study how mechanisms identified in monolingual contexts generalize to multilingual settings, but their focus remains limited to correct prediction cases.

In contrast, our work centers on understanding the internal mechanisms responsible for cross-lingual inconsistencies. By examining both consistent and inconsistent predictions, we uncover how MLMs transition from language-independent to language-specific processing. This approach offers new insights into how MLMs encode and transfer factual knowledge across languages, addressing a key gap in prior research.

3 KLAR Dataset

We focus on the factual knowledge probing task, where a fact is represented as a subject-relation-object triple $\langle s_i, r_i, o_i \rangle$ and expressed in natural language prompts. Given a prompt constructed from the subject s_i and relation r_i, LMs are ex-

[2] English serves as the pivot language due to its central role in many multilingual language models (Held et al., 2023; Zhang et al., 2023).

[3] Our data and code are open-source at `https://github.com/boschresearch/KLAR-CLC`

pected to predict the object o_i. For example, the fact ⟨*Canada, capital, Ottawa*⟩ can be queried as, "What is the *capital* of *Canada*?", and the model should predict the object *Ottawa* as the answer.

Qi et al. (2023) introduce the BMLAMA17 dataset for evaluating multilingual factual knowledge in MLMs. However, in many factual questions in BMLAMA17, the object appears in the middle of the sentence rather than at the end, which is incompatible with knowledge probing for auto-regressive models. Furthermore, BMLAMA17 includes many relations with multiple correct answers,[4] making it difficult to reliably evaluate the correctness of a model's response for a given $\langle s_i, r_i, o_i \rangle$ triple where o_i is only one of the possible answers.

To address these limitations, we construct KLAR, a **K**now**L**edge probing dataset that ensures compatibility with **A**uto-**R**egressive models and provides clarity in factual evaluation. We extract parallel factual knowledge triples in 17 languages from BMLAMA17 and design prompts where the object consistently appears at the end. Relation-specific templates are structured as "<*Question*> The answer is:", e.g., ⟨*Canada, capital, Ottawa*⟩ becomes: "What is the capital of Canada? The answer is:". These templates are initially created in English and translated into 16 other languages using `gpt-3.5-turbo`. To ensure clarity, we exclude relations with multiple correct answers and inspect the semantic clarity in prompt templates manually and/or through back-translation.

The resulting KLAR dataset includes 2,619 parallel factual knowledge triples in 17 languages, covering 20 relation types. Table 1 provides an overview of the languages and sample relations. Detailed statistics are provided in Appendix A.1.

4 Cross-lingual Consistency Evaluation

Models and Languages We analyze two widely used open-source multilingual auto-regressive language models: LLaMA2-7B (Touvron et al., 2023) and BLOOM-560M (Le Scao et al., 2023). LLaMA2 is trained on a multilingual corpus dominated by English, which accounts for 89.7% of the data, whereas BLOOM's training data is more balanced, with English comprising 31.3% of the corpus. Our analysis considers the languages shared between each model and our dataset, covering 12 languages for LLaMA2 and 7 for BLOOM. Details on the selected languages are provided in Table 4 in Appendix A.1.

Languages (17)	
Arabic (*ar*), Catalan (*ca*), Greek (*el*), English (*en*), Spanish (*es*), Persian (*fa*), French (*fr*), Hebrew (*he*), Hungarian (*hu*), Japanese (*ja*), Korean (*ko*), Dutch (*nl*), Russian (*ru*), Turkish (*tr*), Ukrainian (*uk*), Vietnamese (*vi*), Chinese (*zh*)	
Relations (4/20)	**Prompt example**
capital	What is the capital of <subject>? The answer is:
continent	Which continent is <subject> located in? The answer is:
field_of_work	What field does <subject> work in? The answer is:
religion	What is the religious belief of <subject>? The answer is:

Table 1: Overview of the languages and 4 sample relations (out of 20 relations in total) in KLAR.

Evaluation Many prior studies (Geva et al., 2023; Qi et al., 2023; Hernandez et al., 2023) assess correctness based on the model's first predicted token. However, this approach is problematic, especially in multilingual settings with complex tokenization. In many cases, even if the model predicts the correct first token, its complete output can still be incorrect.[5] To address this issue, we evaluate correctness based on the model's full answer to each factual question rather than relying solely on the first token. Following Jiang et al. (2020), we evaluate cross-lingual consistency using the overlap ratio of correct predictions for parallel facts between language pairs.[6]

Results Figure 2 shows the cross-lingual consistency results for LLaMA2 and BLOOM. While LLaMA2 generally performs better than BLOOM, both models face challenges in achieving high consistency across languages, particularly between linguistically diverse pairs. The impact of language scripts is especially evident: Non-Latin scripts, such as Arabic (*ar*), Chinese (*zh*), and Korean (*ko*),

[4] For example, the relation "shares_border_with" (prompt: "Which country does <subject> share a border with?") often involves multiple correct answers, as a country typically shares borders with several others.

[5] For example, given the Chinese prompt "文森山位于哪个大陆？答案是：" ("Which continent is Vinson Massif located in? The answer is:"), the BLOOM model outputs "南美洲" ("South America") instead of the correct answer "南极洲" ("Antarctica"). Although both responses share the same first token, the final prediction is incorrect.

[6] We do not adopt the candidate-based consistency metric proposed by Qi et al. (2023), as it relies on the next-token prediction, which, as discussed in Section 4, is unreliable in a multilingual setup.

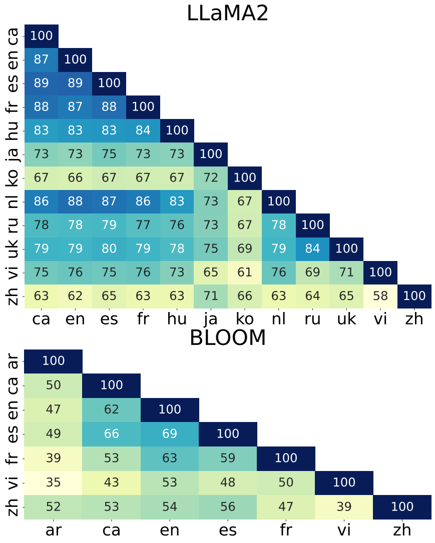


Figure 2: Cross-lingual consistency results across language pairs. The heatmaps show the overlap ratio of correct predictions between language pairs.

consistently show lower consistency scores. This underscores that cross-lingual consistency remains a key limitation for both models, emphasizing the need for more robust approaches to effectively analyze and address this issue.

5 Analyzing Multilingual Factual Recall

To understand how multilingual language models recall factual knowledge across languages, we analyze their internal mechanisms from multiple perspectives: the layer-wise evolution of prediction ranks (§5.1), latent state similarities across language pairs (§5.2), information flow within the model (§5.3), and the composition of the latent concept space (§5.4).

5.1 From the Perspective of Rankings

First, we use Logit Lens (Nostalgebraist, 2020) to project latent states at each layer to the vocabulary (unembedding) and measure the rank (the lower, the better) of the target object at each layer. Specifically, we compare the rank of the correct object in its target language (`rank_target_correct`) and its English equivalent (`rank_en_correct`). This approach allows us to trace how the model processes factual knowledge across layers and transitions between different representation modes.

Figure 3a shows distinct phases of knowledge processing in both models. In the early layers, both ranks remain high, indicating that the models have not begun extracting the target object. Around layer 15 in BLOOM and layer 12 in LLaMA2, both (`rank_target_correct`) and (`rank_en_correct`) drop significantly, marking the beginning of the object extraction phase.

This phase continues until layer 28 in LLaMA2 and layer 19 in BLOOM, where a notable divergence occurs. The English rank (`rank_en_correct`) begins to increase, while the target-language rank (`rank_target_correct`) continues to decrease. This divergence reflects a transition from language-independent object extraction to target language-specific object extraction, where the models adapt the representations to align with the target language.

These findings show that MLMs recall knowledge through an initial concept-space object extraction phase (marked by significant rank drops for both English and target language answers) before transitioning to language-specific object extraction and producing the final output.

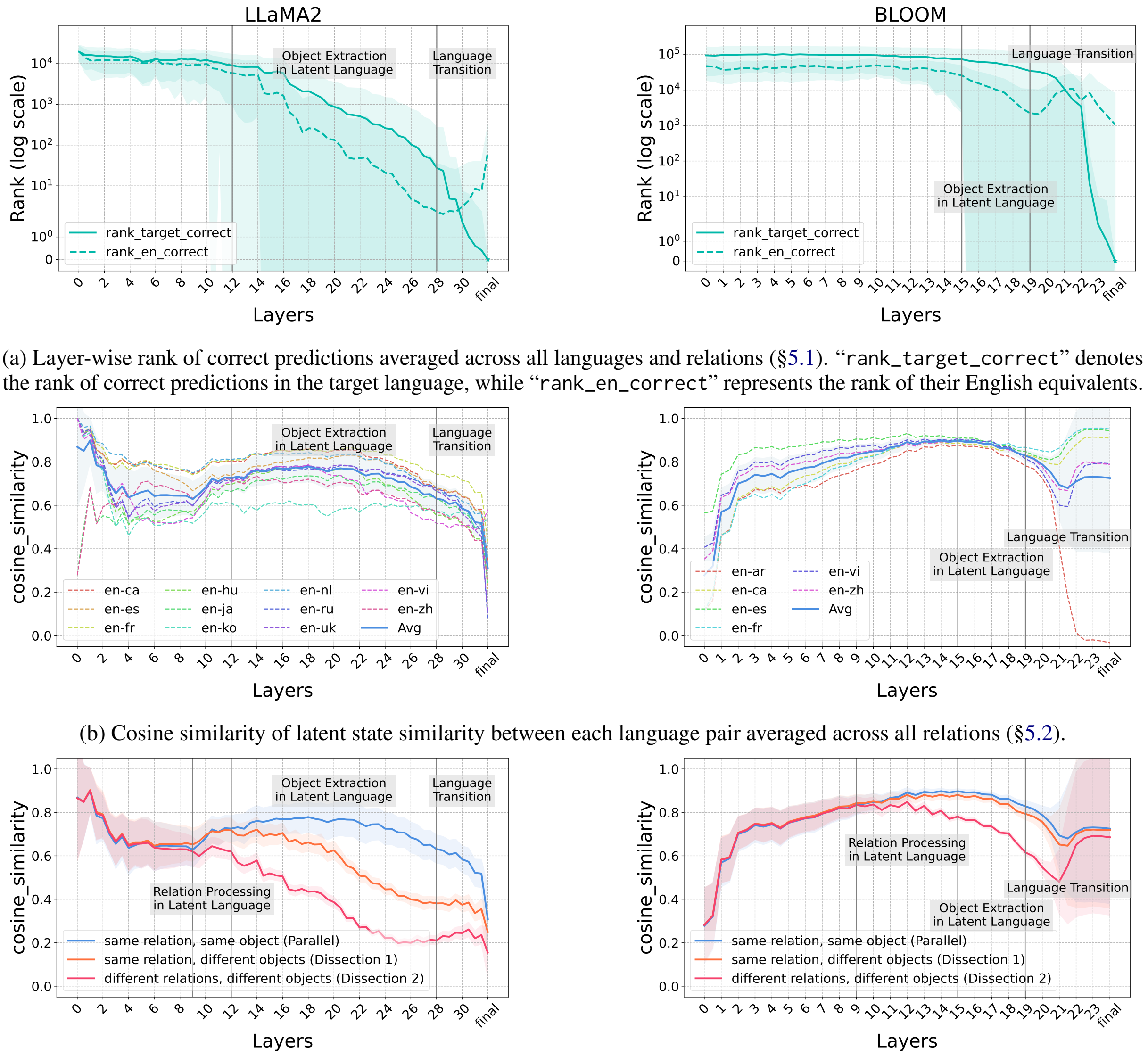
5.2 From the Perspective of Latent States

Moreover, we measure the cosine similarity of latent states between language pairs across layers.

Figure 3b shows the average cosine similarity of latent states between English and individual target languages for LLaMA2 and BLOOM.[7] As information propagates through the layers, similarity increases, peaking around 0.8 in the middle layers for both models.[8] This trend holds even for linguistically diverse pairs, such as English and Arabic, suggesting the formation of a shared concept space where factual knowledge is encoded in the model's latent language which is generic and independent of the input language. In the final layers, similarity decreases, reflecting a transition to language-specific processing. This aligns with the divergence observed in Section 5.1, where the rank changes of the target language object and its English equivalent begin to differ. These observations confirm the model's transition from concept-space object extraction to language-specific adaptations

[7] For clarity, only language pairs involving English are shown here. Complete results for all language pairs are provided in Appendix A.2.1.

[8] Our similarity analysis focuses on the final token, and all prompts end with "The answer is:". In LLaMA2, the colon ":" is typically tokenized as a standalone final token, leading to high early-layer similarity—except in en-ja and en-zh, which use language-specific colon variants. In contrast, BLOOM often fuses the colon with the preceding word (e.g., "is:", "es:", "là:", "是：") or tokenizes it separately (e.g., in ar, ca, fr), causing lower similarity due to mismatched token boundaries. This pattern is also visible in Figure 9.



(a) Layer-wise rank of correct predictions averaged across all languages and relations (§5.1). "`rank_target_correct`" denotes the rank of correct predictions in the target language, while "`rank_en_correct`" represents the rank of their English equivalents.

(b) Cosine similarity of latent state similarity between each language pair averaged across all relations (§5.2).

(c) Comparative study of latent state similarity across language pairs (§5.3). We compare the latent state similarity for parallel facts, non-parallel facts sharing the same relation, and non-parallel facts belonging to different relations, respectively.

Figure 3: Analysis of multilingual knowledge probing of LLaMA2 and BLOOM, including (3a) layer-wise evolution of correct prediction ranks, (3b) latent state similarities across languages, and (3c) the development of latent state similarities in different settings.

in the final layers.

5.3 Information Flow Dissection

While Sections 5.1 and 5.2 demonstrate the presence of a concept space in the middle layers, they do not clarify the type of information contributing to the observed high similarity between language pairs. To disentangle whether this similarity arises from relational information, object information, or both, we perform comparative experiments under three conditions: (1) **Same relation, same object** (***Parallel***, as in Section 5.2): Latent state similarity is calculated using parallel facts between each language pair (e.g., "the capital of Canada" in both English and another language); (2) **Same relation, different objects (*Dissection 1*)**: Similarity is calculated using non-parallel facts sharing the same relation (e.g., "the capital of Canada" in one language versus "the capital of Spain" in another); (3) **Different relation, different objects (*Dissection 2*)**: Similarity is calculated using non-parallel facts from different relations (e.g., "the capital of Canada" versus "the official language of Spain").

Figure 3c shows distinct processing phases. Around layer 9, the *Dissection 2* curve drops significantly in both models, while *Parallel* and *Dissection 1* curves remain close, indicating that models process relational information specific to the current fact's relation. The high similarity during this stage suggests that such relation processing happens in a language-independent concept space.

From layer 12 in LLaMA2 and layer 15 in BLOOM, the *Dissection 1* curve begins to drop, marking a transition to object-specific processing. During layers 12–28 in LLaMA2 and layers 15–19 in BLOOM, the *Parallel* curve remains high, indicating that object information is processed in the model's latent language.

At layer 28 in LLaMA2 and layer 19 in BLOOM, the *Parallel* curve drops significantly, signaling the language transition phase, where the concept-space object representations are adapted to the target language.

Together, the progression shows the models' transitions from relation processing to object extraction and to language-specific adaptation.

5.4 Concept Space Language Composition

To further explore how the concept space encodes information in MLMs, we analyze the language composition of their latent states. Using Logit Lens, we project intermediate layer representations onto the vocabulary space and identify the language of the top-10 predicted tokens at each layer using fasttext (Joulin et al., 2017).[9]

Figure 4 shows the language composition for LLaMA2 and BLOOM with Chinese (zh) as the input language, averaged across factual queries spanning all relations. Results for other input languages are provided in Appendix A.2.3.

In LLaMA2, English dominates the middle-to-upper layers, suggesting that factual knowledge is processed in an English-centric concept space. This is consistent with prior findings that "LLaMA2 models think in English" (Wendler et al., 2024). In contrast, BLOOM shows a more diverse composition in the middle-to-upper layers, comprising primarily Latin-based languages like English, French, Spanish, German, etc.

Within each model, the middle-to-upper layers exhibits similar language compositions across different input languages (see Appendix Figures 10 and 11). This suggests that multilingual models encode factual knowledge in a shared concept space largely independent of the input language. Notably, this space is not necessarily aligned with any single language, indicating that **multilingual LLMs "think" in their own concept space** rather than in the surface form of a particular language.

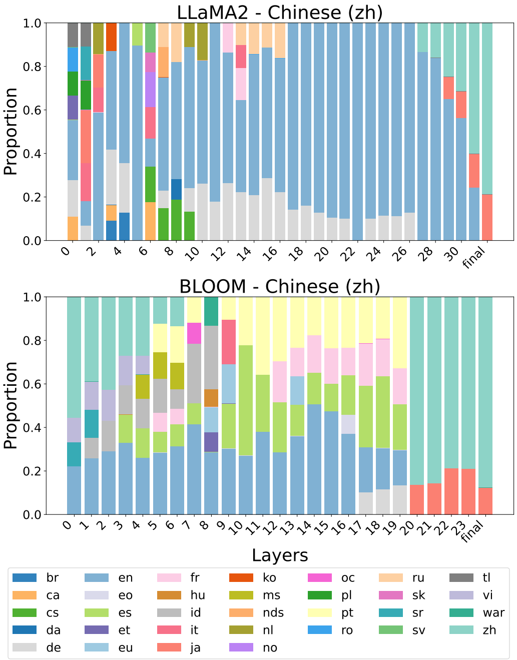


Figure 4: Language composition of latent representations with Chinese as the input language. In LLaMA2, English dominates the middle-to-upper layers, whereas BLOOM has a more diverse language composition.

5.5 Summary

Our analysis reveals a three-stage knowledge recall process in MLMs (as illustrated in Figure 1): first relation processing, then object extraction in the model's latent language, and finally the transition to language-specific processing to adapt the object to the target language. These findings provide a comprehensive view on the mechanisms of multilingual factual recall.

6 Examining the Cause of Cross-Lingual Inconsistency

Next, we analyze incorrect predictions across languages to investigate the causes of cross-lingual inconsistencies in MLMs.

Figure 5 shows the rank evolution for incorrect predictions in LLaMA2 and BLOOM. While the rank of the correct answer decreases significantly in the middle layers (both in the target language and in English) — consistent with the behavior observed in correct predictions (Figure 3a) — the rank of the incorrect answer surpasses that of the correct answer during language transition in the final layers. This suggests that factual knowledge is processed in the concept space in the middle layers

[9] We filter out tokens with confidence scores below 0.5.

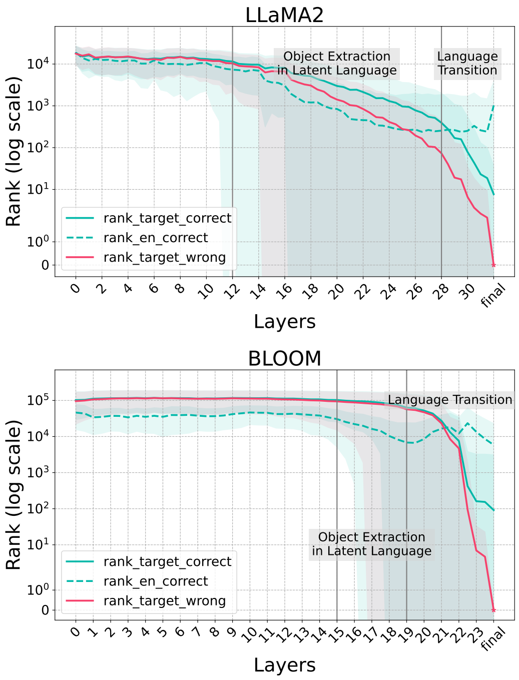
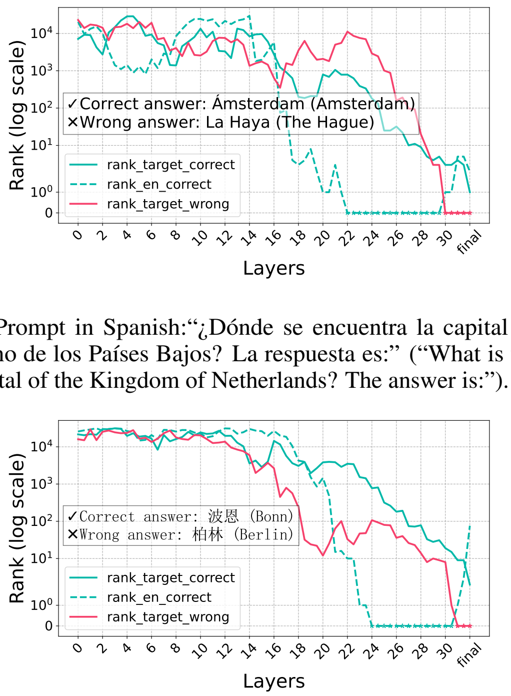


Figure 5: Layer-wise rank of incorrect predictions averaged across all languages and relations. The `rank_target_wrong` curve represents the rank of the model's final incorrect prediction across layers, while `rank_target_correct` and `rank_en_correct` denote the ranks of the correct answer in the target language and the English equivalent, respectively.

as in correct predictions, but errors arise during the transition to language-specific processing.

To further investigate this phenomenon, we examine individual examples of LLaMA2.[10] Figure 6 presents cases in Spanish and Chinese, with additional examples provided in Appendix A.2.3. A consistent pattern emerges: in the middle-to-upper layers, the correct answer in English often ranks lowest (`rank_en_correct`=0), indicating accurate recall during the concept space processing stage. However, in the final layers, the rank of the incorrect target-language answer decreases, surpassing the correct answer during language transition.

This observation underscores the critical role of language transition in cross-lingual inconsistencies. Although MLMs encode correct factual knowledge in the middle-layer concept space, the transition to language-specific processing introduces errors, causing incorrect predictions. Addressing this issue is crucial for improving cross-lingual consistency and robustness of MLMs.



(a) Prompt in Spanish:"¿Dónde se encuentra la capital de Reino de los Países Bajos? La respuesta es:" ("What is the capital of the Kingdom of Netherlands? The answer is:").

(b) Prompt in Chinese: "西德的首都在哪里？答案是：" ("What was the capital of West Germany? The answer is:").

Figure 6: Rank evolution for prompts in Spanish (6a) and Chinese (6b). `rank_target_wrong` represents the rank of the model's final incorrect prediction across layers, while `rank_target_correct` and `rank_en_correct` denote the ranks of the correct answer in the target language and the English equivalent, respectively. The plots show the impact of errors during language transition, where the rank of the incorrect answer surpasses the correct answer in the final layers.

7 Linear Shortcut for Improving Cross-Lingual Consistency

In this section, we propose a linear shortcut method to address language transition errors. Our approach bypasses final-layer computations, directly adapting concept-space representations to the target language, enhancing both prediction accuracy and cross-lingual consistency of MLMs.

7.1 Shortcut with Linear Approximation

The proposed method consists two-step (illustrated in Figure 7): (a) **Deriving the linear shortcut**: Inspired by Hernandez et al. (2023), we hypothesize that the mapping from the model's latent state at layer n to the final layer N, i.e., $h_n \rightarrow h_N$ can be well-approximated by a linear function $f(h_n) = Wh_n + b \approx h_N$. Using m correctly predicted samples per relation, we estimate W and b via first-order approximation, modeling how

[10] LLaMA2's English-biased latent space provides clearer insights into the switch from English to the target language, while BLOOM's latent space is less interpretable, as shown in Figure 4.

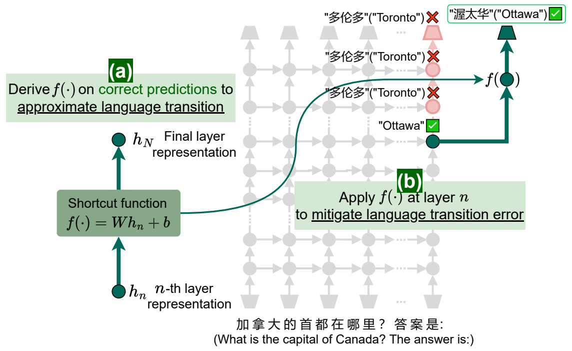

"多伦多"("Toronto") ✗
"渥太华"("Ottawa") ✓
"多伦多"("Toronto") ✗
(a)
Derive $f(\cdot)$ on correct predictions to approximate language transition
$f(\cdot)$
"多伦多"("Toronto") ✗
"Ottawa" ✓
h_N Final layer representation
Shortcut function $f(\cdot) = Wh_n + b$
(b)
Apply $f(\cdot)$ at layer n to mitigate language transition error
h_n n-th layer representation
加拿大的首都在哪里？ 答案是:
(What is the capital of Canada? The answer is:)

Figure 7: Illustration of the proposed shortcut method for mitigating cross-lingual inconsistency. **(a)** The shortcut function is learned on correct predictions to approximate language transition; **(b)** The learned function is then applied to bypass the error-prone final layers. In the example, the shortcut successfully recovers the correct answer, "渥太华" ("Ottawa"), in Chinese.

concept-space representations are adapted to the target language.[11] We optimize one linear shortcut per language, shared across all relations, which aims to capture generalizable patterns in the representation-to-output transition for each language. Further details on the derivation and hyperparameters are provided in Appendix A.3. (b) **Applying the linear shortcut**: At inference time, the learned shortcut $f(\cdot)$ is applied to bypass the original final-layer computations, mitigating errors introduced during language transition.

7.2 Results and Discussion

We evaluate the prediction accuracy and cross-lingual consistency of LLaMA2 and BLOOM, without and with applying the shortcut method, on all KLAR samples.

Baselines. We compare our shortcut method to three translation-based baselines: (1) *translation-en*: We translate all input queries from each language to English using Google Translate, obtain model predictions in English, and then translate them back to the target language. (2) *translation-early-exit*: We use Logit Lens to extract top-predicted tokens from the same layers as the shortcut method, translate them into the target language and evaluate their accuracy. (3) *fine-tuning*: We fine-tune the models using $m = 25$ parallel samples per relation per language and evaluate on the full KLAR dataset, consistent with the settings used for our shortcut method. For efficiency, we applied LoRA-based fine-tuning to LLaMA2 (learning rate $lr = 1\text{e}{-4}$), and full model fine-tuning to BLOOM (learning rate $lr = 5\text{e}{-8}$) due to poor LoRA performance. Both models are trained with a batch size of 4 for 20 epochs.

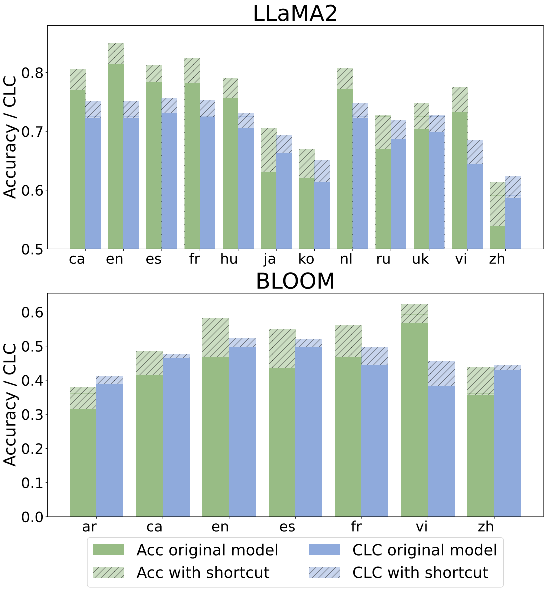


Figure 8: Accuracy (ACC) and cross-lingual consistency (CLC) per language for LLaMA2 and BLOOM, with and without the shortcut method.

Results. Figure 8 shows the effectiveness of the shortcut mapping: It improves prediction accuracy and cross-lingual consistency across models and languages. This demonstrates its ability to adapt concept-space knowledge to target languages for more reliable predictions.

	original	shortcut	trans-en	trans-exit	ft
LLaMA2	71.47	**76.08**	53.88	13.93	72.26
BLOOM	43.24	**51.67**	28.03	15.68	31.73

Table 2: Average accuracy across languages..

As shown in Table 2, both translation-based baseline methods perform poorly (see Table 8 and 9 in Appendix for more details), indicating that existing translators are insufficient for cross-lingual factual prediction.

Fine-tuning also yields unsatisfactory results. For LLaMA2, it slightly outperforms the original model but underperforms our shortcut method in accuracy. For BLOOM, fine-tuning underperforms the original model, with improvements seen only in English. We hypothesize that fine-tuning on a small subset of factual knowledge does not gener-

[11] Layer n and training size m are treated as hyperparameters: $n = 30$ for LLaMA2, $n = 21$ for BLOOM, and $m = 25$ for both models. Details are provided in Appendix A.3.2.

alize well to unseen facts and may even degrade performance due to overfitting.

In contrast, our shortcut method directly adapts latent representations from earlier layers, preserving richer contextual information and thus improving prediction accuracy. Moreover, it is lightweight and efficient, relying only on linear operations, making it easily adaptable to existing MLMs.

8 Conclusion

This study investigates cross-lingual factual inconsistency in multilingual language models, revealing a three-stage knowledge recall process: language-independent relation processing, object extraction, and a final transition to language-specific adaptation. Errors in this transition often lead to incorrect predictions despite accurate object extraction. To address this, we propose a shortcut method that bypasses final-layer computations, improving prediction accuracy and cross-lingual consistency. Our findings enhance understanding of multilingual knowledge processing and introduce an efficient, interpretable solution for mitigating language transition errors.

Future work could expand the investigation to more languages and additional language models to assess broader applicability. Additionally, developing non-linear shortcut methods could better mitigate language transition errors, offering more robust solutions for cross-lingual consistency.

Limitations

First, our cross-lingual consistency analysis assumes English as the pivot language, reflecting the English-centric nature of most multilingual models. While this aligns with prior studies (Wendler et al., 2024; Dumas et al., 2024; Fierro et al., 2024), it may limit applicability to language pairs that do not involve English.

Second, although the KLAR dataset covers 17 languages, it does not fully capture the diversity of world languages. Expanding the analysis to more languages and exploring models with different architectures and sizes could provide deeper insights into cross-lingual inconsistencies.

Additionally, our shortcut method relies on linear approximation for simplicity. Investigating non-linear approaches could better capture complex transformations during language switching and further enhance performance.

Finally, our analysis provides insights relevant to downstream tasks, such as multilingual knowledge localization (Chen et al., 2024; Kojima et al., 2024; Tang et al., 2024) and cross-lingual knowledge editing (Xu et al., 2023; Nie et al., 2024). However, these applications fall beyond the scope of this study and are left for future work.

Acknowledgments

This work was partially supported by Deutsche Forschungsgemeinschaft (project SCHU 2246/14-1). We would like to thank Sebastian Gerstner, Ahmad Dawar Hakimi, Anna Hätty, Lea Hirlimann, Amir Hossein Kargaran, Valentin Knappich, Felicia Körner, Mohsen Mesgar, Ali Modarressi, Philipp Mondorf, Timo Pierre Schrader, Leonor Veloso, Wei Zhou, Yuqicheng Zhu for fruitful discussions.

Ethical considerations

This study investigates cross-lingual factual inconsistencies in multilingual language models. While our focus is diagnostic, incorrect model predictions may propagate misinformation or reflect underlying biases present in the models.

References

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.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Clément Dumas, Veniamin Veselovsky, Giovanni Monea, Robert West, and Chris Wendler. 2024. How do llamas process multilingual text? a latent exploration through activation patching. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer ElShowk, Nicholas Joseph, Nova DasSarma, Ben Mann, Danny Hernandez, Amanda Askell, Kamal Ndousse, Andy Jones, Dawn Drain, Anna Chen, Yuntao Bai, Deep Ganguli, Liane Lovitt, Zac Hatfield-Dodds, Jackson Kernion, Tom Conerly, Shauna Kravec, Stanislav Fort, Saurav Kadavath, Josh Jacobson, Eli Tran-Johnson, Jared Kaplan, Jack Clark, Tom Brown, Sam McCandlish, Dario Amodei, and Christopher Olah. 2022. Softmax linear units. *Transformer Circuits Thread*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2024. How do multilingual models remember? investigating multilingual factual recall mechanisms. *arXiv preprint arXiv:2410.14387*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *Trans. Mach. Learn. Res.*, 2023.

William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. A material lens on coloniality in nlp. *arXiv preprint arXiv:2311.08391*.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2023. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*.

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, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *ArXiv e-prints*, pages arXiv–1607.

Yihong Liu, Runsheng Chen, Lea Hirlimann, Ahmad Dawar Hakimi, Mingyang Wang, Amir Hossein Kargaran, Sascha Rothe, François Yvon, and Hinrich Schütze. 2025. On relation-specific neurons in large language models. *arXiv preprint arXiv:2502.17355*.

Ang Lv, Yuhan Chen, Kaiyi Zhang, Yulong Wang, Lifeng Liu, Ji-Rong Wen, Jian Xie, and Rui Yan. 2024. Interpreting key mechanisms of factual recall in transformer-based language models. *Preprint*, arXiv:2403.19521.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Ercong Nie, Bo Shao, Zifeng Ding, Mingyang Wang, Helmut Schmid, and Hinrich Schütze. 2024. Bmike-53: Investigating cross-lingual knowledge editing with in-context learning. *arXiv preprint arXiv:2406.17764*.

Nostalgebraist. 2020. interpreting gpt: the logit lens.

Chris Olah. 2022. Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuits Thread*, 1(3).

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. 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.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *Preprint*, arXiv:2407.02646.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yifei Wang, Yuheng Chen, Wanting Wen, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. 2024. Unveiling factual recall behaviors of large language models through knowledge neurons. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7402, Miami, Florida, USA. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Wikidata. 2025. Properties - Wikidata. Accessed: 2025-01-09.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. 2023. Language anisotropic cross-lingual model editing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5554–5569, Toronto, Canada. Association for Computational Linguistics.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

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

A Appendix

A.1 KLAR Dataset Details

As discussed in Section 3, BMLAMA17 (Qi et al., 2023) is incompatible with multilingual knowledge probing in auto-regressive models with many objects placed in the middle of sentences, and many relations types with multiple correct answers. To address these limitations, we construct KLAR for reliable multilingual knowledge probing evaluation.

BMLAMA17 does not explicitly specify relation types; however, many factual questions share the same templates. We first group sentences with identical templates and use `gpt-3.5-turbo` to identify the relation for each template and map them to Wikidata property IDs (Wikidata, 2025). We discard the samples which cannot be mapped to any Wikidata property. This process yields a total of 42 relation types.

For each relation, we generate English prompt templates in the format of "<*Question*> The answer is:" as introduced in Section 3, using `gpt-3.5-turbo`. We created five templates per relation and manually verify their clarity. The templates are then translated into 16 additional languages using `gpt-3.5-turbo`. Their quality is manually reviewed for Chinese, Spanish, and Japanese. Back-translation is used to verify clarity and consistency in the remaining languages.

Finally, we remove relation types with multiple correct answers and those with fewer than 30 samples. The resulting KLAR dataset comprises parallel factual knowledge spanning 17 languages and 20 relation types. For the analysis on LLaMA2 and BLOOM models, we use the intersection of languages supported by these models and included in KLAR, covering 12 languages for LLaMA2 and 7 for BLOOM, see Table 4 for the respective language list. Listing 1 illustrates the example of the KLAR dataset structure for the relation *capital* in English.

```
{
    "relation_name": "capital",
    "relation_id": "P36",
    "prompt_templates": [
        "Where is <subject>'s capital
            located? The answer is:",
        "What is the capital of <subject
            >? The answer is:",
        "Which city serves as the
            capital of <subject>? The
            answer is:",
        "Name the capital city of <
            subject>. The answer is:",
        "Where does <subject> have its
            capital? The answer is:"
    ],
    "samples": [
        {
            "subject": "Azerbaijan",
            "object": "Baku",
            "index": 6152
        },
        {
            "subject": "Germany",
            "object": "Berlin",
            "index": 6165
        },
    ]
}
```

Listing 1: Example of KLAR for relation *capital* in English.

A.2 Additional Experimental Results

A.2.1 Latent State Similarity

Here, we present the complete results for latent state similarity across all language pairs in Figure 9.

The plots follow the same trend as in Figure 3b, where similarity across language pairs increases from early to middle layers in both models, indicating that MLMs encode information in a concept space independent of the input language. In the final layers, similarity declines as representations transition to a language-specific form. This pattern holds even for linguistically diverse pairs, highlighting that MLMs initially process factual knowledge in a shared latent space before adapting it to the target language.

A.2.2 Latent Space Language Composition

We examine the language composition of the latent states in LLaMA2 and BLOOM to understand how these MLMs encode information in the concept space. As described in Section 5.4, we apply Logit Lens to project latent states to the vocabulary, and use fasttext to identify the language of the top-10 predicted tokens at each layer.

Figure 10 presents results for languages shared between LLaMA2 and BLOOM, while Figure 11 shows results for languages unique to each model.

LLaMA2's middle-to-upper layers are dominated by English, aligning with prior findings that "LLaMA2 models think in English" (Wendler et al., 2024). In contrast, BLOOM displays a more diverse linguistic composition in these layers.

Across different input languages, both models exhibit similar language distributions in the middle-to-upper layers, indicating that MLMs en-

Relation	**# Facts**	**Prompt Example**
applies_to_jurisdiction	79	Which country has <subject> as a legal term? The answer is:
capital	336	What is the capital of <subject>? The answer is:
capital_of	212	Where is <subject> the capital of? The answer is:
continent	212	Which continent is <subject> located in? The answer is:
country_of_citizenship	60	Which country is <subject> a citizen of? The answer is:
developer	76	Which company is the developer of <subject>? The answer is:
field_of_work	167	What field does <subject> work in? The answer is:
headquarters_location	51	In which city is <subject>'s headquarter located? The answer is:
instrument	46	Which musical instrument is played by <subject>? The answer is:
language_of_work_or_name	108	What is the original language of <subject>? The answer is:
languages_spoken	104	What language did <subject> use to communicate? The answer is:
location_of_formation	66	Where did the formation of <subject> take place? The answer is:
manufacturer	35	Which company manufactures <subject>? The answer is:
native_language	130	What is the native language of <subject>? The answer is:
occupation	46	What is <subject>'s profession? The answer is:
official_language	602	What is the official language of <subject>? The answer is:
owned_by	50	Who is the current owner of <subject>? The answer is:
place_of_birth	35	In which city was <subject> born? The answer is:
place_of_death	79	In which city did <subject> pass away? The answer is:
religion	125	What is the religious belief of <subject>? The answer is:

Table 3: Relations in the KLAR dataset with fact counts and prompt examples used for knowledge probing.

KLAR languages (17)	Arabic (ar), Catalan(ca), Greek (el), English (en), Spanish (es), Persian (fa), French (fr), Hebrew (he), Hungarian (hu), Japanese (ja), Korean (ko), Dutch (nl), Russian (ru), Turkish (tr), Ukrainian (uk), Vietnamese (vi), Chinese (zh)
LLaMA2 overlap (12)	Catalan(ca), English (en), Spanish (es), French (fr), Hungarian (hu), Japanese (ja), Korean (ko), Dutch (nl), Russian (ru), Ukrainian (uk), Vietnamese (vi), Chinese (zh)
BLOOM overlap (7)	Arabic (ar), Catalan(ca), English (en), Spanish (es), French (fr), Vietnamese (vi), Chinese (zh)

Table 4: KLAR dataset languages and their overlap with LLaMA2 and BLOOM.

code knowledge in a concept space largely independent of the input language.

A.2.3 Rank Plots of Wrong Predictions

Figure 12 presents additional examples, one per language, where the correct English answer ranks highest in the middle-to-upper layers but is later surpassed by an incorrect target-language answer during the language transition phase.

A.3 Shortcut Experimental Details

A.3.1 Method

The idea of using linear approximation as a shortcut is inspired by Hernandez et al. (2023), who derive a linear transformation to approximate the mapping from subject to object representations in factual knowledge, showing that relational decoding in transformer models can be effectively modeled with linear functions.

Building on this idea, we apply linear approximation to address cross-lingual inconsistency by bypassing the language transition process in MLMs. We hypothesize that the mapping from the model's latent state at layer n to that at the final layer N, i.e., $h_n \to h_N$ can be well-approximated by a linear function $f(h_n) = Wh_n + b \approx h_N$. Following Hernandez et al. (2023), we use first-order approximation to estimate W_r and b_r as the mean Jacobian and bias across m correctly predicted factual samples $\{h_{n_i}, h_{N_i}\}_{i=1,\dots,m}$. That is, we define:

$$\begin{aligned} W_r &= \mathbb{E}_{h_{n_i},h_{N_i}}\left[\left.\frac{\partial F}{\partial h_n}\right|_{(h_{n_i},h_{N_i})} \right], \\ b_r &= \mathbb{E}_{h_{n_i},h_{N_i}}\left[h_N - \left.\frac{\partial F}{\partial h_n}\right|_{(h_{n_i},h_{N_i})} h_n \right] \end{aligned} \quad (1)$$

As noted in Hernandez et al. (2023), the first-order derivative W_r tends to underestimate the magnitude of changes from h_n to h_N in practice. They attribute this to the use of layer normalization (Lei Ba et al., 2016) in transformers: which does not transmit changes in scale of inputs to changes in scale of output. Specifically, the input h_n at layer n is normalized before being propagated to subsequent layers. To address this underestimation, a

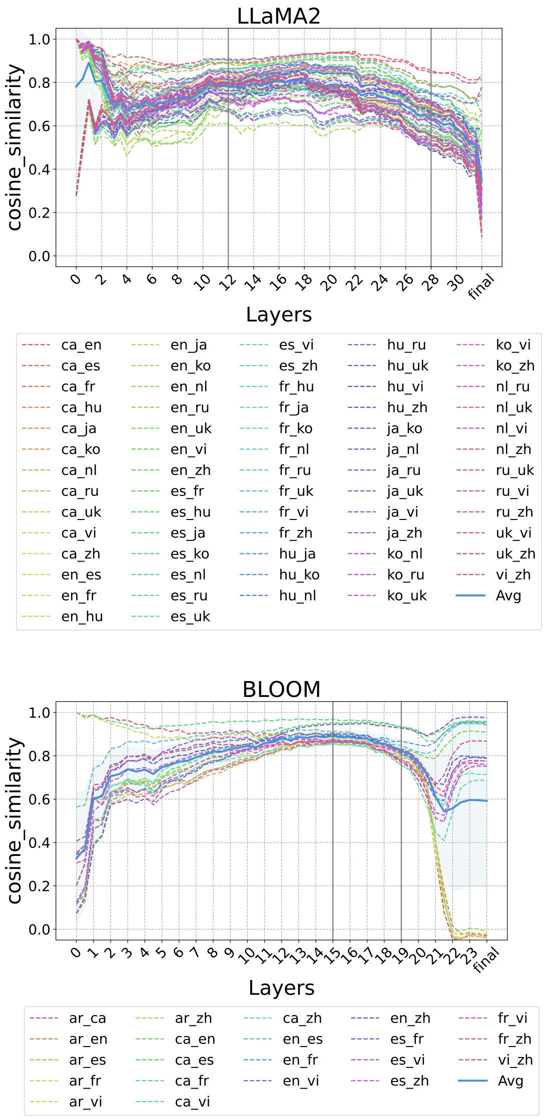


Figure 9: Cosine similarity of latent states between all language pairs averaged across all relation.

scalar constant β is introduced as a hyperparameter and multiplied by W_r as a corrective factor:

$$f(h_n) = \beta W_r h_n + b_r = W h_n + b \qquad (2)$$

A.3.2 Hyperparameters

Several hyperparameters are introduced when determining the linear shortcut $f(\cdot)$: the layer n from which the latent state is extracted for linear approximation, the scalar constant β used to adjust the slope of W_r to account for the underestimation in the first-order approximation of $h_n \to h_N$, and the number of correct samples used to compute $f(\cdot)$. We perform a grid search to select these hyperparameters per language, aiming to maximize prediction accuracy. For the layer n, we search within the range of $[20, 32]$ for LLaMA2 and $[12, 24]$ for BLOOM. The scalar constant β is searched over the range $[0, 5.0]$ in increments of 0.25, following Hernandez et al. (2023). The number of samples m is selected from $[10, 25, 40, 50]$. The hyperparameter search is conducted for each language individually. We find that the optimal β value varies across languages, while the other two hyperparameters — the extraction layer n and the number of samples m — remain consistent across languages. The selected hyperparameters for both models are summarized in Table 5 and 6, respectively.

LLaMA2	n	β	m
ca	30	4.75	25
en		1.50	
es		3.00	
fr		4.25	
hu		2.50	
ja		2.25	
ko		4.50	
nl		3.50	
ru		4.25	
uk		2.25	
vi		1.00	
zh		1.50	

Table 5: Hyperparameters per language for LLaMA2.

BLOOM	n	β	m
ar	21	1.25	25
ca		1.00	
en		1.25	
es		1.00	
fr		0.75	
vi		1.25	
zh		1.50	

Table 6: Hyperparameters per language for BLOOM.

A.3.3 Shortcut Translation Baselines.

As mentioned in Section 7.2, we compare our shortcut method with two translation-based baselines: (1) translation-en (trans-en): We translate all input queries from each language to English using Google Translate, obtain model predictions in English, and then translate them back to the target language to measure accuracy. (2) translation-early-exit (trans-exit): We use Logit Lens to project the latent states at the same extraction layers as in the shortcut method, i.e., layer 30 for LLaMA2 and layer 20 for BLOOM, and extract the top-predicted tokens. These tokens are then translated into the target language using Google Translate, and their accuracy is calculated against the correct object.

As shown in Table 8 and 9, both translation-based methods perform poorly. The low accu-

Relations	Accuracy (acc)			Cross-lingual Consistency (clc)		
	Original	Shortcut	Diff	Original	Shortcut	Diff
applies_to_jurisdiction	92.92	**96.28**	3.36	87.60	**92.84**	5.24
capital	83.06	**88.54**	5.48	80.87	**86.10**	5.23
capital_of	66.16	**70.26**	4.10	71.66	**74.98**	3.32
continent	85.50	**90.37**	4.87	81.34	**86.94**	5.60
country_of_citizenship	71.53	**76.38**	4.85	69.41	**72.91**	3.50
developer	90.90	**94.05**	3.15	84.02	**87.12**	3.10
field_of_work	47.50	**53.39**	5.89	46.34	**53.99**	7.65
headquarters_location	68.79	**74.40**	5.61	62.94	**67.28**	4.34
instrument	60.87	**65.22**	4.35	66.48	**72.76**	6.28
language_of_work_or_name	84.49	**88.09**	3.60	86.14	**90.68**	4.54
languages_spoken	75.48	**83.65**	8.17	71.89	**81.64**	9.75
location_of_formation	49.24	**56.56**	7.32	44.81	**49.58**	4.77
manufacturer	94.28	**96.83**	2.55	91.77	**93.97**	2.18
native_language	91.09	**93.24**	2.15	87.75	**92.50**	4.75
occupation	36.23	**42.22**	5.99	48.18	**56.50**	8.32
official_language	67.64	**71.22**	3.58	74.14	**77.32**	3.18
owned_by	60.50	**64.59**	4.09	57.27	**62.57**	5.30
place_of_birth	53.33	**57.95**	4.62	47.89	**54.26**	6.37
place_of_death	67.62	**72.89**	5.27	66.15	**69.95**	3.80
religion	82.23	**85.33**	3.10	82.15	**86.41**	4.26
AVG	71.47	**76.08**	4.60	70.44	**75.52**	5.07

Table 7: Prediction accuracy (acc) and cross-lingual consistency (clc) of LLaMA2 before and after applying the shortcut method across different relations.

racy of *translation-en* suggests that existing translators struggle with entity translation, especially for languages that are highly dissimilar to English. The poor performance of *translation-early-exit* stems from the inherent unreliability of token-level translations. Overall, these results indicate that translation-based approaches are not a viable solution for cross-lingual factual prediction. In contrast, by directly adapting latent representations from earlier layers, the shortcut method operates at the representation level, capturing richer contextual information. This enables significantly higher prediction accuracy and offers a promising solution for mitigating cross-lingual factual inconsistency.

LLaMA2	original	shortcut	trans-en	trans-exit	ft
ca	76.96	**80.54**	44.95	24.52	70.25
en	81.41	**85.06**	81.41	43.05	80.43
es	78.44	**81.16**	47.77	28.42	75.44
fr	78.14	**82.46**	53.27	24.85	76.58
hu	75.69	**79.04**	64.60	6.91	70.48
ja	63.05	**70.45**	59.59	0.13	63.51
ko	62.14	**66.98**	49.30	0.28	52.38
nl	77.22	**80.77**	62.07	15.24	75.1
ru	67.02	**72.71**	47.58	2.72	67.56
uk	70.46	**74.78**	46.59	5.62	67.18
vi	73.26	**77.56**	39.07	12.70	70.82
zh	53.88	**61.40**	60.38	1.67	52.66

Table 8: Comparison of the prediction accuracy (%) for LLaMA2 across different languages using the original model, the proposed shortcut method, and the translation-based baselines.

BLOOM	original	shortcut	trans-en	trans-exit	ft
ar	31.58	**37.93**	21.87	0.97	23.99
ca	41.50	**48.40**	22.58	15.88	28.93
en	46.81	**58.24**	46.81	26.85	49.97
es	43.56	**54.84**	25.53	11.26	36.17
fr	46.88	**56.03**	26.15	17.97	34.14
vi	56.82	**62.38**	21.98	25.85	29.54
zh	35.54	**43.89**	31.26	10.96	25.44

Table 9: Comparison of the prediction accuracy (%) for BLOOM across different languages using the original model, the proposed shortcut method, and the translation-based baselines.

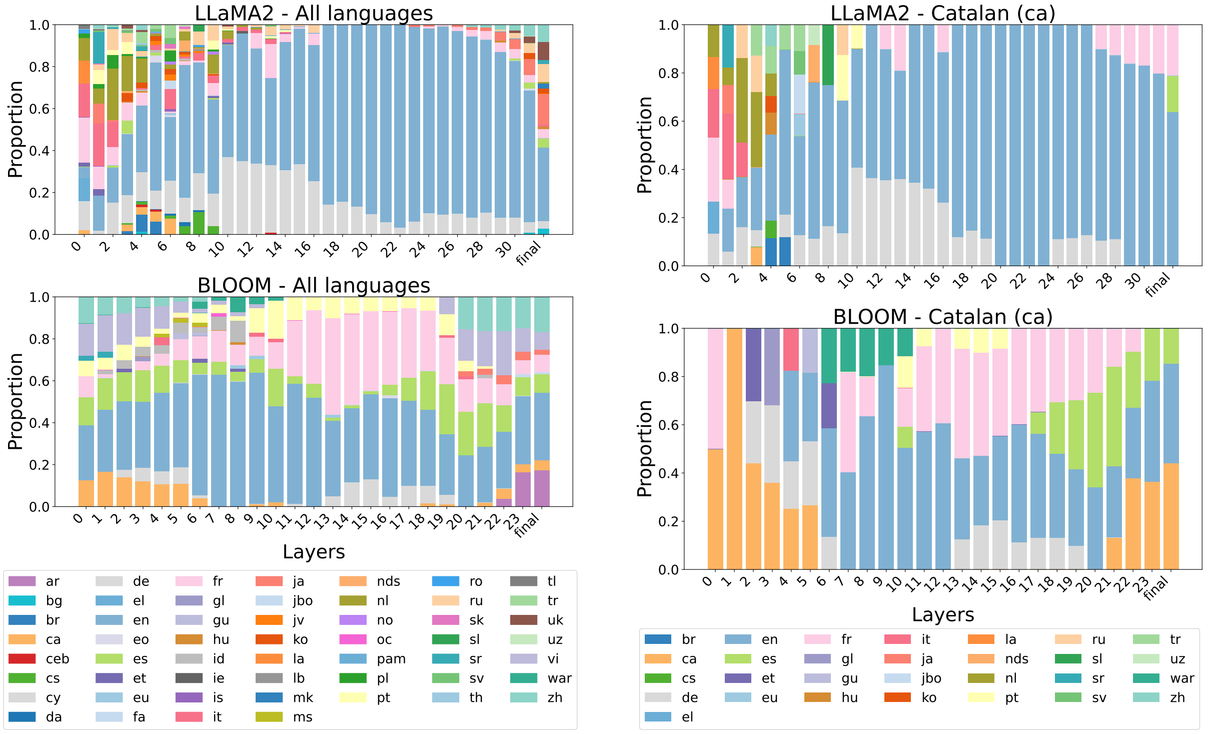
A.3.4 Per-relation Shortcut Performance.

In Tables 7 and 10, we provide a detailed per-relation breakdown of performance for both the original LLaMA2 and BLOOM models and their shortcut-enhanced counterparts, covering prediction accuracy (acc) and cross-lingual consistency (clc).

The results demonstrate that the improvements are not limited to a specific relation, but are consistently observed across a wide range of relation types, underscoring the robustness and generalizability of the proposed shortcut method.

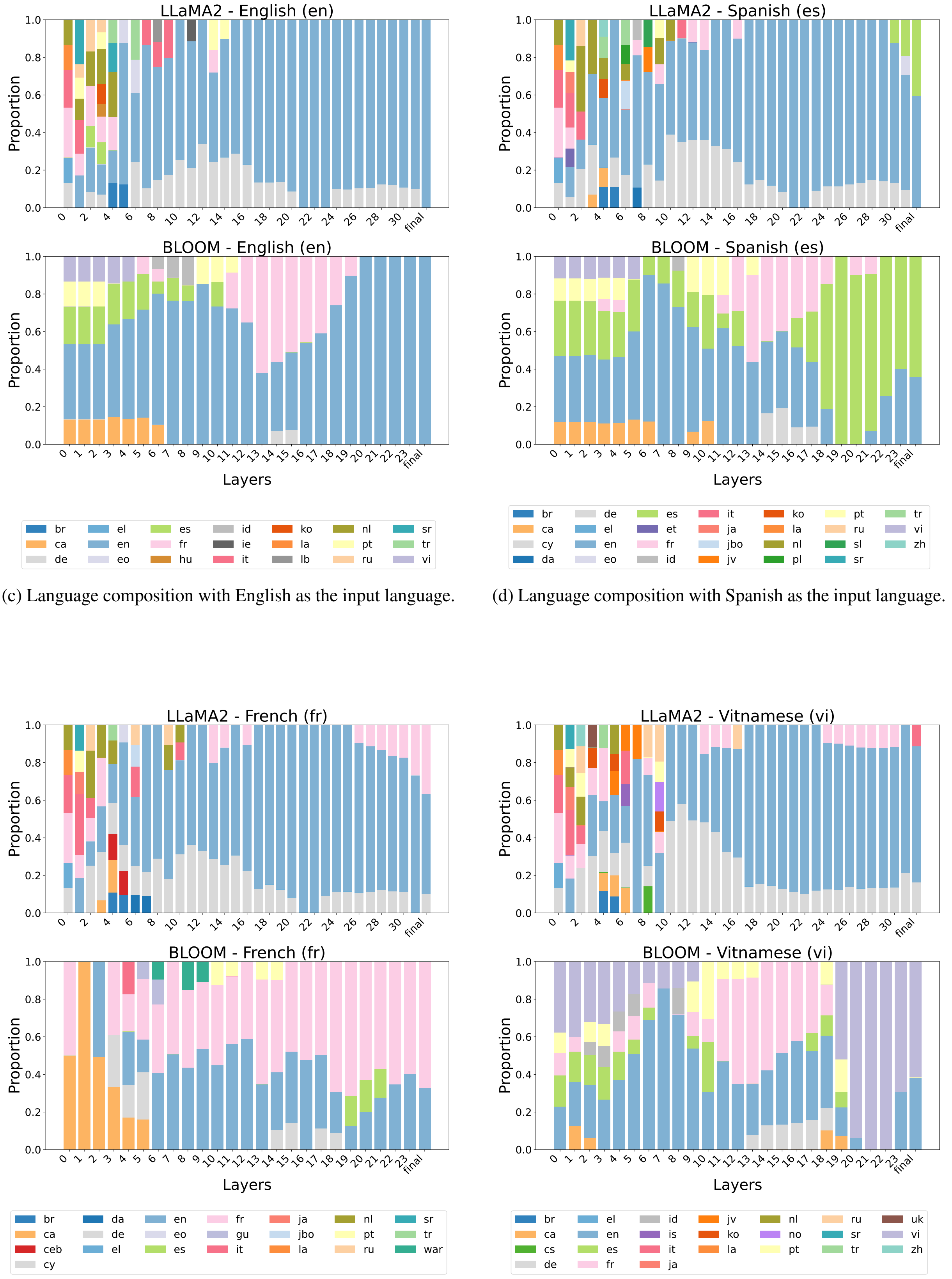
Relations	Accuracy (acc)			Cross-lingual Consistency (clc)		
	Original	**Shortcut**	**Diff**	**Original**	**Shortcut**	**Diff**
applies_to_jurisdiction	88.57	**93.56**	4.99	87.17	**90.03**	2.86
capital	42.86	**48.60**	5.74	47.44	**52.67**	5.23
capital_of	36.28	**42.35**	6.07	40.77	**45.96**	5.19
continent	18.73	**55.73**	37.00	19.68	**39.58**	19.90
country_of_citizenship	32.38	**43.81**	11.43	36.33	**44.86**	8.53
developer	74.06	**78.26**	4.20	67.52	**73.45**	5.93
field_of_work	12.92	**22.50**	9.58	24.05	**34.08**	9.83
headquarters_location	31.09	**36.41**	5.32	49.89	**53.96**	4.07
instrument	46.27	**52.56**	6.29	97.84	**98.87**	1.03
language_of_work_or_name	62.30	**69.04**	6.74	75.24	**79.65**	4.41
languages_spoken	47.66	**53.37**	5.71	59.66	**65.44**	5.78
location_of_formation	17.32	**22.29**	4.97	27.55	**32.41**	4.86
manufacturer	88.16	**92.61**	4.45	83.07	**91.02**	7.95
native_language	54.40	**70.99**	16.59	38.52	**48.71**	10.19
occupation	20.19	**26.25**	6.06	37.56	**42.86**	5.30
official_language	47.01	**54.14**	7.13	44.03	**49.18**	5.15
owned_by	40.00	**45.71**	5.71	52.11	**57.50**	5.39
place_of_birth	19.59	**26.25**	6.66	44.94	**49.03**	4.09
place_of_death	23.91	**40.12**	16.21	53.61	**60.36**	6.75
religion	52.11	**58.12**	6.01	82.93	**84.62**	1.69
AVG	43.24	**51.67**	8.43	54.16	**60.32**	6.16

Table 10: Prediction accuracy (acc) and cross-lingual consistency (clc) of BLOOM before and after applying the shortcut method across different relations.



(a) Language composition aggregated across all languages

(b) Language composition with Catalan as the input language.



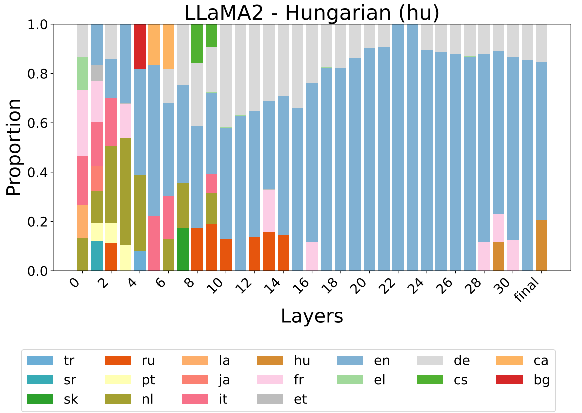
(c) Language composition with English as the input language.

(d) Language composition with Spanish as the input language.

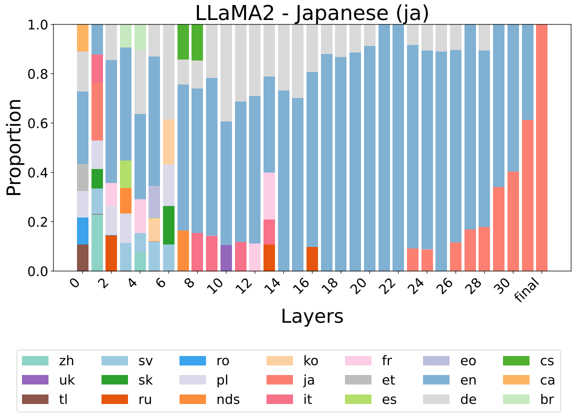
(e) Language composition with French as the input language.

(f) Language composition with Vietnamese as the input language.

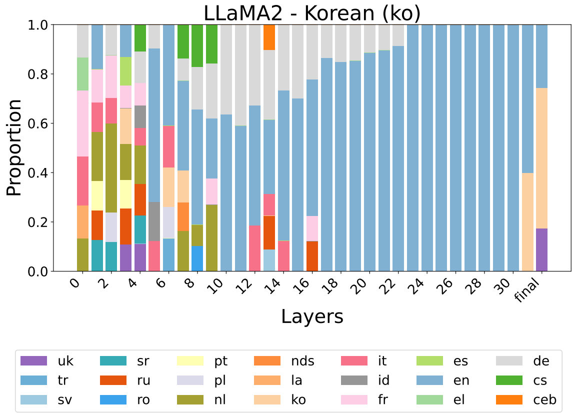
Figure 10: Language composition for languages shared between LLaMA2 and BLOOM.



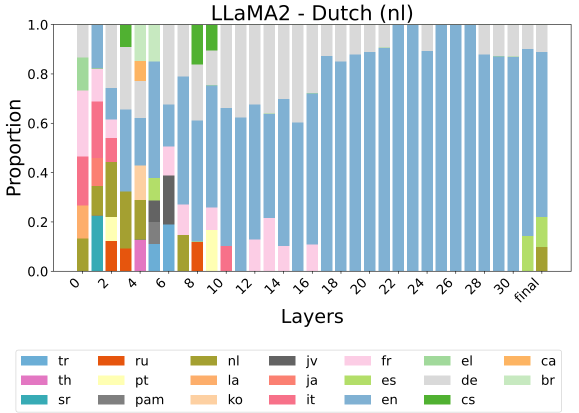
(a) Language composition in LLaMA2 with Hungarian as the input language.



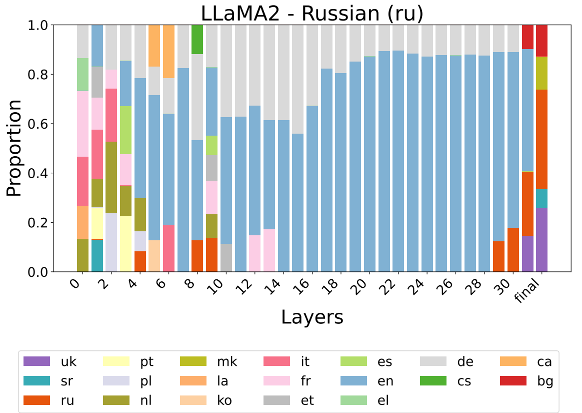
(b) Language composition in LLaMA2 with Japanese as the input language.



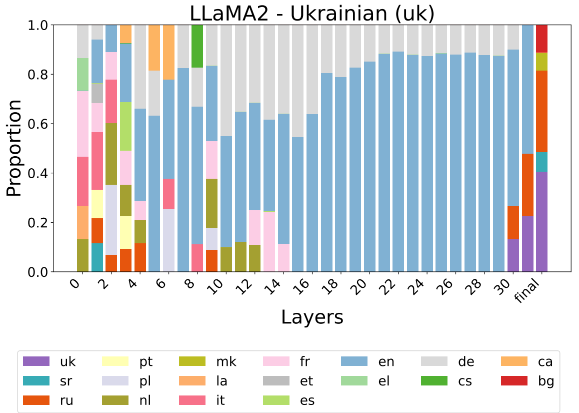
(c) Language composition in LLaMA2 with Korean as the input language.



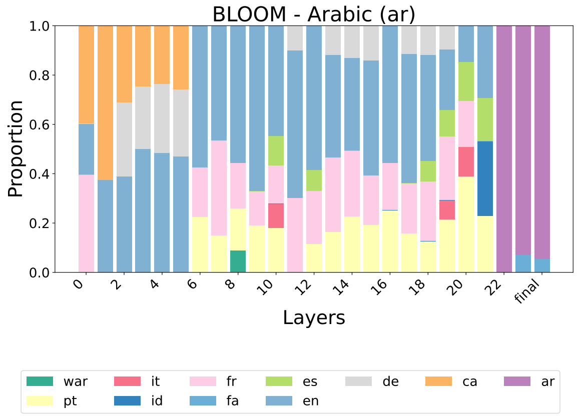
(d) Language composition in LLaMA2 with Dutch as the input language.



(e) Language composition in LLaMA2 with Russian as the input language.

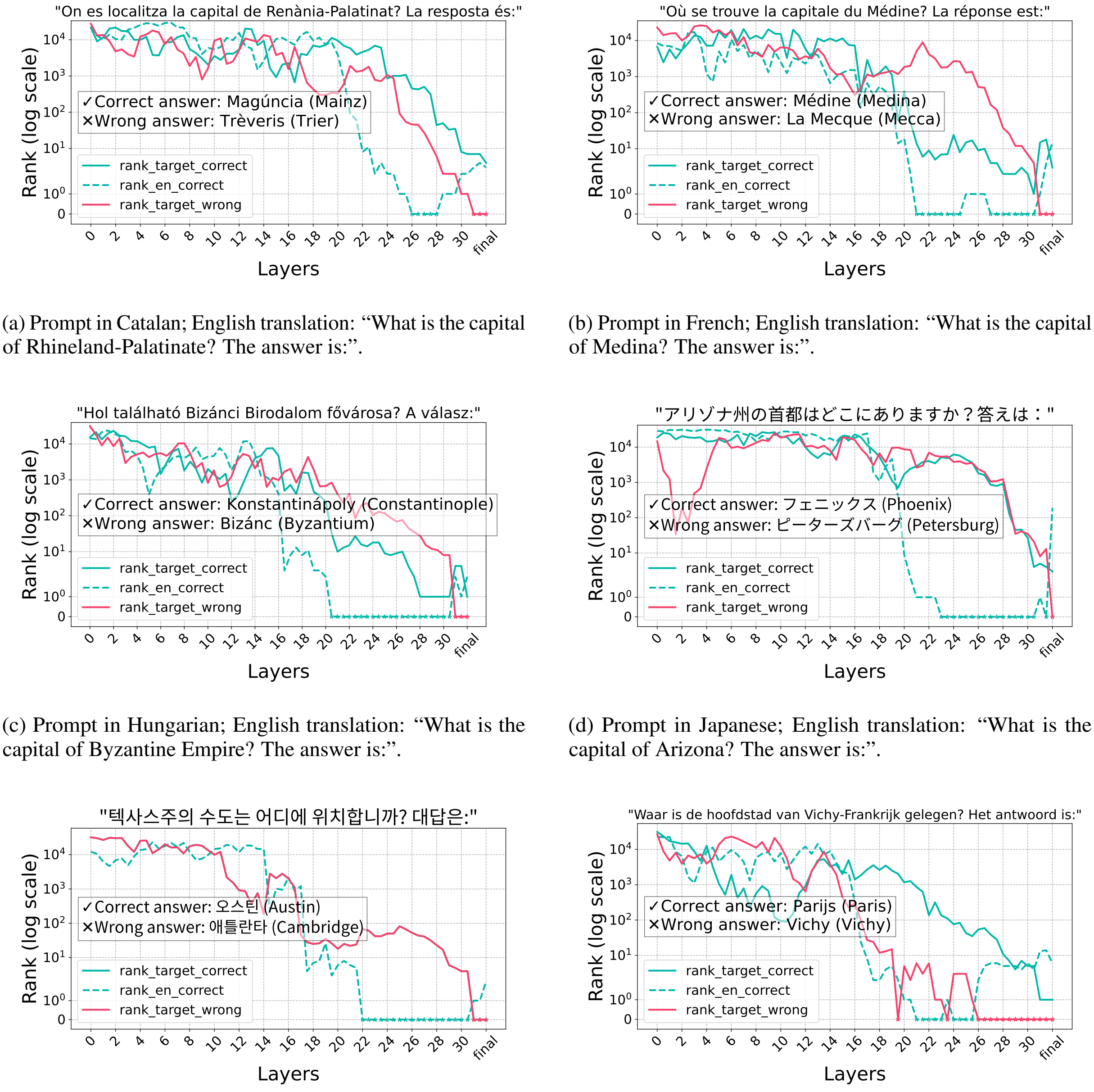


(f) Language composition in LLaMA2 with Ukrainian as the input language.



(g) Language composition in BLOOM with Arabic as the input language.

Figure 11: Language composition for unique languages in LLaMA2 and BLOOM, respectively.



(a) Prompt in Catalan; English translation: "What is the capital of Rhineland-Palatinate? The answer is:".

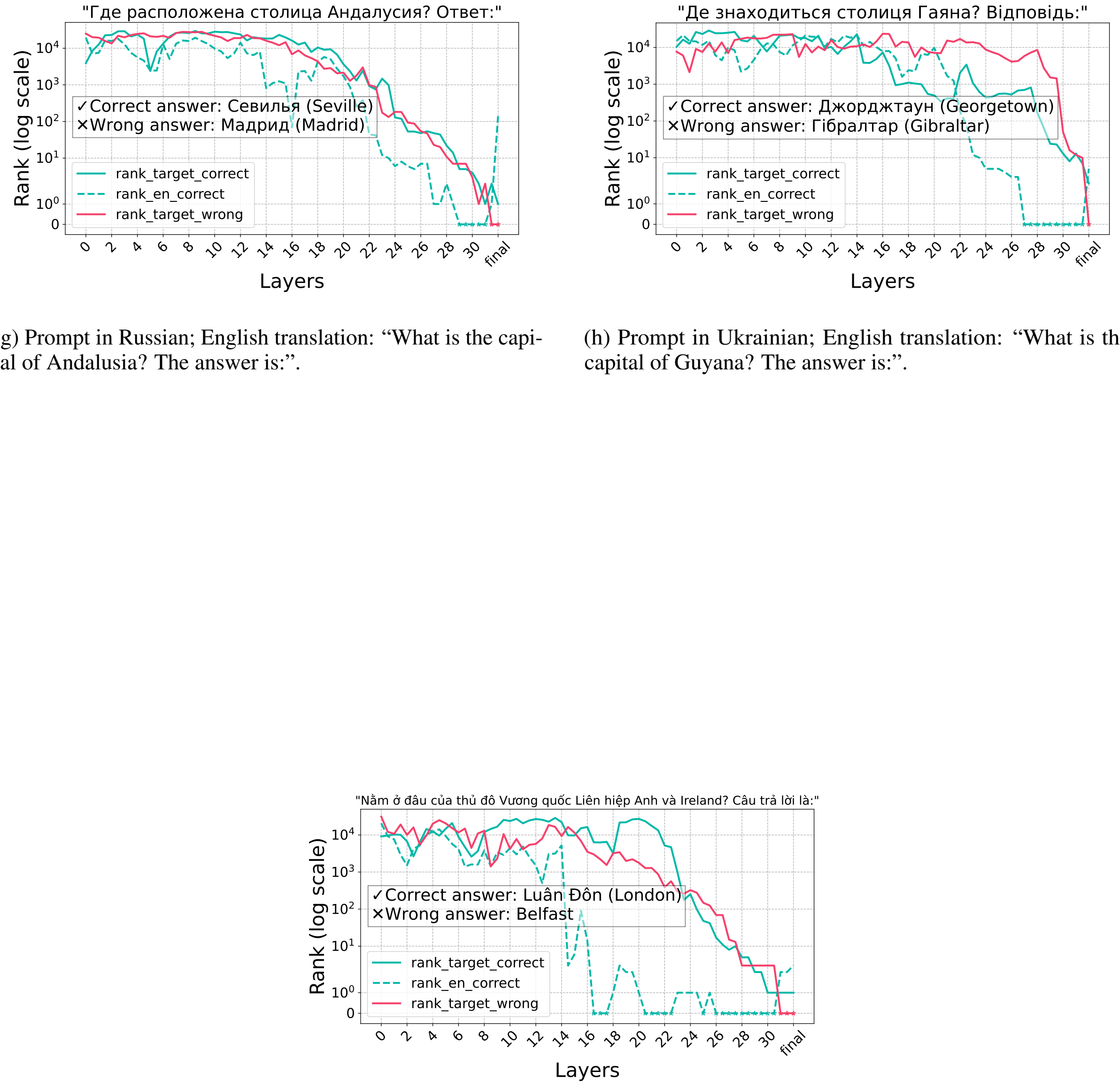
(b) Prompt in French; English translation: "What is the capital of Medina? The answer is:".

(c) Prompt in Hungarian; English translation: "What is the capital of Byzantine Empire? The answer is:".

(d) Prompt in Japanese; English translation: "What is the capital of Arizona? The answer is:".

(e) Prompt in Korean; English translation: "What is the capital of Texas? The answer is:".

(f) Prompt in Dutch; English translation: "What is the capital of Vichy France? The answer is:".



(g) Prompt in Russian; English translation: "What is the capital of Andalusia? The answer is:".

(h) Prompt in Ukrainian; English translation: "What is the capital of Guyana? The answer is:".

(i) Prompt in Ukrainian; English translation: "What is the capital of United Kingdom of Great Britain and Ireland? The answer is:".

Figure 12: Rank evolution for prompts in different languages. `rank_target_wrong` represents the rank of the model's final incorrect prediction across layers, while `rank_target_correct` and `rank_en_correct` denote the ranks of the correct answer in the target language and the English equivalent, respectively. The plots show the impact of errors during language transition, where the rank of the incorrect answer surpasses the correct answer in the final layers.