Are Knowledge and Reference in Multilingual Language Models Cross-Lingually Consistent?

Xi Ai [*] [1], **Mahardika Krisna Ihsani** [*] [†] [2], **Min-Yen Kan** [1]
[1] Web IR / NLP Group (WING), National University of Singapore
[2] Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
barid.x.ai@gmail.com, mahardika.ihsani@mbzuai.ac.ae, kanmy@comp.nus.edu

Abstract

Cross-lingual consistency should be considered to assess cross-lingual transferability, maintain the factuality of the model knowledge across languages, and preserve the parity of language model performance. We are thus interested in analyzing, evaluating, and interpreting cross-lingual consistency for factual knowledge. To facilitate our study, we examine multiple pretrained models and tuned models with code-mixed coreferential statements that convey identical knowledge across languages. Interpretability approaches are leveraged to analyze the behavior of a model in cross-lingual contexts, showing different levels of consistency in multilingual models, subject to language families, linguistic factors, scripts, and a bottleneck in cross-lingual consistency on a particular layer. Code-switching training and cross-lingual word alignment objectives show the most promising results, emphasizing the worthiness of cross-lingual alignment supervision and code-switching strategies for both multilingual performance and cross-lingual consistency enhancement. In addition, experimental results suggest promising result for calibrating consistency in the test time via activation patching.

1 Introduction

Frege's theory of reference (Frege, 1892) indicates that the knowledge conveyed by a sentence depends on the references of the expressions that make up the sentence. A salient aspect of humanity is that, while people may speak different languages, they can share common references and knowledge. Thus, references and knowledge must be consistent across languages, and a multilingual model serving as a knowledge base (Gupta and Srikumar, 2021; Kassner et al., 2021; Hu et al., 2024) should provide consistent knowledge when consulted in different languages. Not only does this theory contribute to cross-lingual performance and maintain knowledge between languages, but it also ensures parity and self-consistency of model performance (Hupkes et al., 2023; Wang et al., 2023). This motivates us to evaluate the knowledge consistency of multilingual language models in all languages when sharing the same references.

Few recent works (Kassner et al., 2021; Fierro and Søgaard, 2022; Qi et al., 2023) focused on translation pair consistency and reported that multilingual models may output knowledge for a particular query that differs with knowledge obtained from the query's translation. We argue that multilingual models show different language biases, leaving a non-trivial confounding factor when evaluating consistency with translation pairs. We hypothesize that for a consistent multilingual model, references, regardless of the surface language, provide energy to constrain the degree of freedom in knowledge recalling. To evaluate this hypothesis (Figure 1), we examine the difference in the output distribution between the original monolingual statement and the corresponding code-mixed coreferential statements, which takes a different angle and is orthogonal to existing works (Kassner et al., 2021; Qi et al., 2023) that rely on translation pairs and the output candidates. This examination explicitly instantiates Frege's theory of reference to check the consistency of knowledge across languages that the same references for sub-sentential expressions, e.g., entities, should result in the identical knowledge. We attempt to answer three questions: *1)* do multilingual language models recall factual knowledge for the coreferential statements in a similar manner, *2)* how does the mechanism of multilingual language models work on the incorporation between entities or references to convey knowledge in cross-lingual settings, and *3)* which factors prevent model consistency in multilingual settings?

In addition to model consistency in cross-lingual

[*]Equal contributions.
[†]Works conducted during the internship in WING@NUS

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 4975–5011

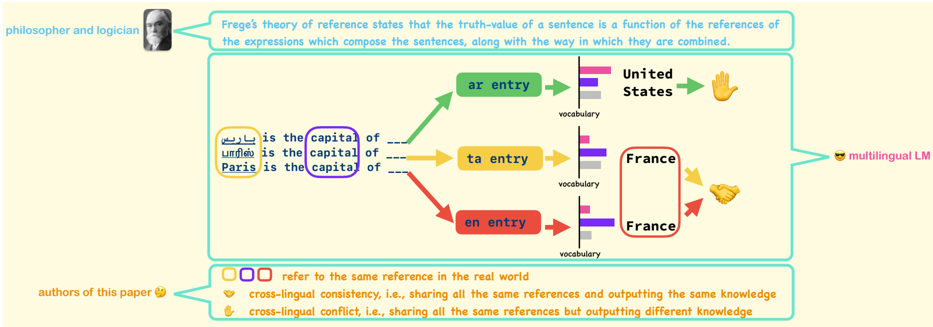


Figure 1: Illustration of Cross-lingual knowledge consistency. Frege's theory of reference defines the reference of a sub-sentential expression as the object singled out by the name. A salient aspect of humanity is that, they can understand knowledge based on references regardless of language.

settings, our study is related to a broader linguistic phenomenon of entity-level code-switching and language interference: an entity code-switches between two languages without changing the reference, as we create code-mixed coreferential statements from monolingual statements by substituting a subject entity with an equivalent one in another language that shares the same reference. More recently, we share a similar goal with knowledge incorporation and editing (Beniwal et al., 2024; Li et al., 2024), since we incorporate a coreferential entity from other languages to recall factual knowledge in cross-lingual settings. Our main findings are as follows.

- We present a code-mixed coreferential task to observe implicit consistency across languages within a sentence. In our experiments, observations and findings are also transferable to explicit cross-lingual consistency across translation pairs.
- We discover consistency bottlenecks and issues tied to language characteristics, scripts, and training biases through layer-wise analyses and interpretability approaches, which potentially prevent cross-lingual consistency improvements and gains from scaling.
- There is a partial causality from adding language biases (of high-resource languages) to improving cross-lingual knowledge consistency. Directly adding bias via representation patching could be a potential method to calibrate consistency in the test time.
- Shared language scripts contribute to cross-lingual consistency, especially for encoder and decoder models, but it is not a necessary condition to achieve it. Reducing script overlaps by expanding vocabulary size slightly improves the consistency yet it helps to improve the consistency for some low-resource languages.
- Cross-lingual supervision can alleviate the consistency bottleneck to enhance alignments between coreferential entities, which can be achieved by training with an explicit alignment objective or a code-switching objective. On the other hand, parallel samples providing cross-lingual generalization supervision offer limited gains to consistency.

Our contribution is to offer an understanding of multilingual language models' limitations under cross-lingual settings and highlight potential research directions to address such issues.

2 Methodology

2.1 Task Definition

We focus on a code-mixed, generative task that forces the multilingual model to condition on coreferential entities across languages to recall a factual answer from its internal knowledge base[1]. We show an example in Figure 1 where en entry "Paris is the capital of ___" is evaluated with its possible code-mixed coreferential statements (ar entry & ta entry). Readers can refer to Appendix §A.1 and §A.2 for details and implementations.

Let $I = \{S^{l1}, \cdots, O, \cdots\} \in l1$[2] be a statement, where $l1$ stands for matrix language (the predominant language), $S^{l1} = \{s_1, \cdots, s_k\} \in l1$ are subject sub-tokens, and $O = \{o_1, o_2, \cdots, o_j\} \in l1$ denote object sub-tokens. This statement

[1] See limitation in §8.

[2] The surface structure is not restricted. We use the common subject–object structure as an example.

is used to format an input I_{mono} by removing O to elicit the internal knowledge K_{θ}^{*} and instruct the model to output the n-gram $Cand(O_{\in V}|I_{mono})$ over the model's vocabulary V, where $Cand(O_{\in V}|I_{mono}) = P(O_{\in V}|K_{\theta}^{*})P(K_{\theta}^{*}|S^{l1}, I_{\backslash(S^{l1}\cap O)})$. Similarly, we create a code-mixed coreferential statement I_{cm} by replacing S^{l1} with a coreferential subject S^{l2} in the embedded language $l2$ to obtain $Cand(O_{\in V}|I_{cm}) = P(O_{\in V}|K_{\theta})P(K_{\theta}|S^{l2}, I_{\backslash(S^{l1}\cap O)})$. I_{cm} and I_{mono} with coreferential subjects S^{l1} and S^{l2} condition the model for recalling knowledge. To measure the knowledge consistency between K_{θ}^{*} and K_{θ}, we calculate the difference between the output $Cand(O_{\in V}|I_{cm})$ and $Cand(O_{\in V}|I_{mono})$ as K_{θ}^{*} and K_{θ} provide energies to constrain the degree of freedom in generation. Additionally, we also evaluated the baseline setting of I_{cm} by removing the subject entities to obtain the model's default outputs with no references for comparison.

We analyze the **consistency evolution** as the layer goes deeper to trace the consistency and understand the models' behavior. Specifically, we apply LogitLens (nostalgebraist, 2019) for encoder and encoder-decoder models or DecoderLens (Langedijk et al., 2023) for decoder models to computing the layer-wise output distributions from the layer representations, retrieving layer-wise $Cand(O_{\in V}|I_{cm})$ and $Cand(O_{\in V}|I_{mono})$.

2.2 Metric Function and Interpretability

Readers can refer to Appendix §A.3 for more details, e.g., equations.

Output Distributions Consistency. Top@1 Accuracy and RankC (i.e., weighted Precision@5) (Qi et al., 2023) are used to capture the difference between two output distributions, $Cand(O_{\in V}|I_{mono})$ and $Cand(O_{\in V}|I_{cm})$. In contrast to the previous works (Kassner et al., 2021; Qi et al., 2023), we do not constrain the output candidate or domain. Instead, the output distribution over the full vocabulary is examined. Since the experimental results in Top@1 and RankC are similar, Top@1 are moved to the Appendix §A.4.

Cross-lingual Representations Similarity. We hypothesize that cross-lingual generalization across languages results in cross-lingual consistency to some extent. To evaluate this hypothesis, we examine the contextualized representation similarity for our correferential statements by computing batch-wise CKA similarity scores (Kornblith et al., 2019) between them over each layer.

IG^2 Score. We adapted IG^2 (Liu et al., 2024) to interpret the impact of each feed-forward neuron on the output where the higher the value is, the more critical the neuron is to predict the ground truth object. This examination is used to analyze the correlation between cross-lingual consistency and shared neurons across languages.

2.3 Dataset and model

Dataset. We use mLAMA dataset (Kassner et al., 2021) that provides parallel triples (object, predicate, subject) in 53 languages written in cloze, completion task format (e.g., "Paris is the capital of ") to query knowledge in zero-shot settings. In our experiments, $l1$ is set to English for all pairs, and $l2$ is the other 52 languages to report an overall result, where $l2$ languages are categorized into two separate categories for each of the three factors (geographics, writing scripts, and language family) using ISO-639 language codes information from "localizely"[3]. For an in-depth analysis, we examine 2 similar $l2$ languages (De, Id) and 2 dissimilar $l2$ languages (Ar, Ta) to observe the consistency evolution from early layers to later ones[4].

Models. We examine distinct model families: encoder models (xlm-r from 0.3B to 10B) (Conneau et al., 2020)), encoder-decoder models (mT0 from 0.6B to 3.7B (Muennighoff et al., 2023), mT5 from 0.6B to 3.7B) (Xue et al., 2020)), and decoder models (Llama3-instruct 1B & 8B) (Grattafiori et al., 2024)). In our experiments, we obtain consistent findings across model families and sizes. Therefore, we show essential results in the main text and move the rest to the Appendix §A.4.

3 Observing Consistency

3.1 Consistency on All Languages

From Figure 2, dissimilar $l2$ tends to have lower consistency than those similar to $l1$ across all factors. The difference in writing scripts plays the most important role in both encoder and decoder models. However, surprisingly, encoder-decoder models are more tolerant to any kind of factors.

[3] https://localizely.com/language-code

[4] While Id does not belong to the same language family as En, it has many loanwords from En (Krause, n.d.). Ar and Ta are not considered as Indo-European languages and also do not use latin scripts.

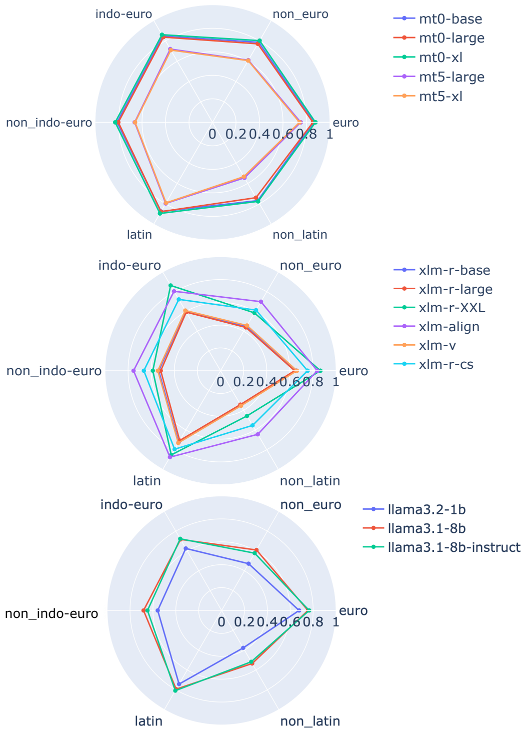


Figure 2: Cross-lingual consistency of output distribution in different model types (top: encoder-decoder, middle: encoder, bottom: decoder) grouped by 3 factors (geographics: europe & non_europe, language family: indo-european & non_indo-european, writing scripts: latin & non_latin). (*cf.* §A.4.2.)

Another intriguing finding is that geographic factor affects consistency, and this could be attributed to common culture and vocabulary (Zhao et al., 2024a). On the other hand, we suppose that other linguistic factors contributing to cross-lingual performance, such as similarity in linguistic characteristics (Chronopoulou et al., 2023), or borrowing (Tsvetkov and Dyer, 2016), could also affect cross-lingual knowledge consistency. However, such factors are difficult to quantify, leaving such analyses for future work. Note that language families and writing scripts have an impact on vocabulary, and we will confirm it in a later section.

3.2 Consistency Evolution across Layers

To better understand the cross-lingual consistency bottleneck, we examine the layer-wise consistency patterns in different model sizes and types, as presented in 1st and 2nd Row of Figure 3. For encoder and encoder-decoder models, the noticeable difference lies in the initial consistency, whereby dissimilar language pairs have low consistency scores. The consistency gap between dissimilar and similar languages starts to close at some specific layer while widening again later. Meanwhile, for decoder models, the pattern is more distinct, where there is a consistent degradation for smaller model in dissimilar language pairs and baseline, as for the larger model, it interestingly manages to recover the consistency starting from middle layer yet we can notice a bottleneck in last layer. This observation provides evidence for empirical studies that scaling benefits downstream task performance (Conneau et al., 2020), for example, XNLI, but offers limited gain for cross-lingual consistency, as we can observe in Figure 2.

Layer-wise analyses help us to understand the model behaviors. However, the question remains as to why such behaviors could happen. To answer that question, we analyze contextualized representation similarity across layers from the 3rd and 4th Row of Figure 3, which shows different patterns from the cross-lingual consistency. In general, for encoder-decoder and encoder models, there is a degradation of similarity scores until the middle layer (except for xlm-r-base, where the growth is slightly fluctuating). In contrast, the small decoder model shows more stable similarity over the layers, and there is also a monotonic increase until the middle layer for the larger decoder model. This finding suggests that the cross-lingual representation similarity improved via model scaling might be a necessary condition rather than a sufficient condition to achieve cross-lingual consistency. Some other factors, such as isotropy and contextualization (Ethayarajh, 2019), might impact cross-lingual consistency other than the cross-lingual representation similarity. In addition, dissimilar languages have low similarity scores that are quite similar to the baseline setting, which is also observed in the layer-wise consistency scores.

3.3 Correlation and Interpretability

To understand the model behaviors, we analyze the contribution of every neuron within MLP on the correferential statements based on findings from (Geva et al., 2020). Specifically, we inspect the IG^2 scores of all the feed-forward neurons at all the layers. Our analysis for this factor could show a moderate correlation with the cross-lingual consistency, as shown in Table 1. In Figure 4, the IG^2 scores for similar language pairs are almost the same, while there is a subtle difference for the dissimilar language pairs. This disparity on neurons could explain why the multilingual model is only highly consistent for certain language pairs.

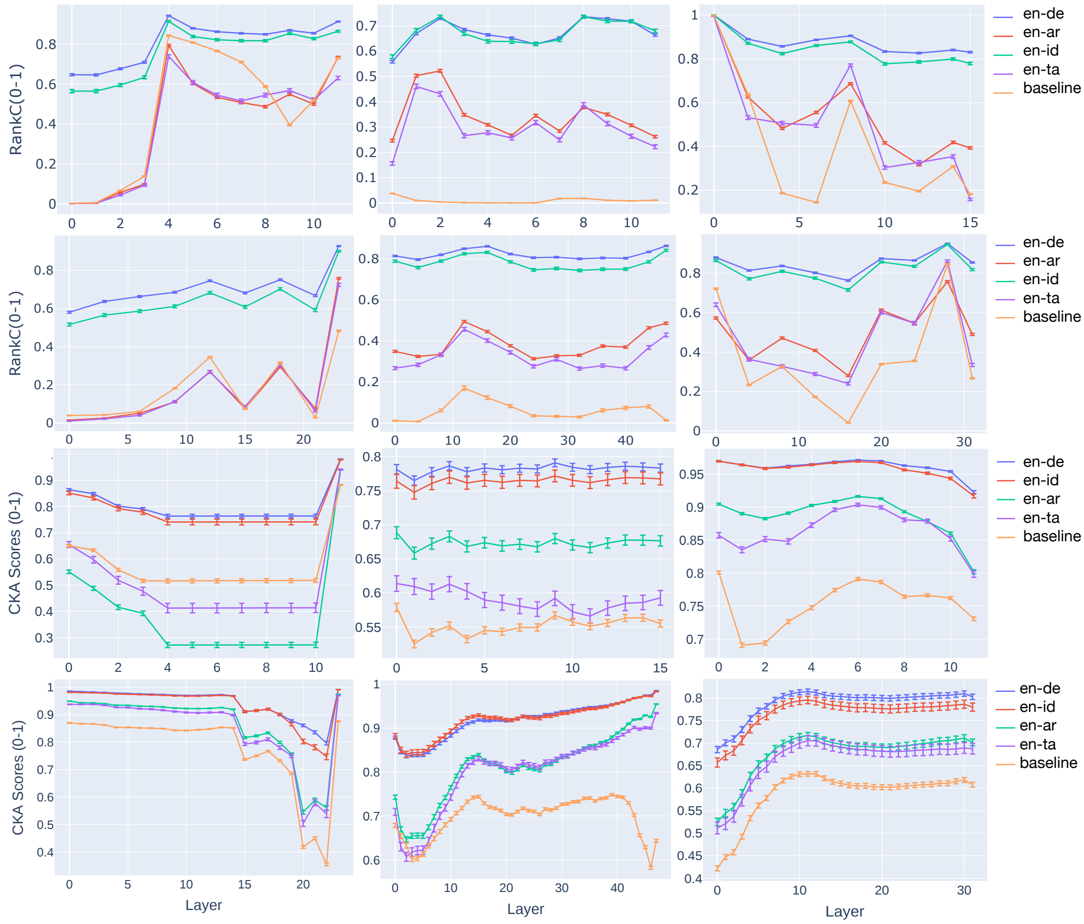


Figure 3: Consistency evolution (1st and 2nd row: consistency score for small and large models, 3rd and 4th row: CKA score for small and large models) in different model types (L: encoder-decoder, M: encoder, R: decoder). For each model family, scaling models is not a promising strategy in general to mitigate consistency bottlenecks when observing 1st row vs 2nd row and 3rd row vs 4th row (except for the xlm-r-xxl CKA similarity). (*cf.* §A.4.1)

	IG^2	
Model	RankC	Acc
mT0-base	0.528^*	0.519^*
mT0-large	0.705^*	0.699^*
xlm-r-base	0.400^*	0.397^*
xlm-r-large	0.508^*	0.481^*
llama3.2-1B-Instruct	0.544^*	0.489^*

Table 1: Statistical spearman ρ correlation ($\alpha = 0.05$) between average scores of layers with the patterns on each language model's IG^2 absolute difference.

Model	Codemixing Language	Patched FFN Layers
mt0-base	en–ta	[0,3,10,11]
	en–ar	[0,1,9,10]
mt0-large	en–ta	[0,1,19,20,21]
	en–ar	[0,1,19,20,21]
xlmr-base	en–ta	[5,8,9,10]
	en–ar	[5,7,8,10]
xlmr-large	en–ta	[0,2,5,19,20]
	en–ar	[17,18,19,20,21]
Llama 3.2-1B	en–ta	[2,5,10,12]
	en–ar	[2,5,10,12]
Llama 3.1-8B	en–ta	[5,10,15,18,20]
	en–ar	[5,10,15,18,20]

Table 2: Causal Intervention Hyperparameters Setup

4 Correlation between Consistency, Language Bias, and Cross-lingual Bias

4.1 Can Language Bias Calibrate Consistency in The Test Time?

From previous findings, we think of one question: *Can we add biases from I_{mono} to the feed-forward layers for consistency calibration in the test time?* Considering that two different patterns (on IG^2 scores) are discovered from our experiments and IG^2 score is moderately correlated with the consistency score, we perform one causal intervention on the feed-forward network to align the output of I_{cm} closer to the output of I_{mono} by patching I_{mono}'s activations of all tokens to I_{cm} in selected feed-forward neurons based on IG^2 (Vig et al., 2020; Geiger et al., 2021). This experiment measures whether each pattern has a causal relationship with cross-lingual consistency.

Specifically, we consider $a_i^{(l,p)}$ as the activation

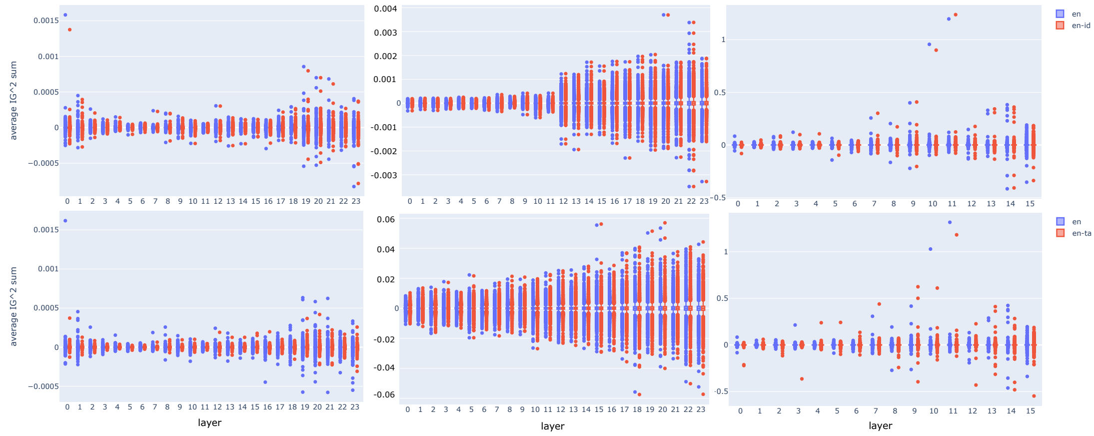


Figure 4: IG^2 scores in across different model types (L: encoder-decoder, M: encoder, R: decoder) for en–id (1st row) and en–ta (2nd row). The distribution is more contrastive on dissimilar languages (en–ta) than the similar languages (en–id). (*cf.* §A.4.4.)

of the i-th token on I_{mono} produced by the p-th neuron in the feed-forward network of the l-th layer, and then the patched activation value for the i-th token on I_{cm} is $\bar{a}_i^{(l,p)} = a_i^{(l,p)}$, in which we apply this to every new token. We intervene 4 different layers for base models and 5 different layers that have language-sensitive neurons based on IG^2 (i.e., layer which has noticeable IG^2 distribution difference between I_{mono} and I_{cm}). Table 2 lists the hyperparameters used in this experiment.

In Figure 5, there is a potential causal relationship between the activation intervention and consistency, subject to model architectures and sizes. Specifically, for encoder-decoder models, the intervention approach can increase the consistency scores in the middle-later layers only in the larger model, while such intervention does not offer substantiate gains for the smaller model. Similarly, we observe the effectiveness of the intervention in large encoder models but not in small encoder models. In contrast, the intervention shows effectiveness for small decoder models, but not for the large decoder models.

4.2 Vocabulary Expansion and Script Overlapping to Cross-lingual Consistency

We hypothesize that vocabulary size plays a crucial role in improving consistency, as 1) it allows a language model to potentially align semantics better due to preventing the model from latching onto shallow local signals or restoring words from subtokens (Levine et al., 2021)[5] and 2) it impacts the script overlapping across language. To test this hypothesis, we consider two similar language models, xlm-r-base and xlm-v-base (Liang et al., 2023), where xlm-v-base has a larger vocabulary (901,629 tokens) than xlm-r-base (250,002 tokens).

The vocabulary expansion offers a slight consistency improvement in any categorization, which is evident from the consistency difference between xlm-v-base and xlm-r-base in Figure 2. This finding challenges the conclusion in previous works (Kassner et al., 2021; Fierro and Søgaard, 2022; Qi et al., 2023), where sharing script is the key to cross-lingual consistency. Specifically, the base model shows better cross-lingual consistency in early layers due to the surface alignments via possible shared scripts. This can be observed from the study of representation similarity in Figure 6, where the base model shows strong alignments in early layers before final contextualization. However, such cross-lingual consistency cannot propagate to later layers. Compared to that, vocabulary-expanded models rely on deep semantic alignments in later layers for cross-lingual consistency. In Figure 6, the layer-wise consistency drops significantly in the base model's last layers but increases in the vocabulary-expanded model's last layers. On the other hand, more samples are required to generalize in the pre-training phase for the vocabulary expansion. Therefore, it alone cannot improve consistency significantly, especially for low-resource languages with limited corpora, but it still benefits dissimilar languages with lower consistency in the last layers to alleviate the consistency bottle-

[5]e.g., if the tokenizer splits the word "Tokyo" into ["To," "Kyo"], the token "To" is polysemous making thus the alignment of this word would be one-to-many, on the other hand, if a tokenizer keeps the word as it is, the tokenized form of the word is monosemous making it less ambiguous.

Figure 5: FFN intervention scores in different model types (L: encoder-decoder, M: encoder, R: decoder) with different model sizes (1st row: small models, 2nd row: large models). (*cf.* §A.5.1).

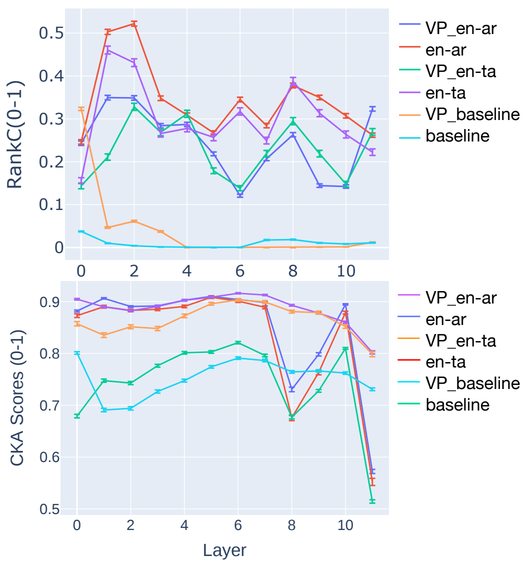


Figure 6: Effects of vocabulary expansion to consistency scores (top) and CKA scores (bottom). (*cf.* §A.5.2)

neck to some extent because of the deep semantic alignment regardless of the script overlapping. Overall, this finding shares the insight from Zhao et al. (2024a) where they found that the one-token P@1 of Afrikaans is higher than the Japanese due to segmentation and tokenization[6]. Additionally, we studied whether transliteration could help, and found that such a factor does not boost consistency, which we could attribute to the lack of semantic alignments in later layers (*cf.* §A.5.6).

[6]See discussions about a token parity issue in Figure 24.

4.3 Cross-lingual Supervisions to Cross-lingual Consistency

Lastly, we analyze how different training supervisions could contribute to the cross-lingual consistency. For this factor, we evaluated several training approaches: additional cross-lingual word alignment training (Chi et al., 2021), code-switching training (Whitehouse et al., 2022), multilingual multitask instruction tuning (Wei et al., 2021), and multilingual chat instruction tuning (Grattafiori et al., 2024). The former two strategies provide explicit alignments across languages, while the latter two strategies leverage the cross-lingual generalization from parallel samples for implicit alignments.

Overall, as presented in Figure 2, instruction tuning does not offer significant gains, but code-switching and word alignment training objectives improve the consistency significantly, especially for non-Latin script languages, which is not surprising as these objectives encourage models to align word knowledge and writing systems across languages. In addition, the alignment might also improve robustness for handling non-standard spellings and orthographic variations, which is observed in our case study for "transliteration vs translation" presented in §A.5.6. This finding may show the importance of adding explicit cross-lingual alignment in the training objective.

In the layer-wise analysis, from Figure 7, code-switching (2nd column, xlm-r-cs) and word align-

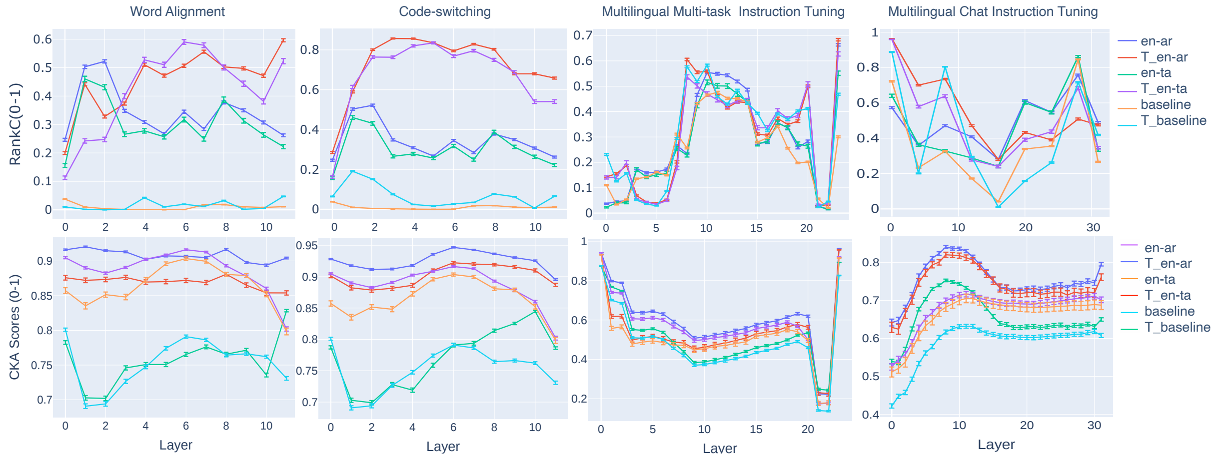


Figure 7: Consistency evolution (1st row: consistency score, 2nd row: CKA score) with different pre-training objectives (1st col: word alignment, 2nd col: code-switching training, 3rd col: instruction tuning on mt5, 4th col: instruction tuning on llama3.1-8b)

ment (1st column, xlm-align) training objectives could contribute to alleviating the consistency bottleneck occurring in middle layer onward, with word alignment showing the best effect. Such an attribute could cause the cross-lingual representation to be more consistently high as shown in 2nd row of Figure 7. On the other hand, instruction tuning with parallel samples, including the 3rd column (mt0 tuned from mt5) and the 4th column (Llama 3.1-8b-instruct tuned from Llama 3.1-8b) in the figure, does not offer a universal solution to the consistency bottleneck across model types. Specifically, it manages to slightly improve the cross-lingual consistency for the encoder-decoder, as shown in the 1st row of Figure 2 with mt0-base showing better consistency over mt5-xl. This could be attributed to additional parallel samples used in the instruction tuning, e.g., multilingual task datasets used for mt0. However, this is not a successful strategy for decoder models, where Llama 3.1-8b-instruct is not more cross-lingually consistent than the base model Llama 3.1-8b. For further analysis of the effect of each supervision on consistency, readers can refer to §A.5.3, §A.5.4, and §A.5.5.

5 Transferable Findings to Other Language Bias

Throughout this paper, studies are conducted on code-mixed coreferential statements between English and other languages. However, references and knowledge are universal. This raises a question: are all findings transferable to other coreferential entities and statements in non-English-centric scenarios? To answer this question, we conduct experiments for Llama-3.1-1B-Instruct, mt0-base, and xlm-r-base, using fr, vi and hy as the matrix (L_1) languages. Experimental results in Figure 8 are consistent with our main findings, providing evidence that our findings can be transferred to other language bias.

6 Related Work

Kassner et al. (2021) extended LAMA (Petroni et al., 2019) to a multilingual version multilingual version, mLAMA, and discovered that the language's relational knowledge capability varies in different languages, sharing similar findings with Schott et al. (2023); Zhao et al. (2024a) and other benchmarks (Wang et al., 2024a; Qi et al., 2023). Fierro and Søgaard (2022); Zhao et al. (2024a) studied the final predictions in different languages and reported inconsistencies across languages, especially for low-resource languages. Mousi et al. (2024) quantified the entity alignment in the shared space for the consistency goal, and Gao et al. (2024); Hua et al. (2024) further traced the alignments emerged from multilingual training. We take a different angle from those works in which we evaluate the consistency against code-mixed coreferential statements in cross-lingual settings.

Bhattacharya and Bojar (2023); Kojima et al. (2024); Tan et al. (2024); Miao and Kan (2025) discovered that a considerable portion of language-agnostic neurons encode universal concepts and utilize the latent language (in this case English). Zhao et al. (2024b); Wang et al. (2024c); Zhang et al. (2024) further showed that the cross-lingual downstream performance is potentially propor-

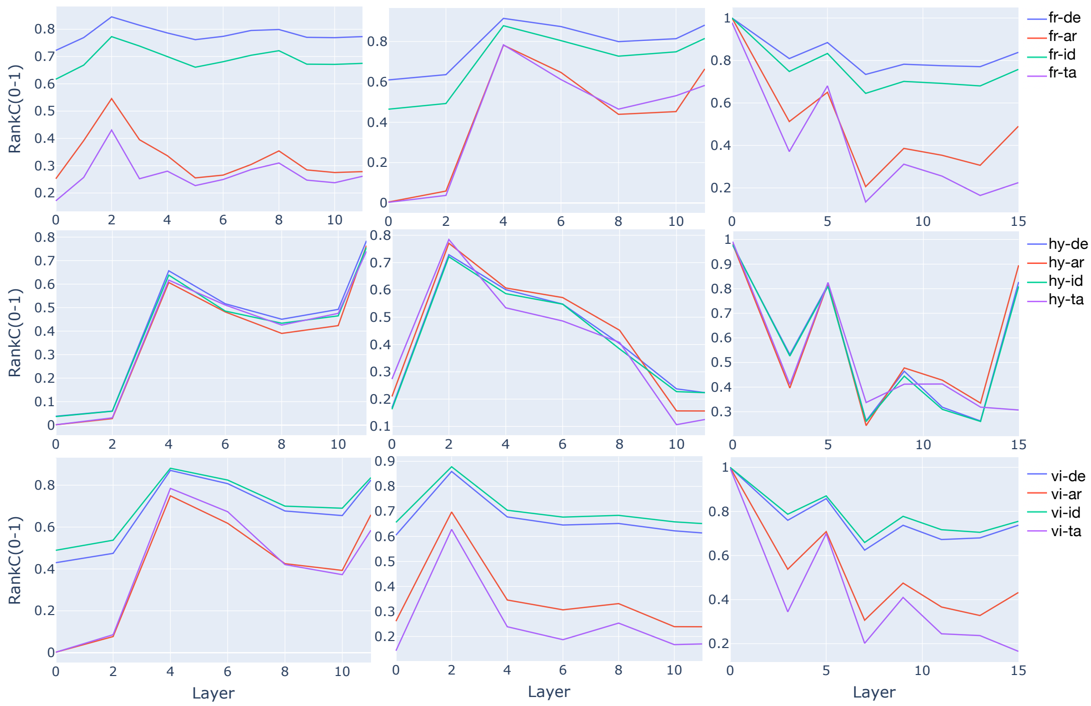


Figure 8: Consistency evolution across different model types (1st row: fr, 2nd row: hy, 3rd row:vi) on different non-english matrix languages (L: encoder-decoder, M: encoder, R: decoder). (*cf.* Figure 17)

tional to the number of language-agnostic neurons. Ferrando and Costa-jussà (2024) discovered a shared circuit or sub-network that is responsible for subject-verb agreement task for English & Spanish, and Stanczak et al. (2022); Wang et al. (2024b) found that morpho-syntax attributes have noticeable neuron overlapping degree over notable amount of language pairs. Wang et al. (2025) discovered three stages of cross-lingual factual recall in which the inconsistency occurred in the last stage called translation stage happening in later layers. In addition, they discovered that language models are able to recall the correct knowledge in the middle layers using the English concept, which is consistent with Wendler et al. (2024); Dumas et al. (2024). We trace consistent information and knowledge throughout the layers in cross-lingual settings, attempting to understand and interpret how commonly used strategies to improve multilingual models for downstream tasks could impact the cross-lingual knowledge consistency.

7 Conclusion

Our analysis reveals that knowledge consistency is highly dependent on model architectures, training strategies, deep semantic alignments, and language-specific information. Our layer-by-layer analysis of multilingual models uncovers a consistency bottleneck whereby the consistency does not grow monotonically on each layer. Our work highlights promising directions in the test-time calibration and training with cross-lingual alignment objectives to achieve knowledge consistency across languages, which will better preserve parity of language model performance and also alleviate such bottleneck. Cross-lingual representations, shared scripts and parallel samples might contribute to the cross-lingual consistency but are not a sufficient condition to achieve it.

We encourage researchers to work on representation learning approaches that induce cross-lingual alignment inductive bias explicitly to enhance alignments between coreferential entities. We also suggest test-time approaches that calibrate output distributions for knowledge consistency across languages. These methods can alleviate the consistency bottleneck and enhance alignments between coreferential entities, potentially improving both multilingual performance and cross-lingual consistency.

8 Limitations

A promising avenue for this work is evaluating cross-lingual knowledge consistency on other language models. Moreover, we only analyze each crucial component independently due to the time constraint and left scrutinizing the interaction between each component for future work (In particular, one can run any automatic circuit discovery algorithm (Syed et al., 2023; Conmy et al., 2023) to find subnetwork responsible for cross-lingual consistency and evaluate its performance). In the future, we may expand this work by analyzing how the interaction among these components could affect the cross-lingual consistency of multilingual models. Another thing is that our causal intervention method needs to be done manually, and we suspect that this method could produce a side effect on the model because the representations encoded by language models are more likely to be polysemous. In addition, we only evaluate language models in context-independent settings. Thus, in the future, we plan to evaluate the consistency of the models' knowledge and observe whether language models utilize their parametric knowledge more or emphasize the knowledge from the given context under the cross-lingual setting. Another thing to consider is that we only evaluate our solution using some particular models due to the time constraint. One interesting thing to explore in this aspect is to see whether adversarial training and multi-agent setting could help to enhance cross-lingual consistency. Moreover, we use an assumption that one reference is represented as a single English object entity to make the evaluation tractable; hence, we do not take into account the real-world setting where one reference can be interpreted in different ways on multiple languages (e.g., "China" is written as "ZhongGuo" in Chinese rather than "China"). Lastly, our research scope assumes that the knowledge we want to evaluate is factual and not dependent on subjective aspects (e.g., cultural context). With that assumption, we assume that references here generally have one-to-one mapping to representation in one language where the representation here is considered common knowledge.

9 Ethics Statement

This work aims to evaluate the consistency of the language model across different senses (particularly between a monolingual input and its code-mixed counterparts) and the impact of different factors on that metric. Doing such a study could shed light on the limitations of language models and think of the mitigations of such matters.

10 Reproducibility Statements

We used open-source pretrained models and also dataset for all of the reported experiments thus no undisclosed assets utilized in our work. Additionally, we also provide necessary experiments' output and codes on `https://github.com/baridxiai/knowledgeConsistencyAndConflict`.

11 Acknowledgments

We would like to thank several group members from the Web Information Retrieval / Natural Language Processing Group (WING) at NUS and also some group members from the Research in Text Understanding and Analysis of Language (RiTUAL) at MBZUAI for providing constructive feedbacks on our paper especially Prof. Thamar Solorio. We also grateful to NUS and MBZUAI for providing computing resources to run experiments on this paper.

References

Himanshu Beniwal, Kowsik D, and Mayank Singh. 2024. Cross-lingual editing in multilingual language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2078–2128, St. Julian's, Malta. Association for Computational Linguistics.

Sunit Bhattacharya and Ondrej Bojar. 2023. Unveiling multilinguality in transformer models: Exploring language specificity in feed-forward networks. *arXiv preprint arXiv:2310.15552*.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. *arXiv preprint arXiv:2106.06381*.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2023. Language-family adapters for low-resource multilingual neural machine translation. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 59–72, Dubrovnik, Croatia. Association for Computational Linguistics.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 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, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Javier Ferrando and Marta R Costa-jussà. 2024. On the similarity of circuits across languages: a case study on the subject-verb agreement task. *arXiv preprint arXiv:2410.06496*.

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. Association for Computational Linguistics.

Gottlob Frege. 1892. On sense and reference.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. 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 6101–6117, Mexico City, Mexico. Association for Computational Linguistics.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xin-feng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2024. Towards understanding factual knowledge of large language models. In *The Twelfth International Conference on Learning Representations*.

Tianze Hua, Tian Yun, and Ellie Pavlick. 2024. mOthello: When do cross-lingual representation alignment and cross-lingual transfer emerge in multilingual models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1585–1598, Mexico City, Mexico. Association for Computational Linguistics.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5:1161–1174.

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.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR.

Wayne B. Krause. n.d. English & indonesian similarities & differences. Accessed: 2024-10-01.

Anna Langedijk, Hosein Mohebbi, Gabriele Sarti, Willem Zuidema, and Jaap Jumelet. 2023. Decoderlens: Layerwise interpretation of encoder-decoder transformers. *arXiv preprint arXiv:2310.03686*.

Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. 2021. {PMI}-masking: Principled masking of correlated spans. In *International Conference on Learning Representations*.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2024. Unveiling the pitfalls of knowledge editing for large language models. In *The Twelfth International Conference on Learning Representations*.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models. *arXiv preprint arXiv:2301.10472*.

Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. 2024. The devil is in the neurons: Interpreting and mitigating social biases in language models. In *The Twelfth International Conference on Learning Representations*.

Yisong Miao and Min-Yen Kan. 2025. Discursive circuits: How do language models understand discourse relations? In *The 2025 Conference on Empirical Methods in Natural Language Processing*.

Basel Mousi, Nadir Durrani, Fahim Dalvi, Majd Hawasly, and Ahmed Abdelali. 2024. Exploring alignment in shared cross-lingual spaces. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6326–6348, Bangkok, Thailand. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

nostalgebraist. 2019. Interpreting gpt: The logit lens. Accessed: 2024-08-04.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553.

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.

Tim Schott, Daniel Furman, and Shreshta Bhat. 2023. Polyglot or not? measuring multilingual encyclopedic knowledge in foundation models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11253, Singapore. Association for Computational Linguistics.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.

Aaquib Syed, Can Rager, and Arthur Conmy. 2023. Attribution patching outperforms automated circuit discovery. *arXiv preprint arXiv:2310.10348*.

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. *arXiv preprint arXiv:2404.11201*.

Yulia Tsvetkov and Chris Dyer. 2016. Cross-lingual bridges with models of lexical borrowing. *Journal of Artificial Intelligence Research*, 55:63–93.

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

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024a. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. 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 370–390, Mexico City, Mexico. Association for Computational Linguistics.

Hetong Wang, Pasquale Minervini, and Edoardo M Ponti. 2024b. Probing the emergence of cross-lingual alignment during llm training. *arXiv preprint arXiv:2406.13229*.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. 2025. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. *arXiv preprint arXiv:2504.04264*.

Weixuan Wang, Barry Haddow, Minghao Wu, Wei Peng, and Alexandra Birch. 2024c. Sharing matters: Analysing neurons across languages and tasks in llms. *arXiv preprint arXiv:2406.09265*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.

Chenxi Whitehouse, Fenia Christopoulou, and Ignacio Iacobacci. 2022. EntityCS: Improving zero-shot cross-lingual transfer with entity-centric code switching. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6698–6714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Unveiling linguistic regions in large language models. *arXiv preprint arXiv:2402.14700*.

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 Appendix

A.1 Our Task Definition

We focus on a code-mixed context-independent cloze task that forces the multilingual model to rely on its internal knowledge base to recall the common knowledge shared by coreferential entities across languages due to cross-lingual generalization[7]. In the following introduction, we will define the evaluation task mathematically. Let $I = \{S^{l1}, \cdots, O, \cdots\} \in l1$[8] be a statement, where $l1$ stands for matrix language (the predominant language), $S^{l1} = \{s_1, \cdots, s_k\} \in l1$ are subject sub-tokens, and $O = \{o_1, o_2, \cdots, o_j\} \in l1$ denote object sub-tokens. This statement is used to create a cloze task input $I_{mono} = \{S^{l1}, \cdots, M, \cdots\}$, where M is the mask token used to substitute O in I (i.e., the mask $M = \{mask_1, \cdots, mask_j\}$ in encoder models, the sentinel token $M =< extra\_id\_0 >$, or the next token in decoder models). We define n-gram prediction for O from M, denoted as $Cand(O_{\in V} | I_{mono})$, as the top-k n-gram candidates obtained from beam search decoding over the model's vocabulary V. Similarly, we can create a code-mixed coreferential statement I_{cm} by

[7] See limitation in §8.

[8] The surface structure is not restricted. We use the common subject–object structure as an example.

replacing S^{l1} with a coreferential subject S^{l2} in the embedded language $l2$ (the subsidiary language) in order to obtain $Cand(O_{\in V}|I_{cm})$. Therefore, I_{cm} and I_{mono} are coreferential and expected to recall the same knowledge. Finally, we define the consistency of cross-lingual knowledge as $0 \leq f_{metric}(Cand(O_{\in V}|I_{mono}), Cand(O_{\in V}|I_{cm})) \leq 1$, where f_{metric} is a consistency metric defined in the next subsection. If $f_{metric} = 1$, it implies that multilingual language models recall the factual knowledge for the coreferential statements I_{mono} and I_{cm} in an identical manner. The coreferential statements are fully inconsistent if $f_{metric} = 0$. Note that we do not consider whether the prediction is correct. Instead, f_{metric} evaluates the parity and consistency across languages in which the model is expected to produce similar candidates for I_{mono} and I_{cm}.

From a probability view, we can define our task as measuring the difference between two distributions, $Cand(O_{\in V}|I_{cm}) = P(O_{\in V}|K_\theta)P(K_\theta|S^{l2}, I_{\setminus(S^{l1}\cap O)})$ and $Cand(O_{\in V}|I_{mono}) = P(O_{\in V}|K^*_\theta)P(K^*_\theta|S^{l1}, I_{\setminus(S^{l1}\cap O)})$, where K_θ is the knowledge recalled from the model given the preceding context, and $I_{\setminus(S^{l1}\cap O)}$ stands for I without both the subject and the object. Then, cross-lingual knowledge consistency between K^*_θ and K_θ reflects on the measured difference.

The high-level idea of this evaluation task is illustrated in Figure 1 where en entry "Paris is the capital of ___" is evaluated with its possible code-mixed statements (ar entry & ta entry). Additionally, we also evaluated the baseline setting of I_{cm} by removing the subject entities to obtain the model's default object tokens for comparison. In this example, S^{l1}, $I_{\setminus(S^{l1}\cap O)}$, and S^{l2} are "Paris", "is the capital of", and the ar or ta entry for "Paris", respectively. If coreferential subject entries are trained to generalize across languages, we could observe the cross-lingual consistency. In addition, we are aware of a baseline from this probability view. Specifically, we define the baseline as the difference between $Cand(O_{\in V}|I_{mono})$ and $Cand(O_{\in V}|I_{\setminus(S^{l1}\cap O)}) = P(O_{\in V}|K^\alpha_\theta)P(K^\alpha_\theta|I_{\setminus(S^{l1}\cap O)})$, measuring agnostic consistency without the coreferential subjects S^{l1} and S^{l2} in cross-lingual settings. In implementation, we mask the both subject and object entities to create the "code-mixed" counterpart as the baseline. Readers can refer to Appendix §A.2 for our implementation.

A.2 Input Format

In our task definition, we introduce our evaluation task in both intuition and math perspective. Here is the input sample in Table 3, 4, 5. Meanwhile, as presented in the task definition, we do not consider whether predictions are true but focus on the same prediction distributions regardless of languages. Note that we did not perturb the surface structure in order to minimize variables to affect factual knowledge recall because S^{l2} "switches-in" at grammatically correct point as the new subject (Pratapa et al., 2018).

A.3 Metric Function and Interpretability Approach

RankC RankC (Qi et al., 2023) is used to evaluate the cross-lingual knowledge consistency. Given a set of statements S where each of the statement having each own I_{mono} and I_{cm}, the number of candidates $Cand(O_{\in V}|I_{mono})$ of i-th statement N_i, $mono^j$ stands for the j-th candidate of $Cand(O_{\in V}|I_{mono})$, cm^j stands for the j-th candidate of $Cand(O_{\in V}|I_{cm})$, and the RankC score of $Cand(O_{\in V}|I_{mono})$ concerning $Cand(O_{\in V}|I_{cm})$ can be written as

$$RankC(cm, mono) \tag{1}$$

$$= \frac{\sum_{i=1}^{|S|}\sum_{j=1}^{N_i} \frac{e^{N_i-j}}{\sum_{k=1}^{N_i} e^{k-j}} * P@j}{|I_{mono}|}, \tag{2}$$

$$P@j = \frac{1}{j}|\{cm_i^1, cm_i^2, \cdots, cm_i^j\}\cap \tag{3}$$

$$\{mono_i^1, mono_i^2, \cdots, mono_i^j\}|. \tag{4}$$

Top@1 Accuracy The Top@1 accuracy is defined as the average number of exact matches between the top-1 predictions given I_{mono} and I_{cm}.

IG^2 Score If $w_j^{(l)}$ is the activation value of j-th neuron in the l-th layer of a particular input (either code-mixed or not), m is the approximation step, and t as a token of the whole ground truth object entity, the score for a given I_{mono} or I_{cm} is defined as

$$IG^2(w_j^{(l)}) = \sum_{t\in T} \frac{\frac{w_j^{(l)}}{m}\sum_{k=1}^{m} \frac{\partial P(t|\frac{k}{m}w_j^{(l)})}{\partial(\frac{k}{m}w_j^{(l)})}}{|T|} \tag{5}$$

A.4 Findings in Details

A.4.1 Layer-wise Consistency

Refer to Figure 9, 10, and 11.

	xlm-r input
I_{mono}	Paris is the capital of <**mask**>
I_{cm}	باريس is the capital of <**mask**>
$I_{\setminus(S^{l_1}\cap O)}$	<mask> is the capital of <**mask**>

Table 3: Input sample for the evaluation task for xlm-r. We only predict the object in bold. $I_{\setminus(S^{l_1}\cap O)}$ is the baseline input.

	mt0 input
I_{mono}	Paris is the capital of <**extra_id_0**>
I_{cm}	باريس is the capital of <**extra_id_0**>
$I_{\setminus(S^{l_1}\cap O)}$	<extra_id_0> is the capital of <**extra_id_1**>

Table 4: Input sample for the evaluation task for mt0. We only predict the object in bold. $I_{\setminus(S^{l_1}\cap O)}$ is the baseline input.

	Llama input
I_{mono}	Finish the cloze question with words. Do not give additional comments. Question: Paris is the capital of _. Answer:
I_{cm}	Finish the cloze question with words. Do not give additional comments. Question: باريس is the capital of _. Answer:
$I_{\setminus(S^{l_1}\cap O)}$	Finish the cloze question with words. Do not give additional comments. Question: _is the capital of _. Answer:

Table 5: Input sample for the evaluation task for llama 3. We only predict the object in bold. $I_{\setminus(S^{l_1}\cap O)}$ is the baseline input.

A.4.2 Overall Consistency of Output distribution

Refer to Figure 12, 14, and 15.

A.4.3 Consistency of Non-English Matrix Languages

Refer to Figure 12, 14, and 15.

A.4.4 Feed-Forward Neurons' Gradients Sum

Refer to Figure 18, 19, and 19

A.5 Improving Consistency

A.5.1 Adding Monolingual Bias

A.5.2 Impact of Larger Vocabulary

When expanding the vocabulary size, we found on Figure 26 that such method causes marginal improvement. Furthermore we conducted correlation analysis and based on Figure 24, we discovered no correlation between token parity seen in table and consistency improvement and this explains why we observed such limited improvement.

A.5.3 The Effect of Cross-Lingual Word Alignment Training Objective to The Cross-lingual Consistency

Another possible hypothesis is that there might be an entanglement of features between linguistic and knowledge features. (Elhage et al., 2022) discovered that a neural network could fit multiple features into one dimension at the price of more entangled features, and this entanglement could cause tokens not cross-lingually aligned, as there may be an entanglement between syntactic and semantic features within one dimension. Inspired by that, we suspect that this might hinder the consistency of language models. To test this assumption, we evaluated two similar language models in which one model is trained solely on MLM objective (xlm-r), and another similar model is trained on one additional objective to align word translations (xlm-align (Chi et al., 2021)), where this word alignment could be helpful in aligning references across languages.

Word alignment increases cross-lingual consistency monotonically to alleviate the cross-lingual bottleneck. Similar to the vocabulary expansion, this strategy does not improve the consistency for the baseline as we would expect. The aligned model outperforms the baseline starting from the middle layers in Figure 27. Multiple pre-training objectives that could approximately disentangle different features can help preserve the model's knowledge of different languages. We could also confirm this finding by observing the overall cross-lingual

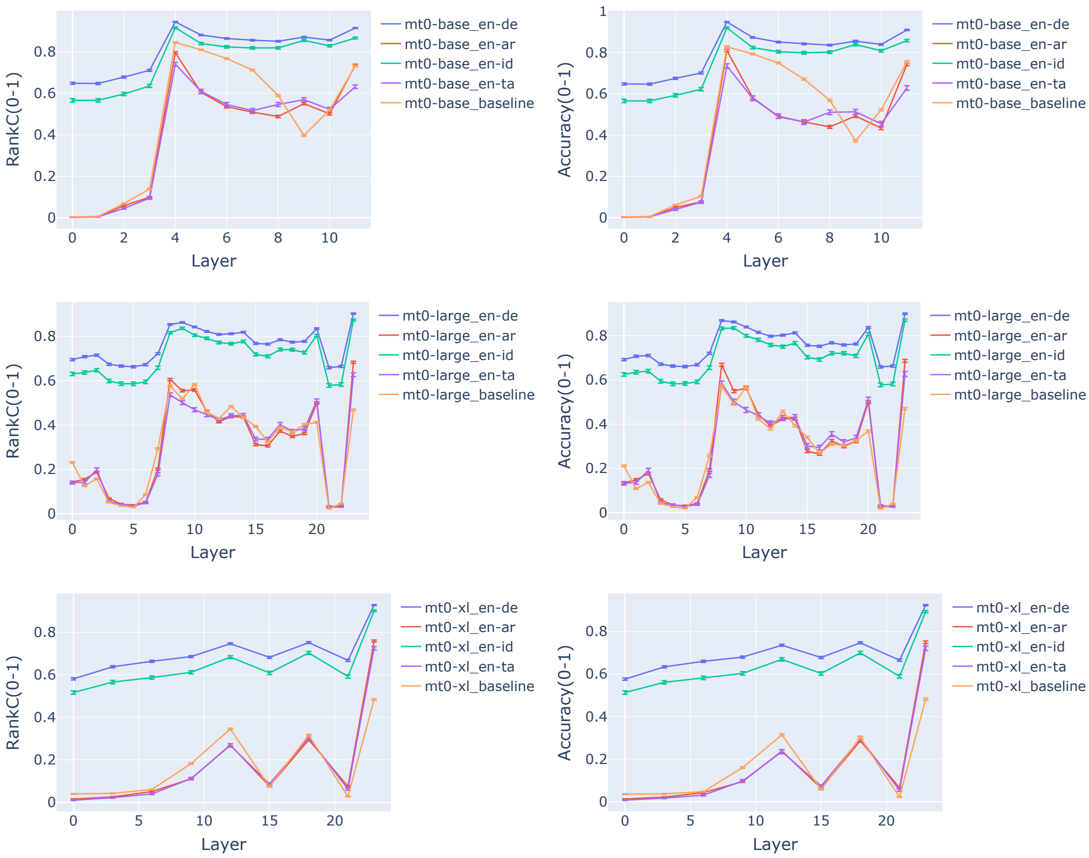


Figure 9: mT0 (base, large, XL) layer-wise cross-lingual consistency scores (left: RankC, right: Top@1)

consistency result in. In addition, word alignments improve consistency for transliterations or similar orthographical forms, contributing to model's robustness against orthographic variations and non-standard spellings, but vocabulary expansion can not offer such gains.

A.5.4 The Effect of Code-switching Training to The Cross-lingual Consistency

Inspired by the experiment on cross-lingual supervision, we further hypothesize that code-switching training, which substitutes an entity with alternatives from other languages for intra-sentential alignments in cross-lingual settings, can help the model understand common knowledge across languages for cross-lingual consistency to some extent. To evaluate this hypothesis, we study xlm-r and xlm-r-cs (Whitehouse et al., 2022), where xlm-r-cs is continuously trained on code-switching corpus from xlm-r-base and shows high performance in multilingual fact-checking. From Figure 29, we observe a shift in the consistency bottleneck from the middle layers to the later layers of xlm-r-cs, where the consistency gap between dissimilar and similar languages narrows in xlm-r-cs compared to xlm-r in the middle layers. Overall, code-switching can offer significant gains to the cross-lingual consistency, even without additional objectives.

A.5.5 The Effect of Multi-task Fine-tuning to The Cross-lingual Consistency

We hypothesize that method of fine-tuning can improve the cross-lingual consistency due to improved cross-lingual generalization across similar tasks in different languages, as opposed to word-level alignments discussed in previous sections. Surprisingly, multi-task fine-tuning can not offer significant gains to the layerwise cross-lingual consistency. As presented in Figure 31 and Figure 32, the consistency patterns are quite similar for both type of model families (decoder is represented by llama3.1-8b-instruct, encoder-decoder is represented by mt0-large). Intriguingly, we can more salient enhancement on encoder-decoder models as shown in Figure 33 than decoder models as evident in Figure 34 which might suggest the possibility of ratio of multilingual examples in the pretraining corpora could play role on such improvement.

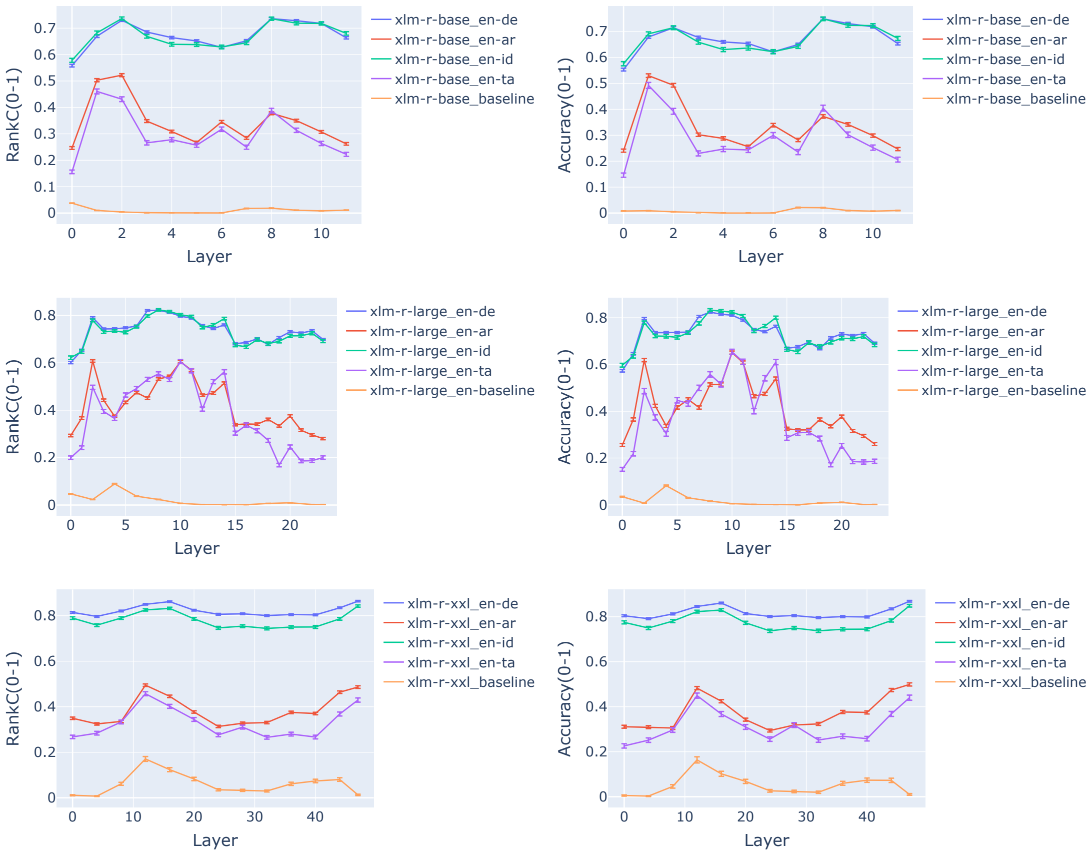


Figure 10: xlm-r (base, large, XXL) layer-wise crosslingual consistency scores (left: RankC, right: Top@1)

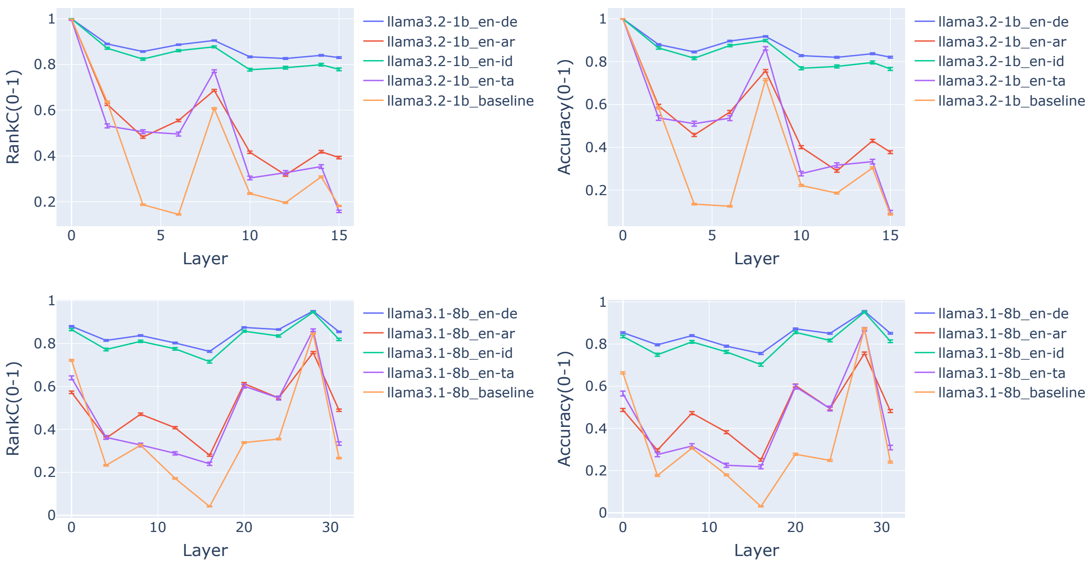


Figure 11: llama 3 (1B, 8B) layer-wise cross-lingual consistency scores (left: RankC, right: Top@1)

A.5.6 Case Study for Transliteration

Instead of using translations, we transliterate bn[9] and ar[10] to understand the impact of writing systems, particularly transliterations. As presented in Figure 35, word alignments (or the similar effect from CS training) contribute to the model's cross-lingual consistency against writing systems because xlm-align and xlm-r-cs show similar performance in both original and transliteration settings.

[9]https://github.com/shhossain/BanglaTranslationKit
[10]https://github.com/hayderkharrufa/arabic-buckwalter-transliteration

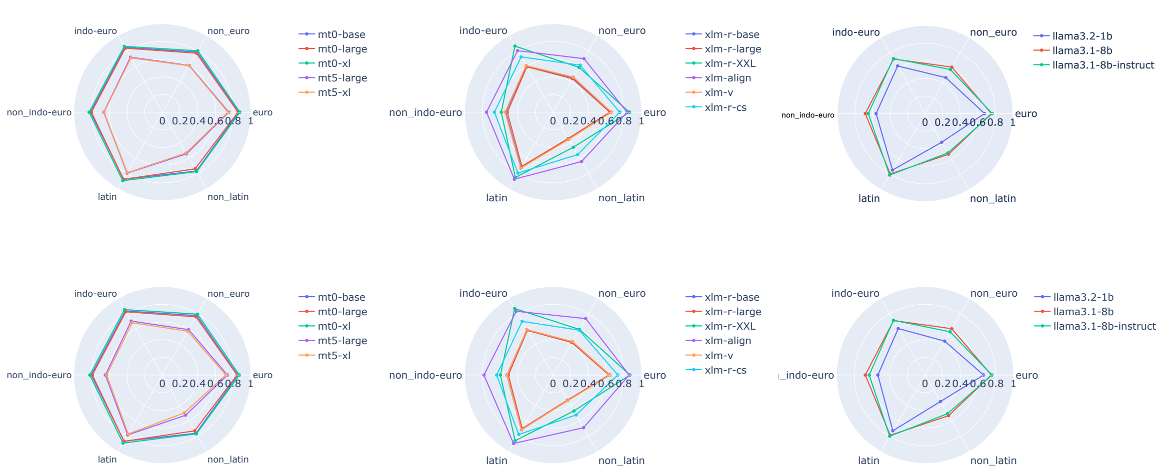
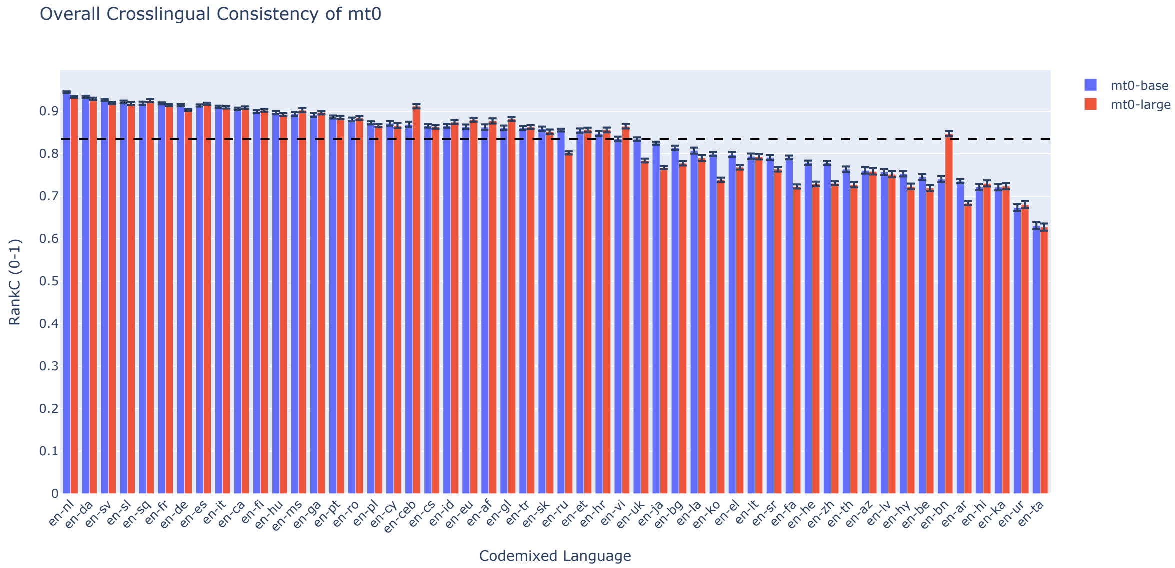


Figure 12: Overall cross-lingual consistency across different transformer types (left: encoder-decoder, middle: encoder, right: decoder) grouped by 3 factors (geographics: europe & non_europe, language family: indo-european & non_indo-european, writing scripts: latin & non_latin).

Meanwhile, we can observe that xlm-align and xlm-r-cs significantly improve the overall performance for non-Latin scripts in §A.5.3 & §A.5.4. This is reasonable as word alignments or CS training help the model link original words with their translations or transliterations, depending on the training corpus, thereby enhancing cross-lingual consistency. We suspect that these word alignments might also improve robustness for handling non-standard spellings and orthographic variations. However, xlm-v-base and xlm-r-base without word alignment benefit from transliterations, which means that xlm-v-base and xlm-r-base do not sufficiently align original words with their transliterations to main cross-lingual consistency. It is also confirmed by the overall performance of vocabulary expansions in §A.5.2, where vocabulary expansions can not offer significant gains for cross-lingual consistency. Overall, the evaluation task does not inadequately boost consistency for languages using Latin script because word alignments resulting in cross-lingual consistency are the main factor.



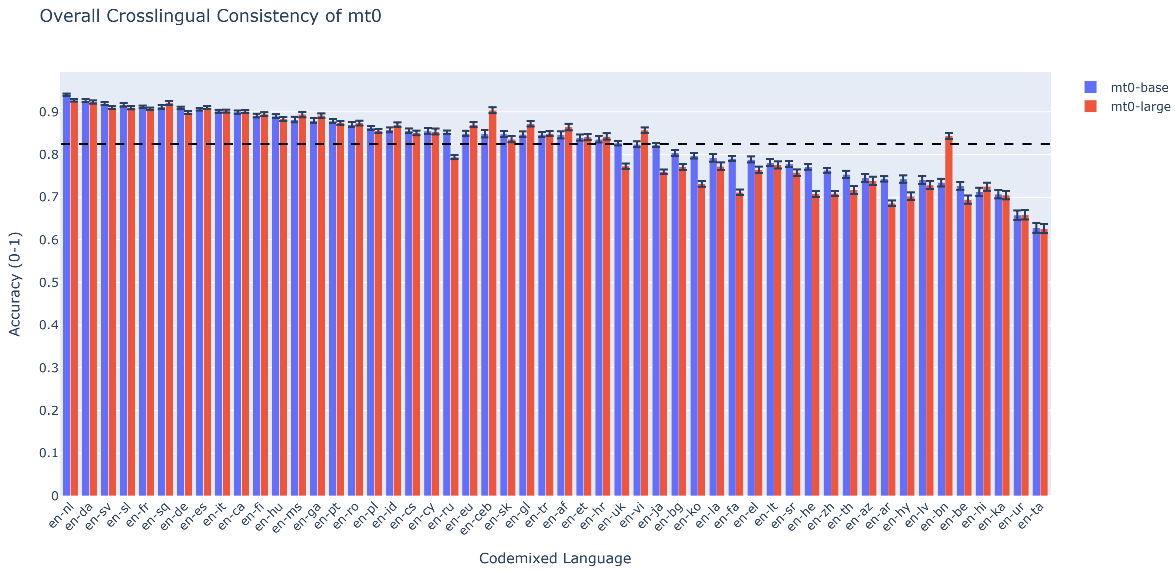
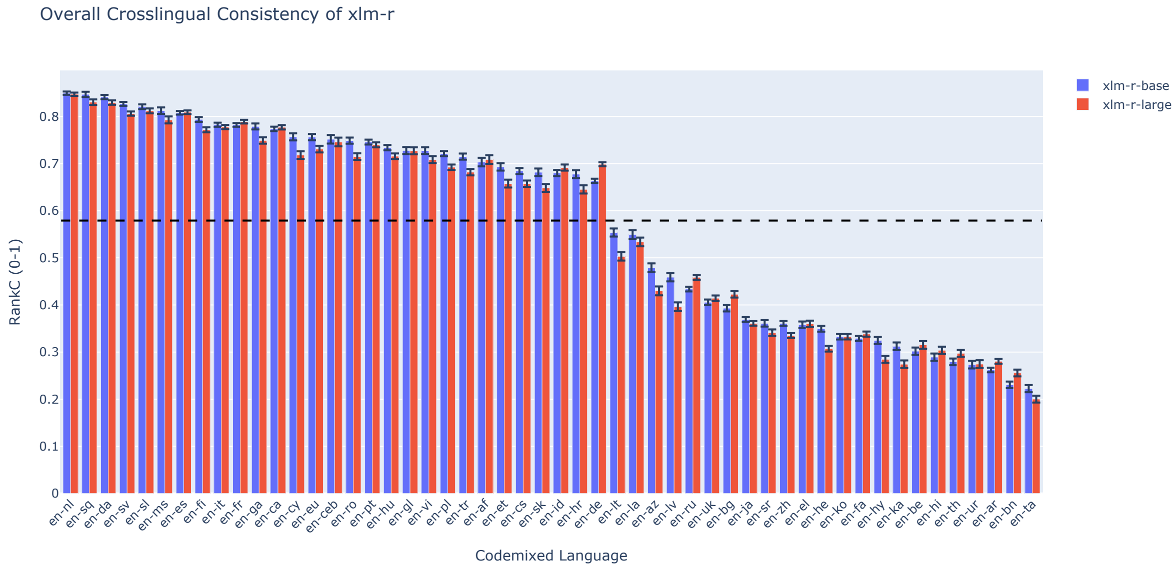


Figure 13: Cross-lingual consistency scores across languages of mt0 (top: RankC, bottom: Top@1 Accuracy). Note: The dashed line here is the average corresponding consistency scores of mt0-base across languages



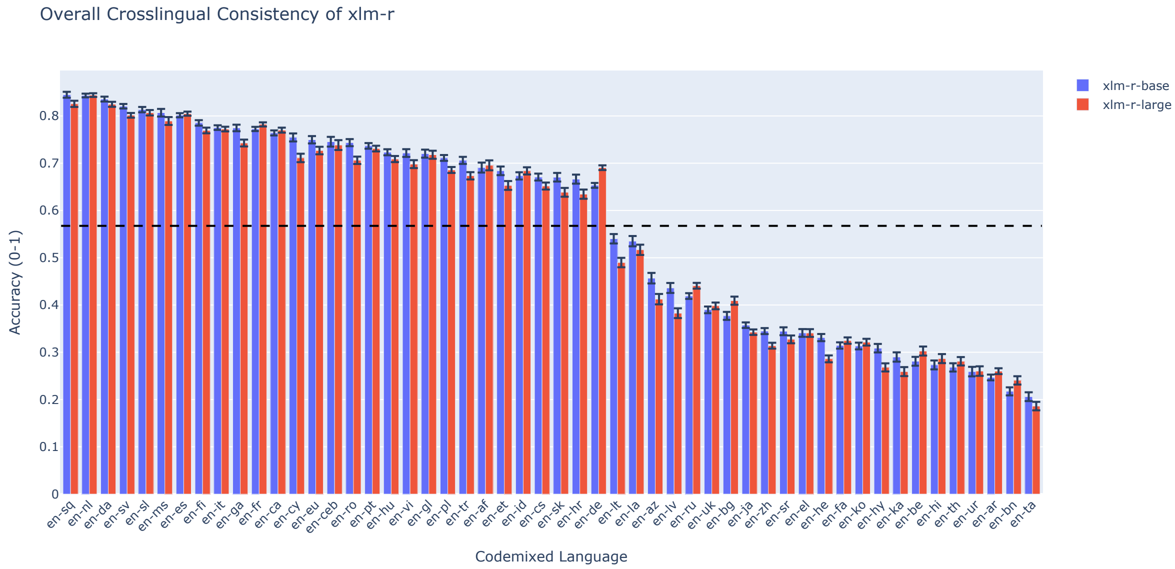
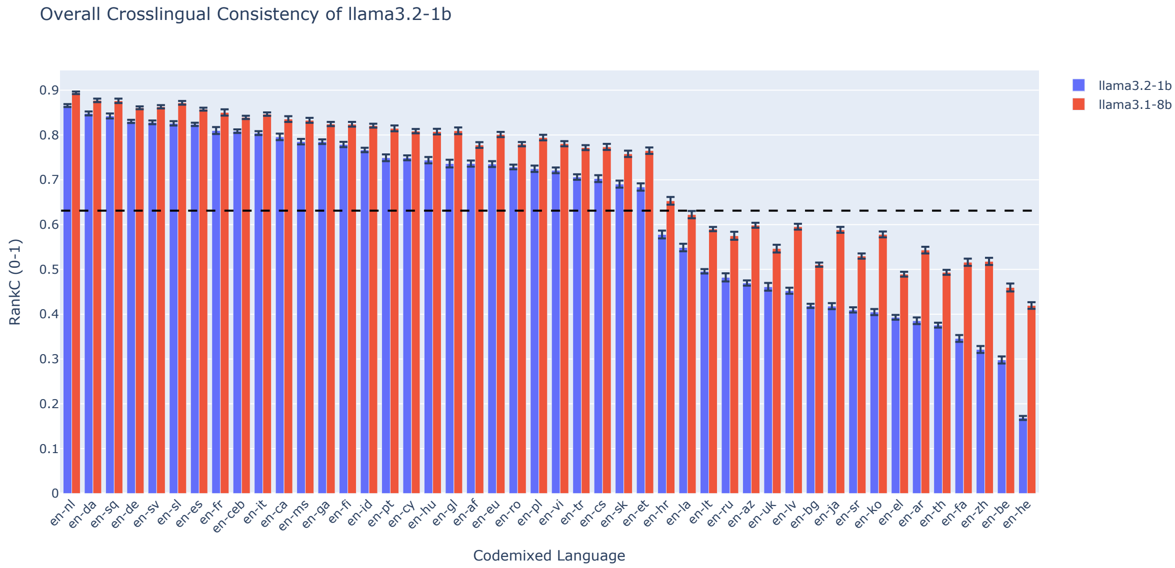


Figure 14: Cross-lingual consistency scores across languages of xlm-r (top: RankC, bottom: Top@1 Accuracy). Note: The dashed line here is the average corresponding consistency scores of xlm-r-base across languages



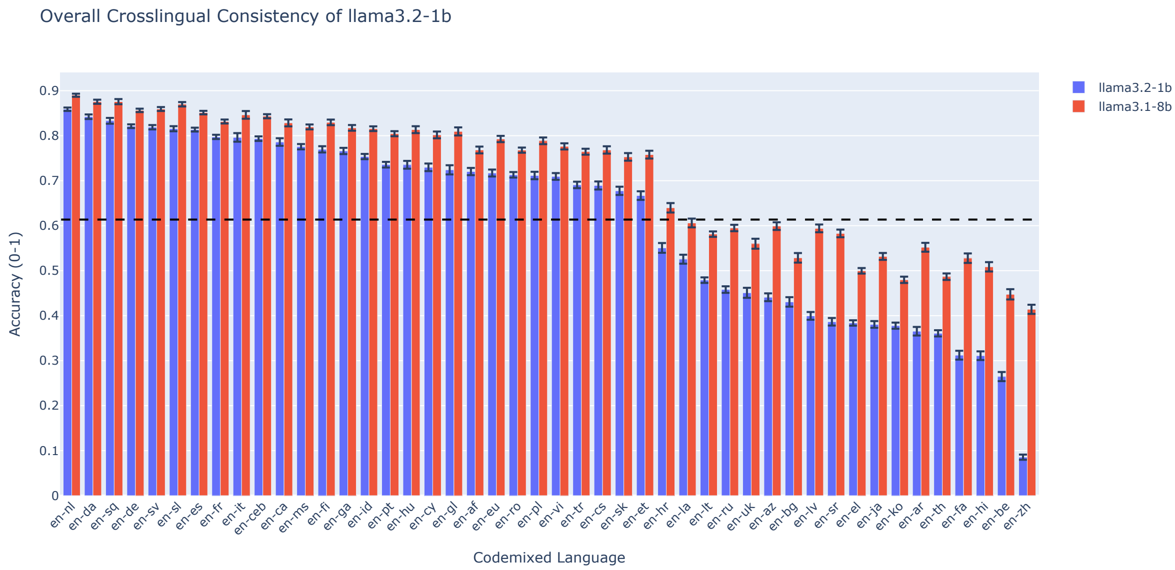


Figure 15: Cross-lingual consistency scores across languages of llama 3 (top: RankC, bottom: Top@1 Accuracy). Note: The dashed line here is the average corresponding consistency scores of llama3.2-1b across languages

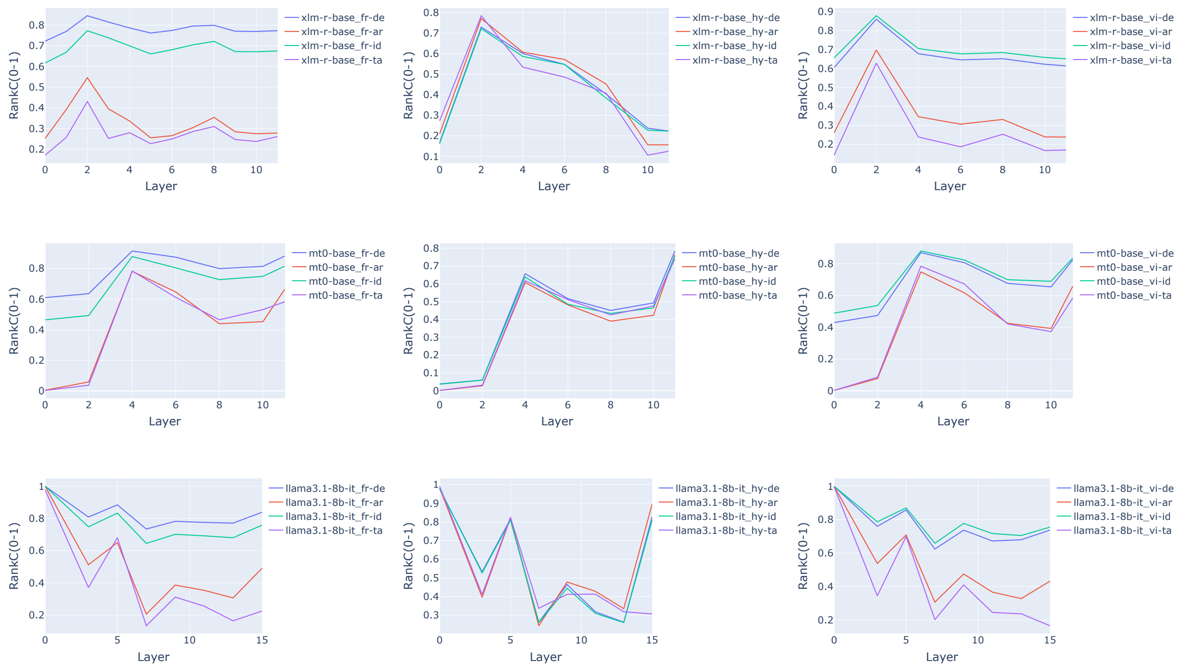


Figure 16: Layerwise cross-lingual consistency (rankC) across different transformer types (top: encoder, middle: encoder-decoder, bottom: decoder) on different non-english matrix languages (left: french, center: armenian, right: vietnamese).

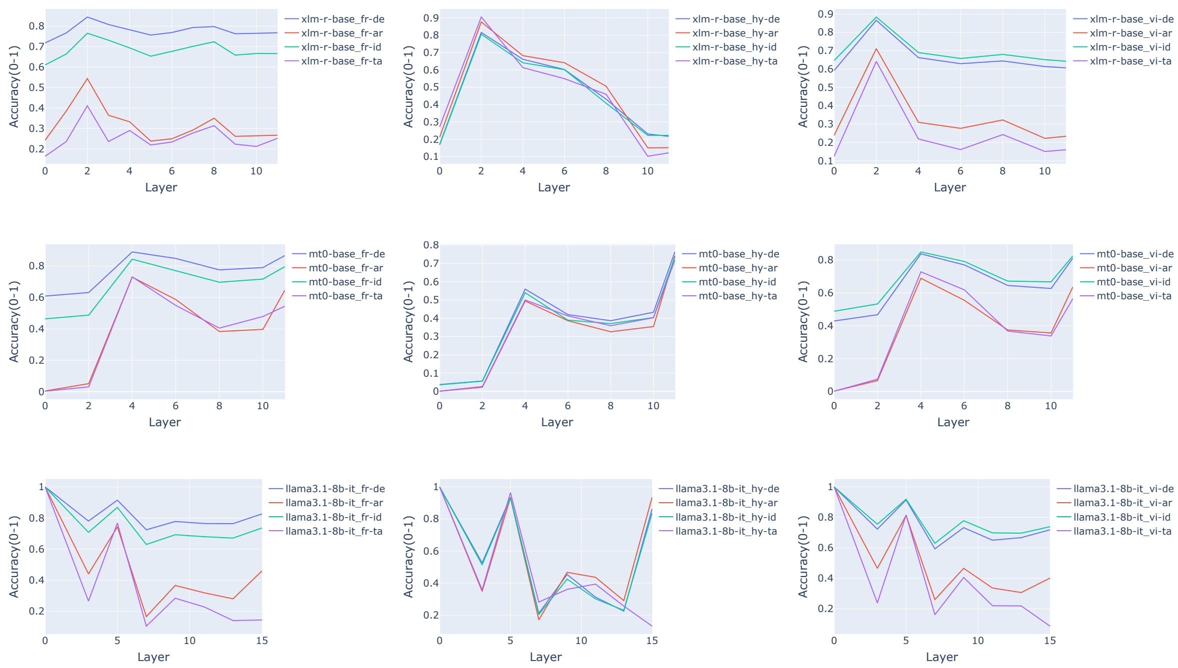


Figure 17: Layerwise cross-lingual consistency (rankC) across different transformer types (top: encoder, middle: encoder-decoder, bottom: decoder) on different non-english matrix languages (left: french, center: armenian, right: vietnamese).

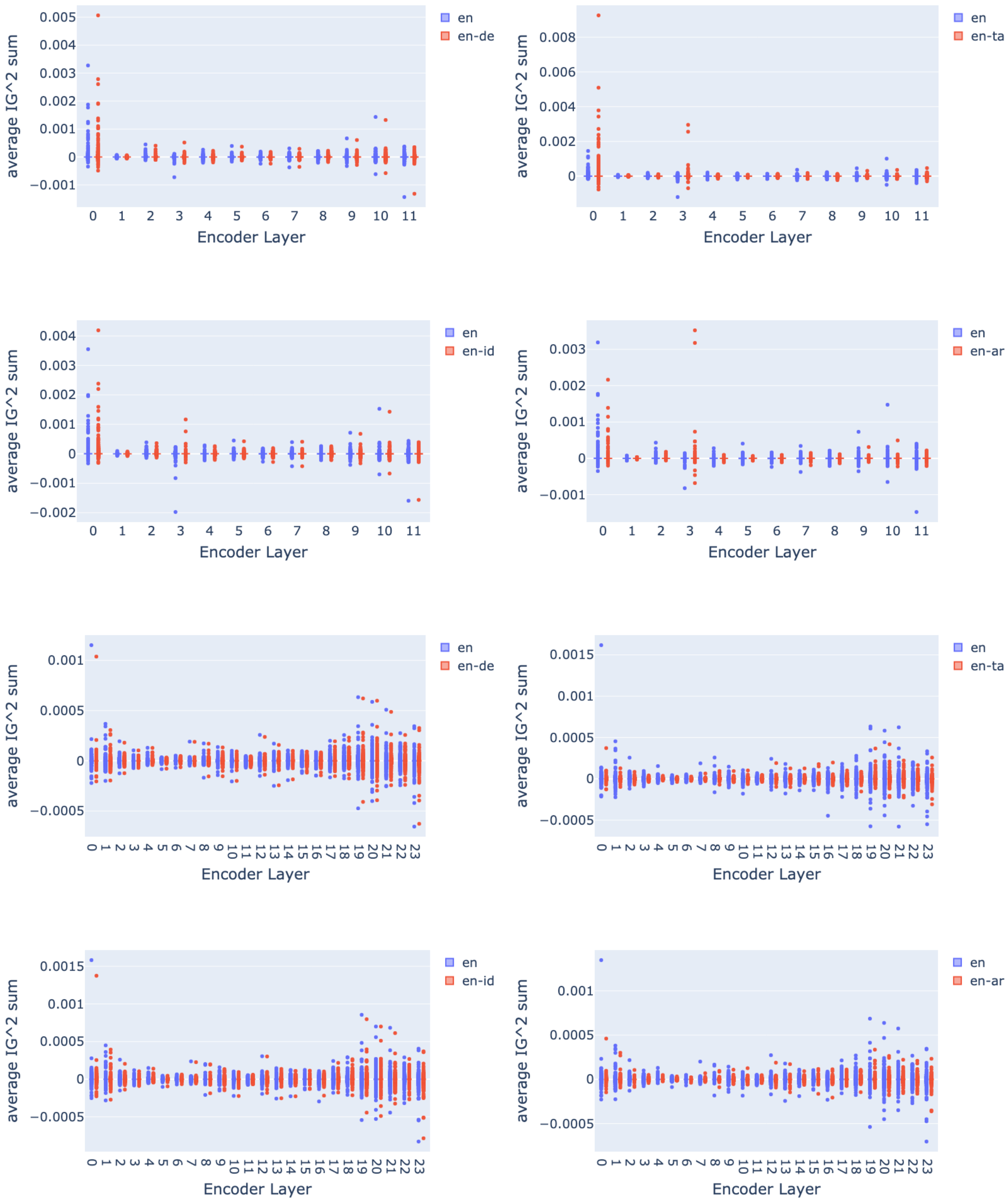


Figure 18: IG^2 scores in mt0 for en–de, en–ta, en–id, and en–ar. Models legend: upper two rows: mt0-base, lower two rows: mt0-large.

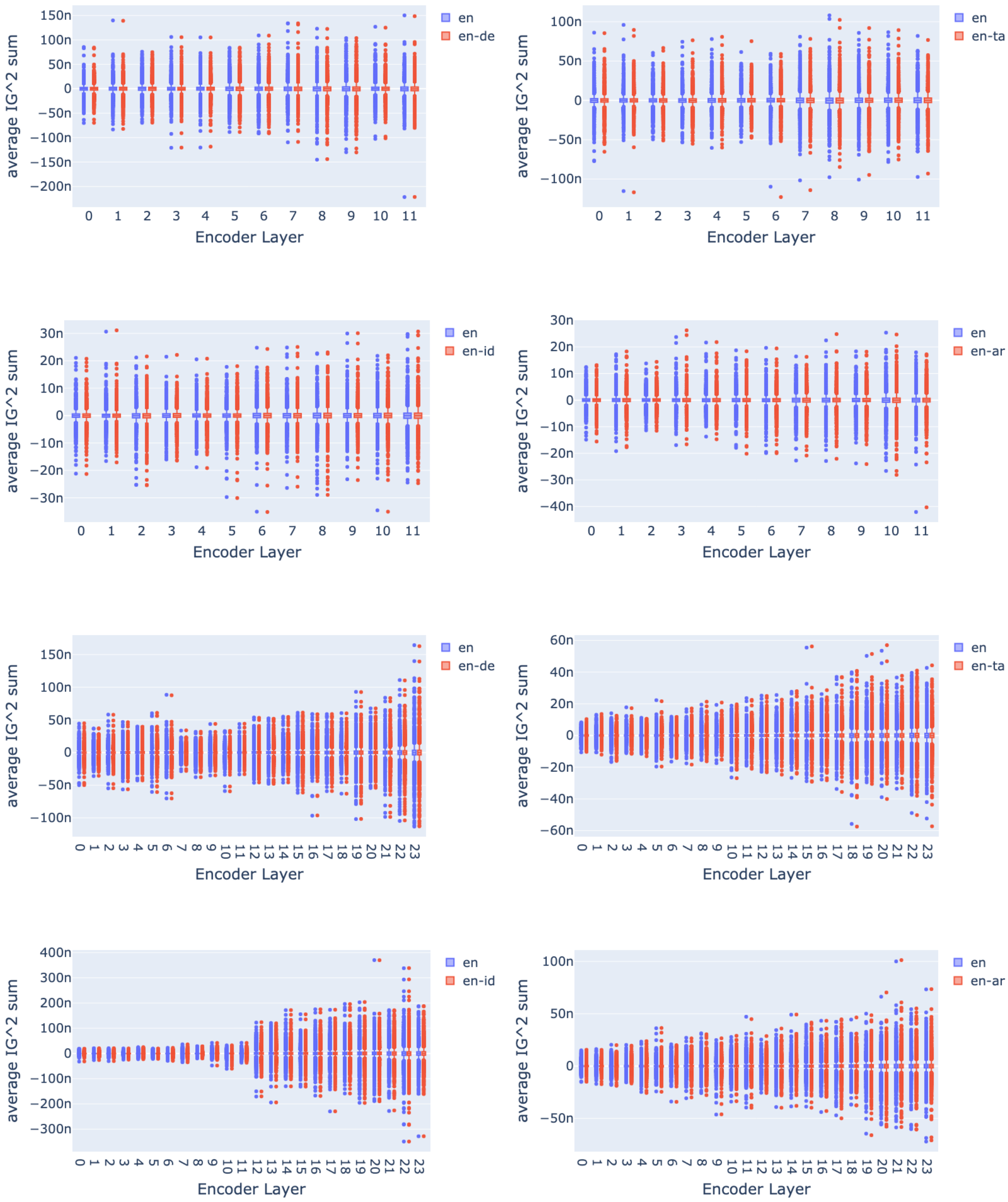


Figure 19: IG^2 scores in xlm-r for en–de, en–ta, en–id, and en–ar. Models legend: upper two rows: xlm-r-base, lower two rows: xlm-r-large

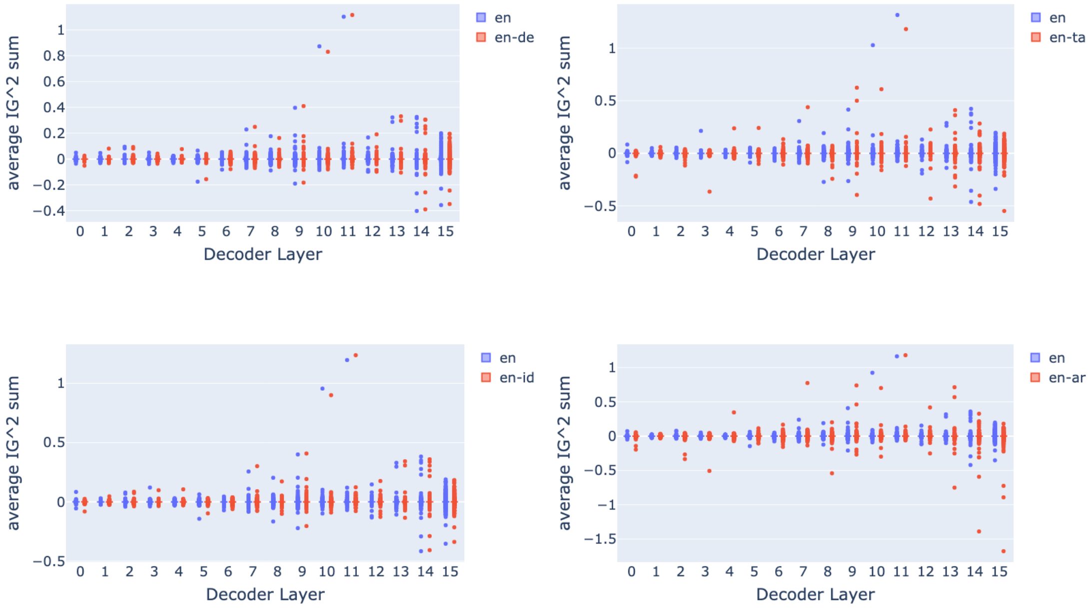


Figure 20: IG^2 scores in llama3.2-1b for en–de, en–ta, en–id, and en–ar.

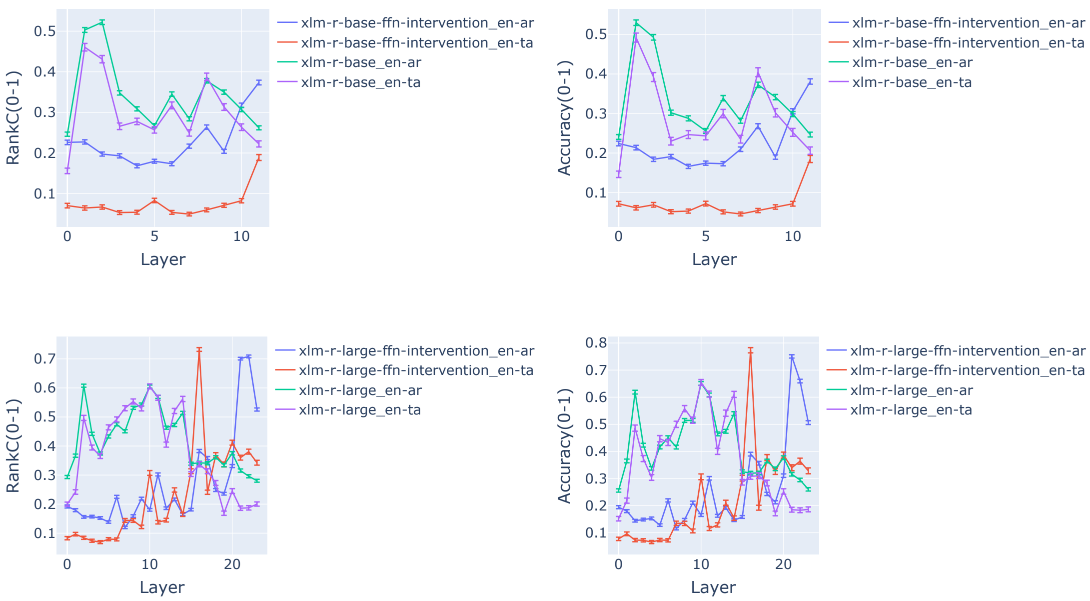


Figure 21: Intervention scores across all models. Metrics legend: left: RankC, right: Top@1 Accuracy. Model family: xlm-r

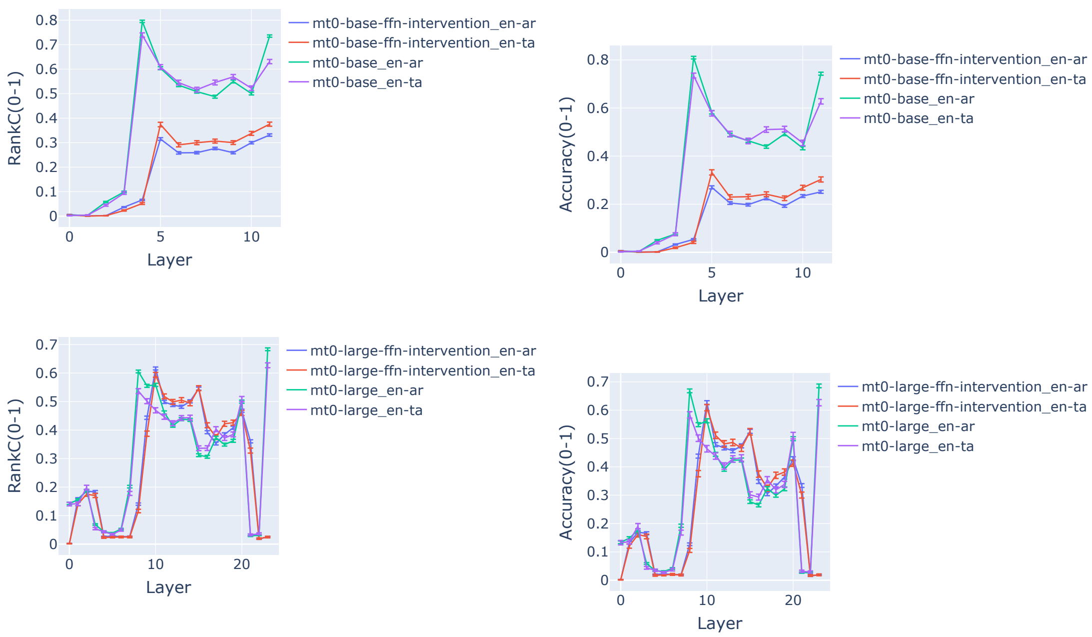


Figure 22: Intervention scores across all models. Metrics legend: left: RankC, right: Top@1 Accuracy. Model family: mt0.

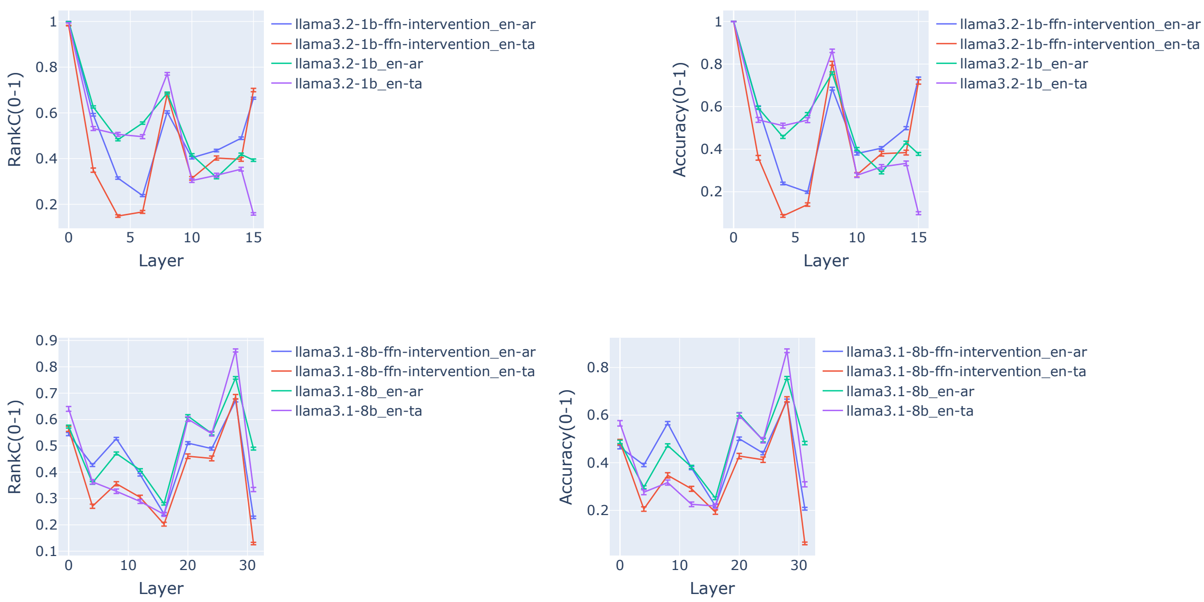


Figure 23: Intervention scores across all models. Metrics legend: left: RankC, right: Top@1 Accuracy. Model family: Llama 3.

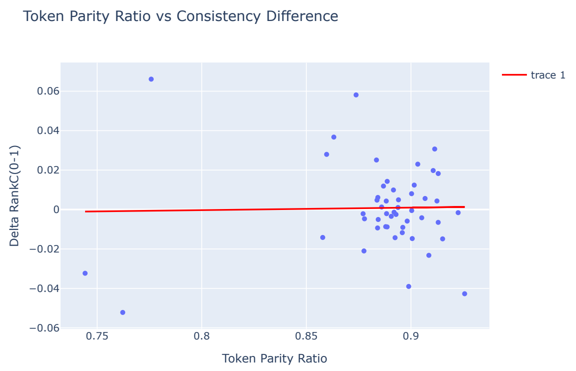


Figure 24: Regression analysis between parity ratio and RankC improvement offered by xlm-v to xlm-r. Spearman $\rho = 0.06$. We define parity ratio as the token length ratio between tokenized subjects for xlm-v-base and xlm-r-base. Our analysis discovers that many languages have a token parity ratio average within 0.8-1, which means that many of the subject entities are known on both tokenizers of the models.

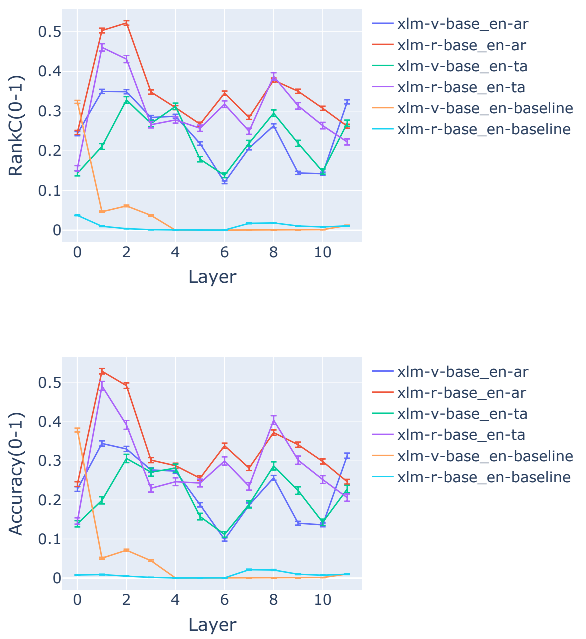


Figure 25: Layer-wise cross-lingual knowledge consistency of xlm-v vs xlm-r-base

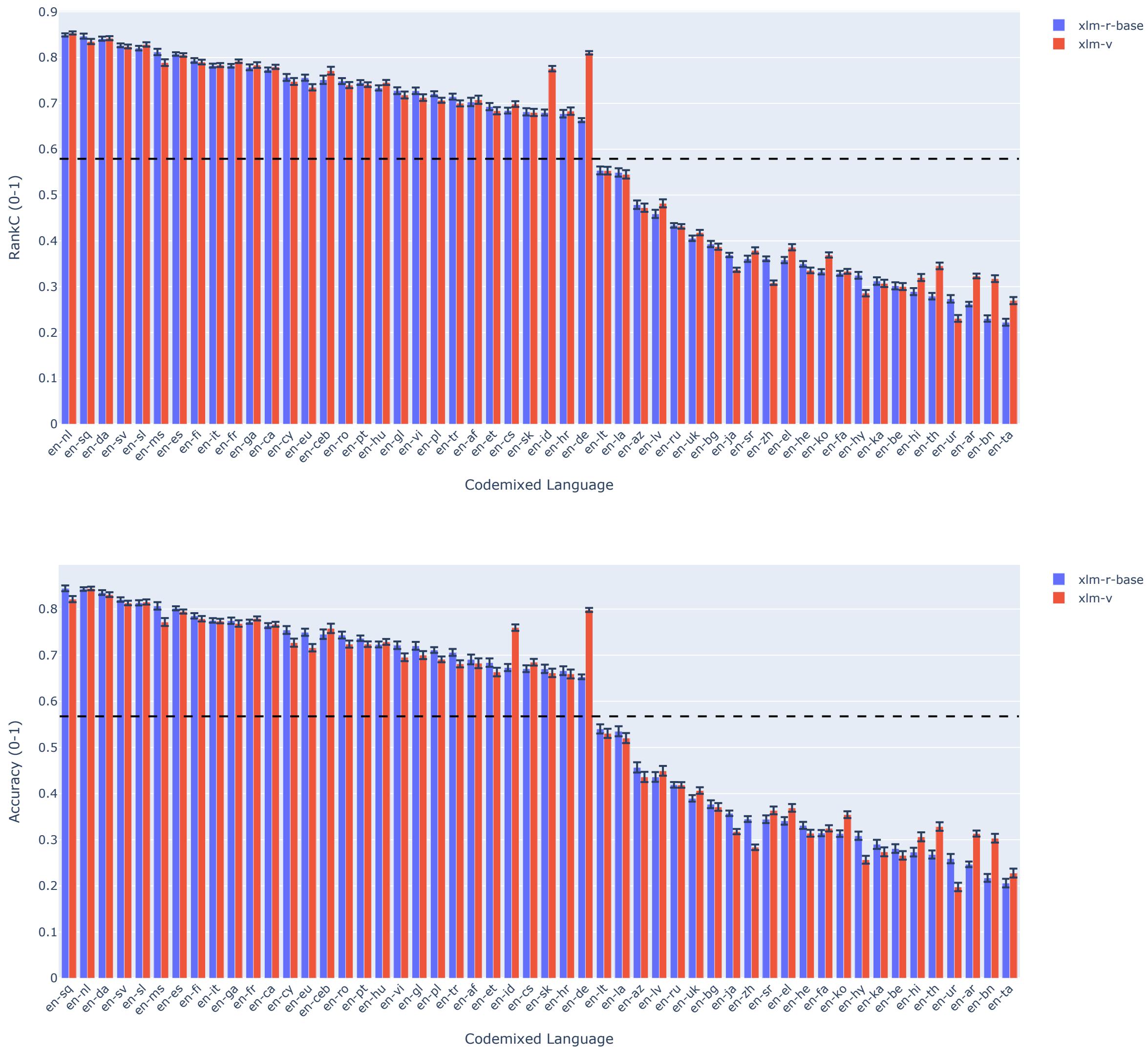


Figure 26: Effects of vocabulary expansion to overall cross-lingual consistency (top: RankC, bottom: Top@1 Accuracy). Note: The dashed line here is the average corresponding consistency scores of xlm-r-base across languages

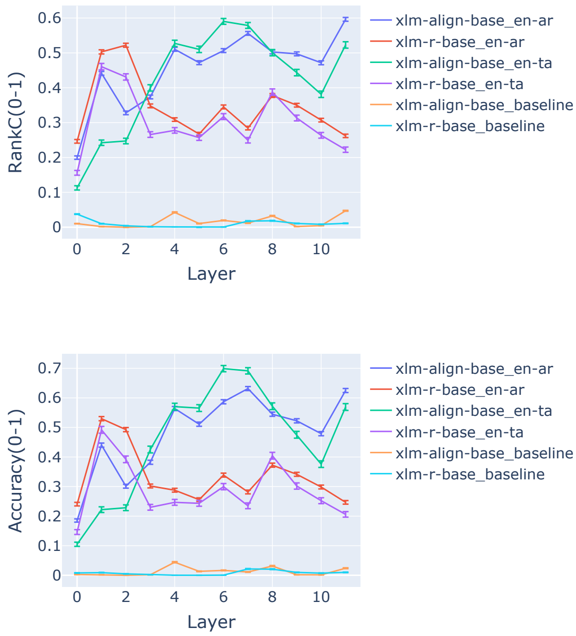
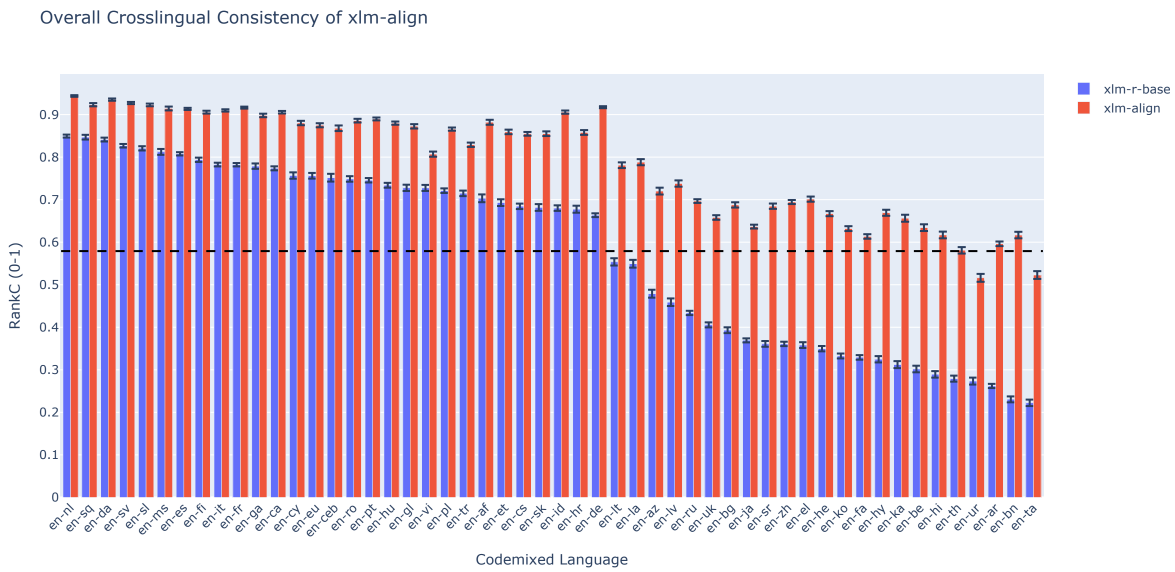


Figure 27: Effects of cross-lingual word-alignment training on the layer-wise consistency.



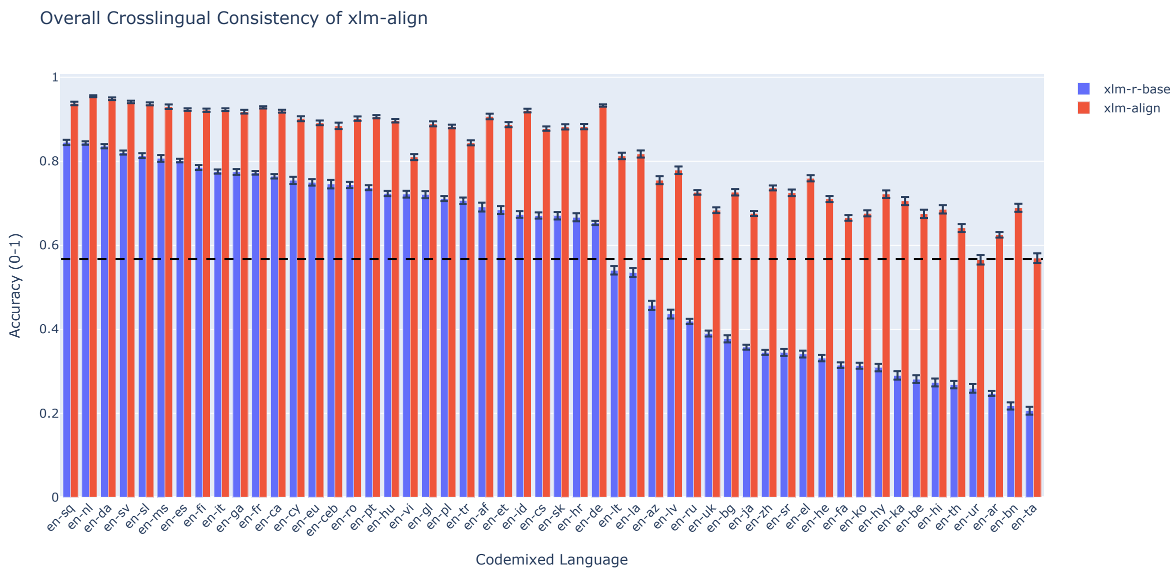


Figure 28: Effects of additional cross-lingual word alignment to overall cross-lingual consistency (top: RankC, bottom: Top@1 Accuracy). Note: The dashed line here is the average corresponding consistency scores of xlm-r-base across languages

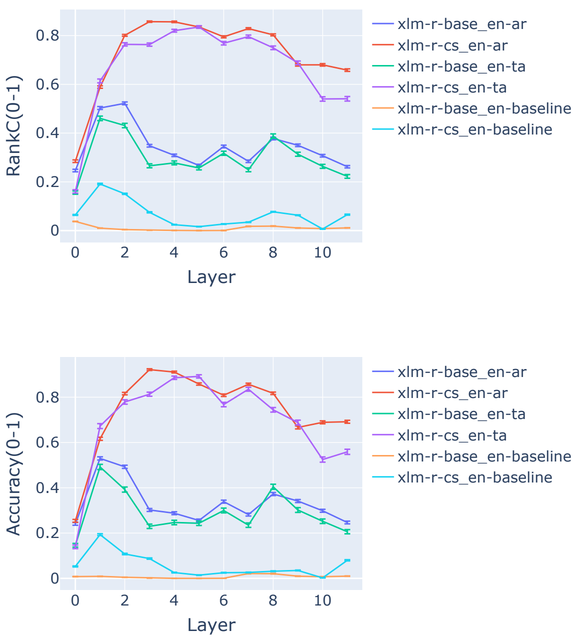


Figure 29: Effects of code-switching training to layer-wise cross-lingual consistency (top: RankC, bottom: Top@1 Accuracy).

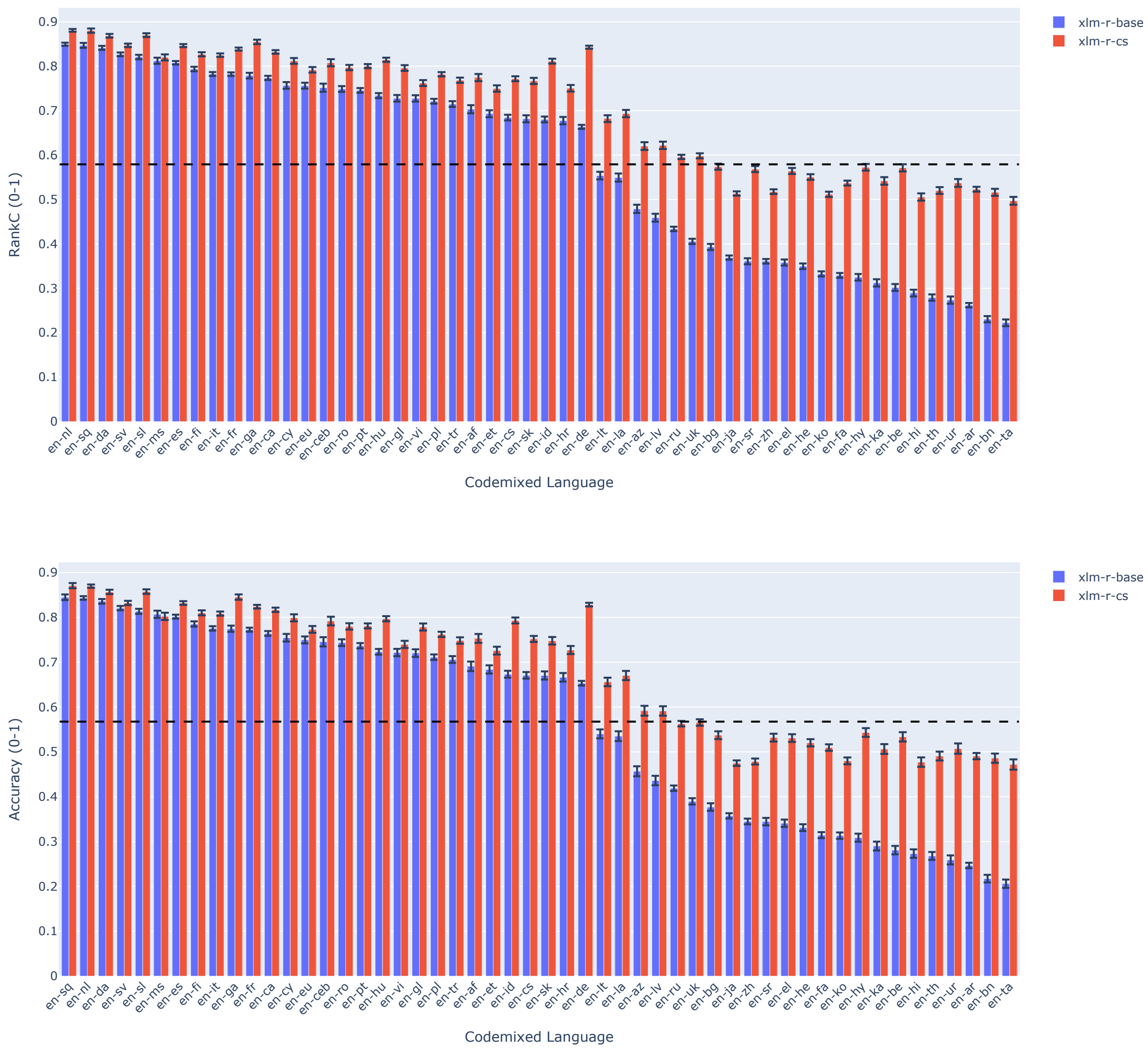


Figure 30: Effects of code-switching training to overall cross-lingual consistency (top: RankC, bottom: Top@1 Accuracy). Note: The dashed line here is the average corresponding consistency scores of xlm-r-base across languages

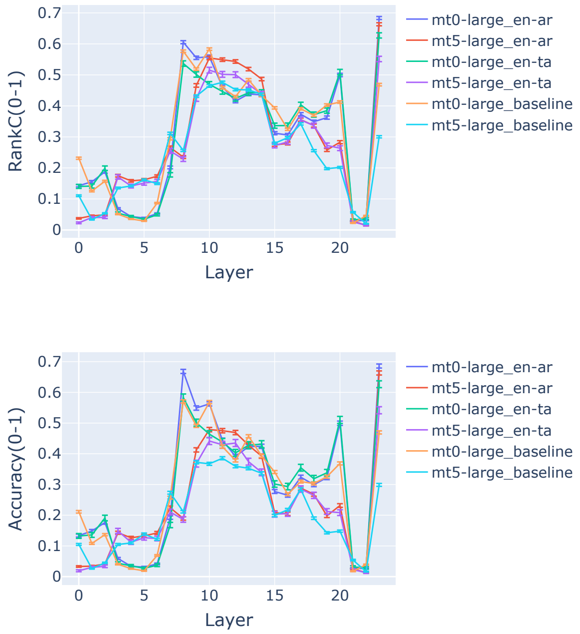


Figure 31: Effects of multi-task instruction tuning on the layer-wise consistency in encoder-decoder models.

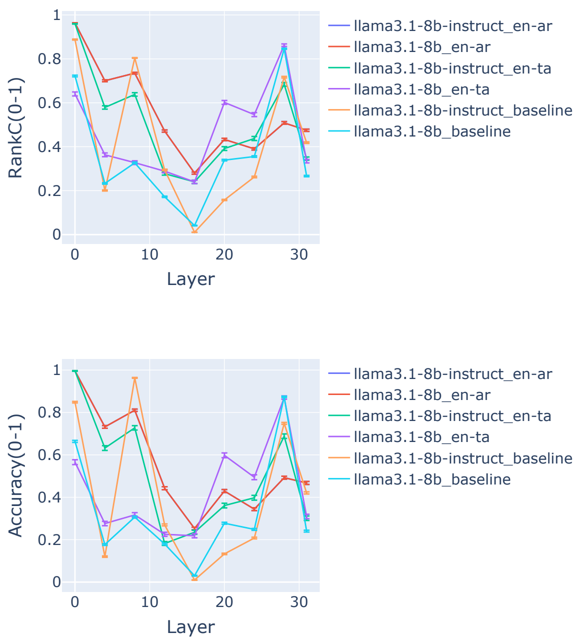


Figure 32: Effects of multi-task instruction tuning on the layer-wise consistency of decoder models.

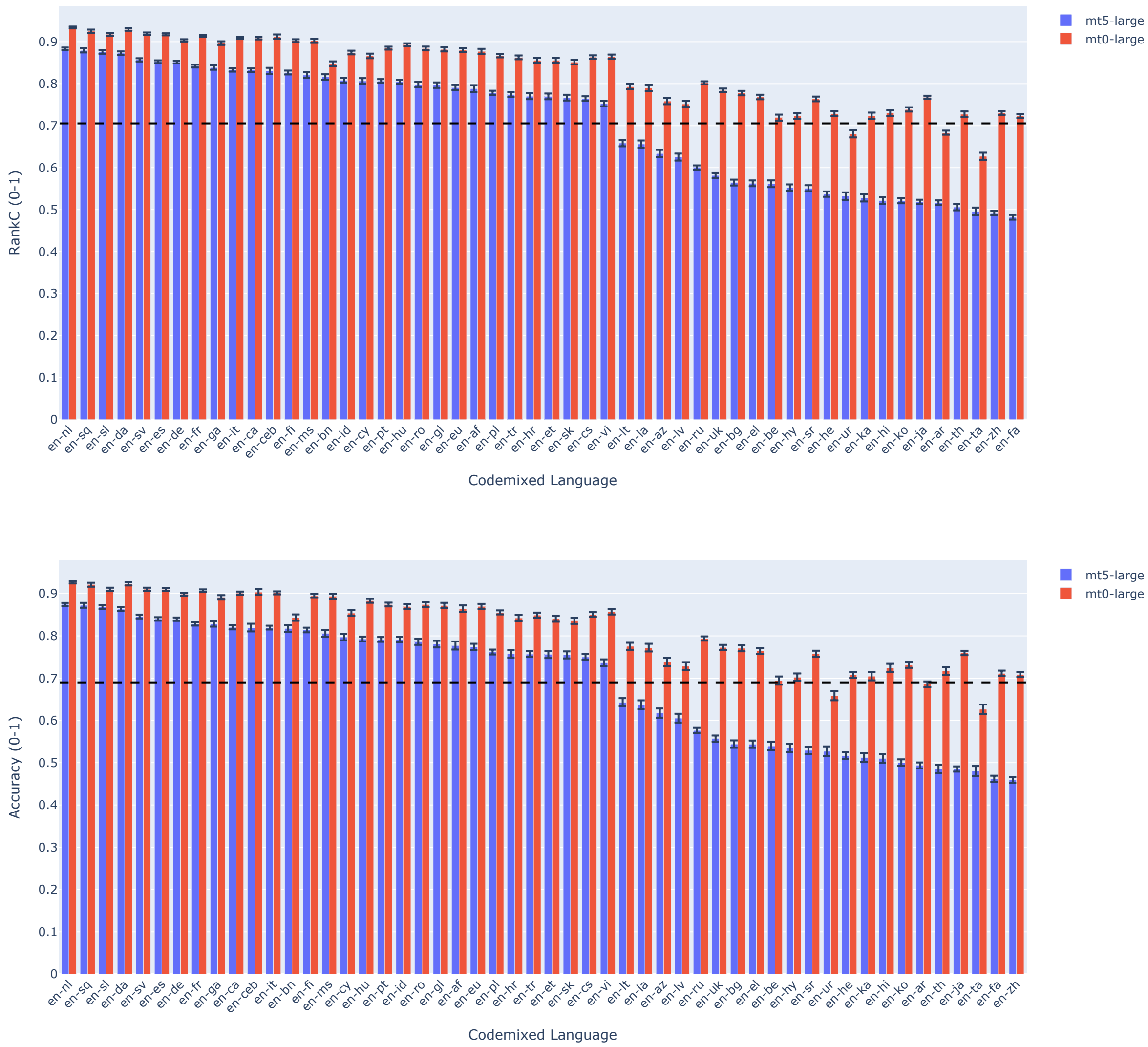


Figure 33: Effects of multi-task instruction tuning to overall cross-lingual consistency (top: RankC, bottom: Top@1 Accuracy). Note: The dashed line here is the average corresponding consistency scores of mt5-large across languages

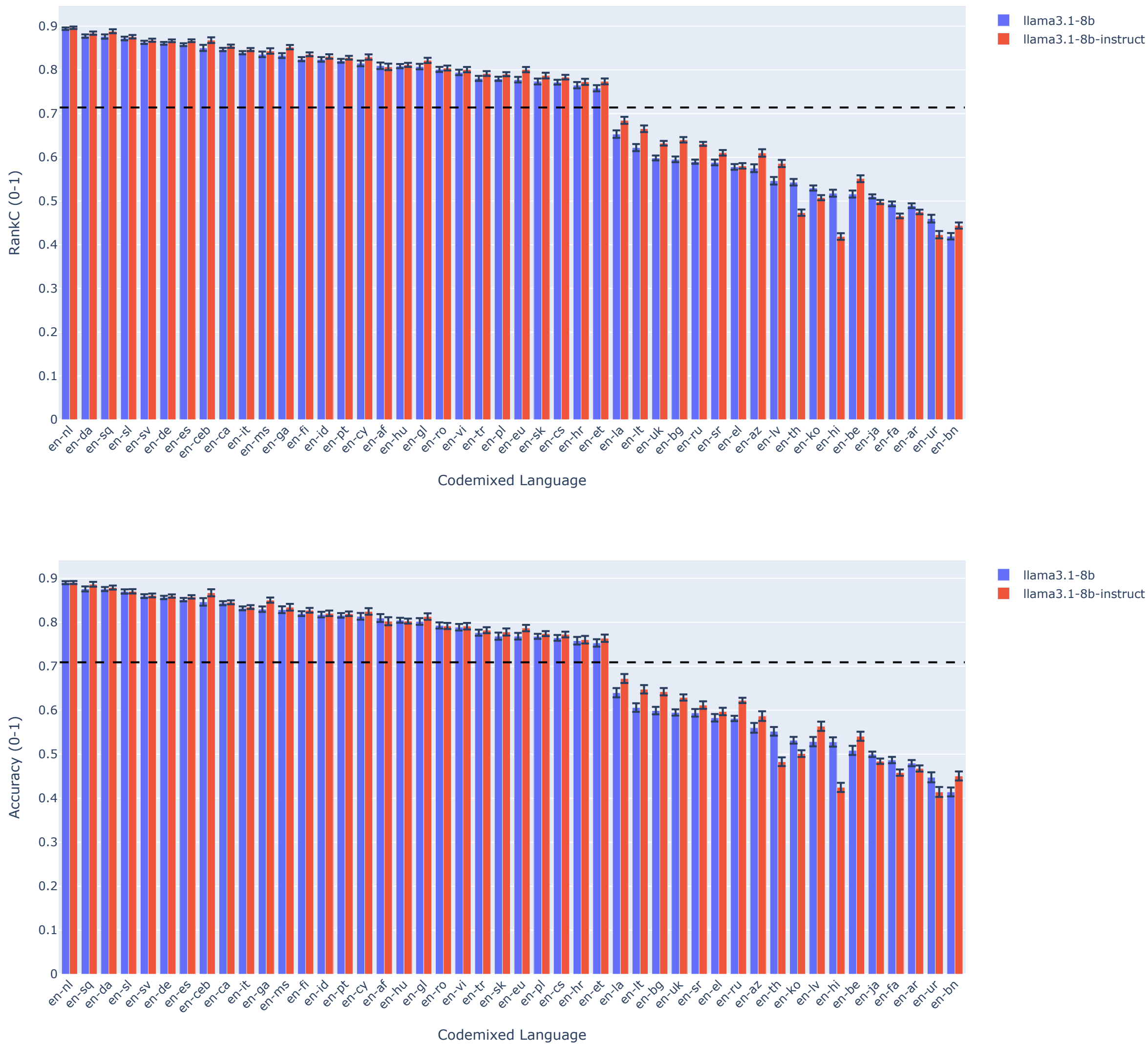


Figure 34: Effects of multi-task instruction training to overall cross-lingual consistency (top: RankC, bottom: Top@1 Accuracy). Note: The dashed line here is the average corresponding consistency scores of llama3.1-8b across languages

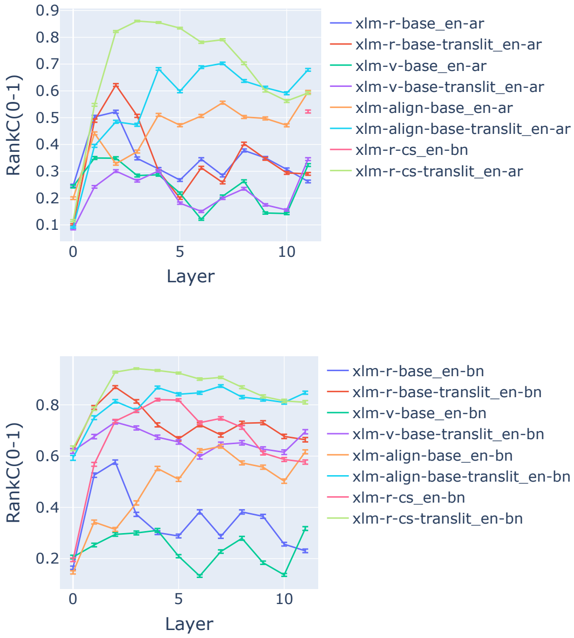


Figure 35: Impact of Transliterations.