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

Xi Ai*, Mahardika Krisna Ihsani*, Min-Yen Kan

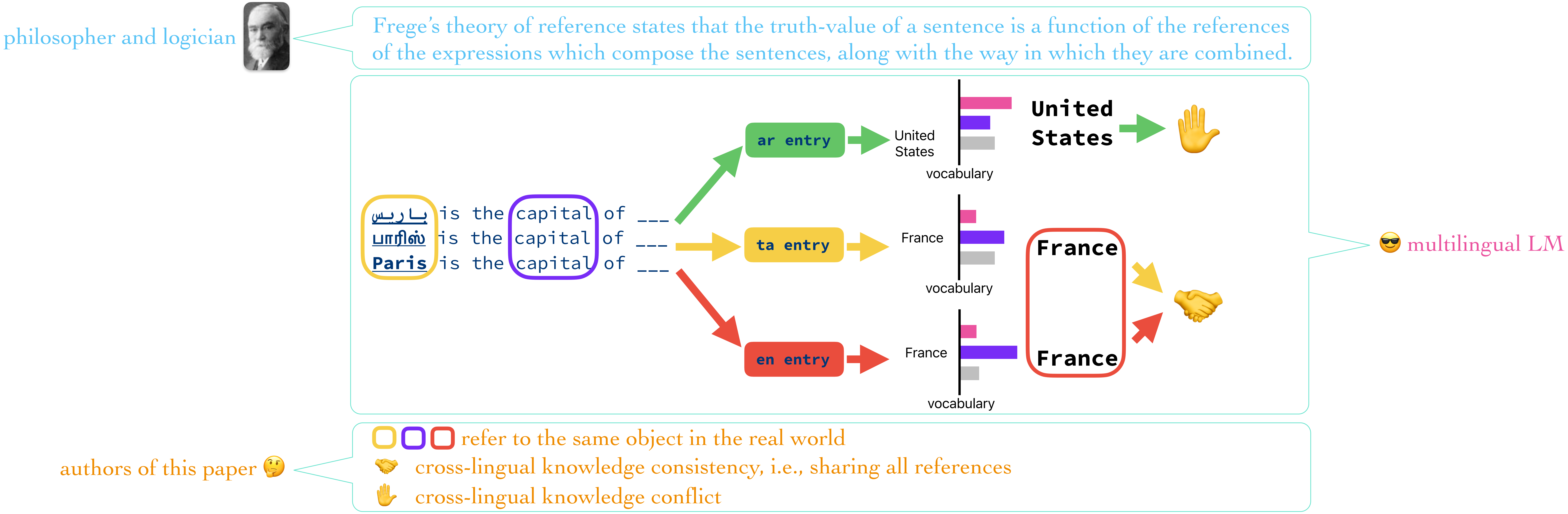
NUS Computing
National University of Singapore

WING
Web Information Retrieval / Natural Language Processing Group

Mohamed bin Zayed University of Artificial Intelligence

EMNLP 2025
Suzhou, China | 中国苏州

Language Diversity and Knowledge Consistency

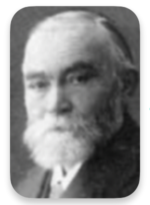


A salient aspect of humanity is that, while people may speak different languages, they can share common references and knowledge. We are thus interested in analyzing, evaluating, and interpreting cross-lingual consistency for factual knowledge.

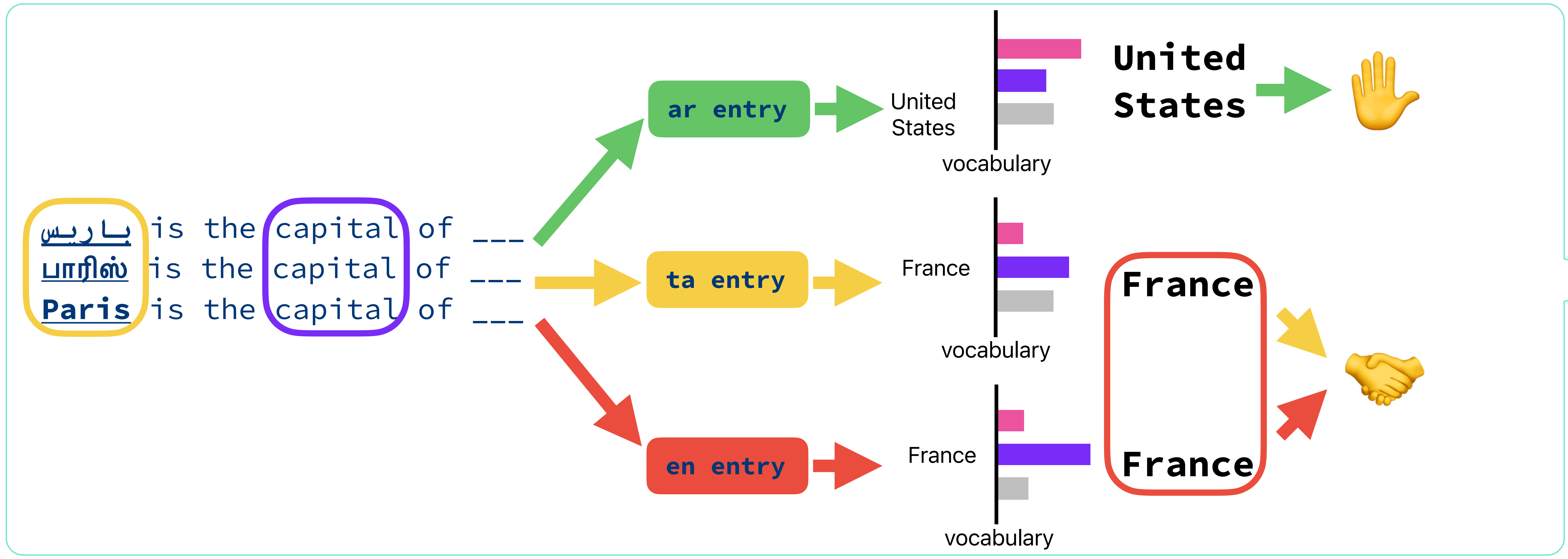
In this work, we measure consistency evolution by quantifying the distribution difference between two code-mixed, coreferential statements for each layer's output.

Method

Measure consistency evolution by quantifying the distribution difference between two code-mixed, coreferential statements for each layer's output.

philosopher and logician 

Frege's theory of reference states that the truth-value of a sentence is a function of the references of the expressions which compose the sentences, along with the way in which they are combined.



😎 multilingual LM

authors of this paper 🤔

- ▢▢▢ refer to the same object in the real world
- 🤝 cross-lingual knowledge consistency, i.e., sharing all references
- ✋ cross-lingual knowledge conflict

•

Setup

Model

- Encoder models (xlm-r from 0.3B to 10B)
- Encoder-Decoder models (mT0 from 0.6B to 3.7B, mT5 from 0.6B to 3.7B)
- Decoder models (Llama3-instruct 1B \& 8B)

Dataset

- **mLAMA** 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.**

Consistency Metrics

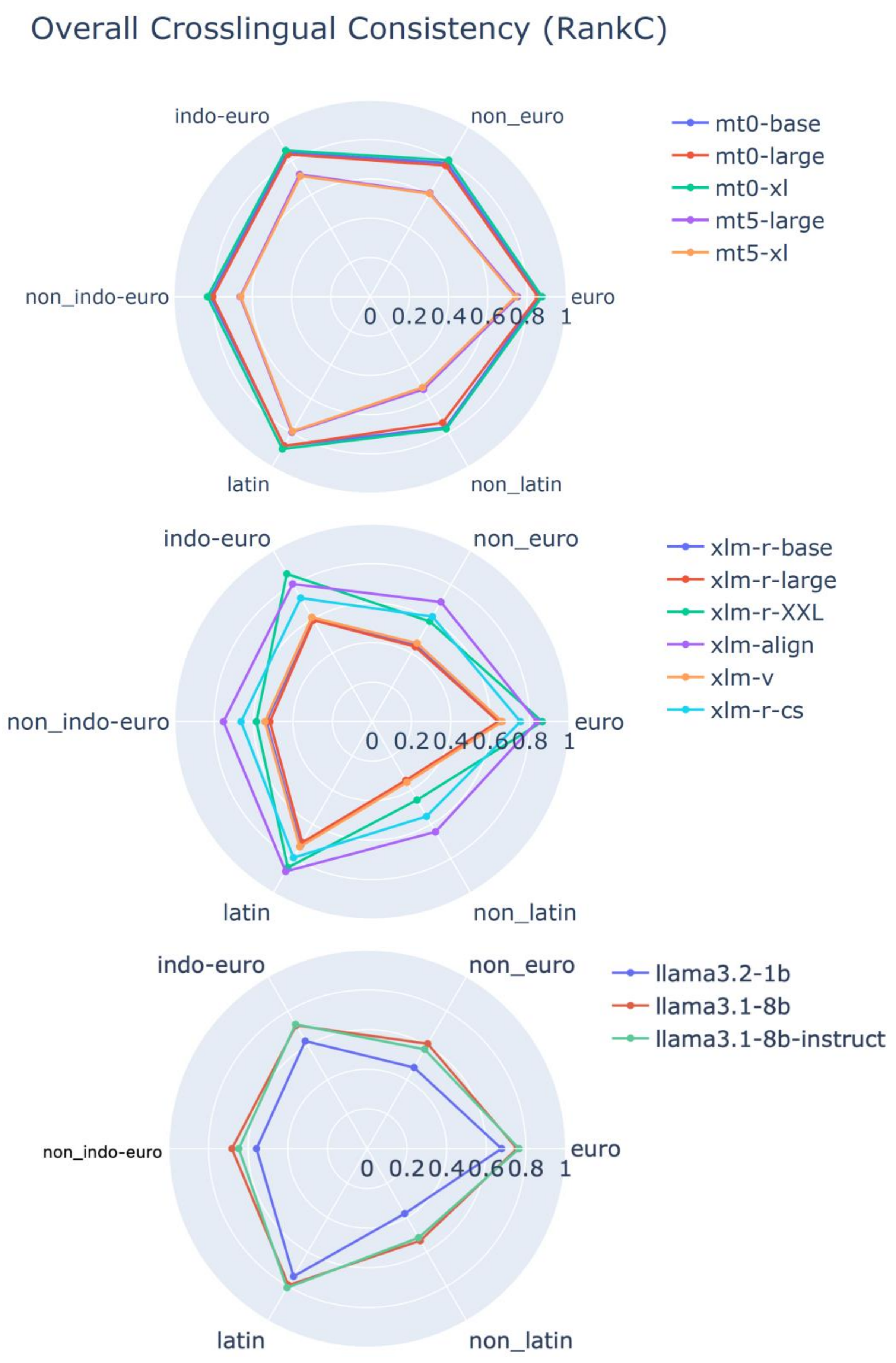
- Logit Lens (multiply representations by embeddings for distributions over the vocabulary)
  - RankC without output domain (Weighted Top-5 distritbuion)
  - Accuracy (Top-1)
- CKA similarity

Findings 1/5

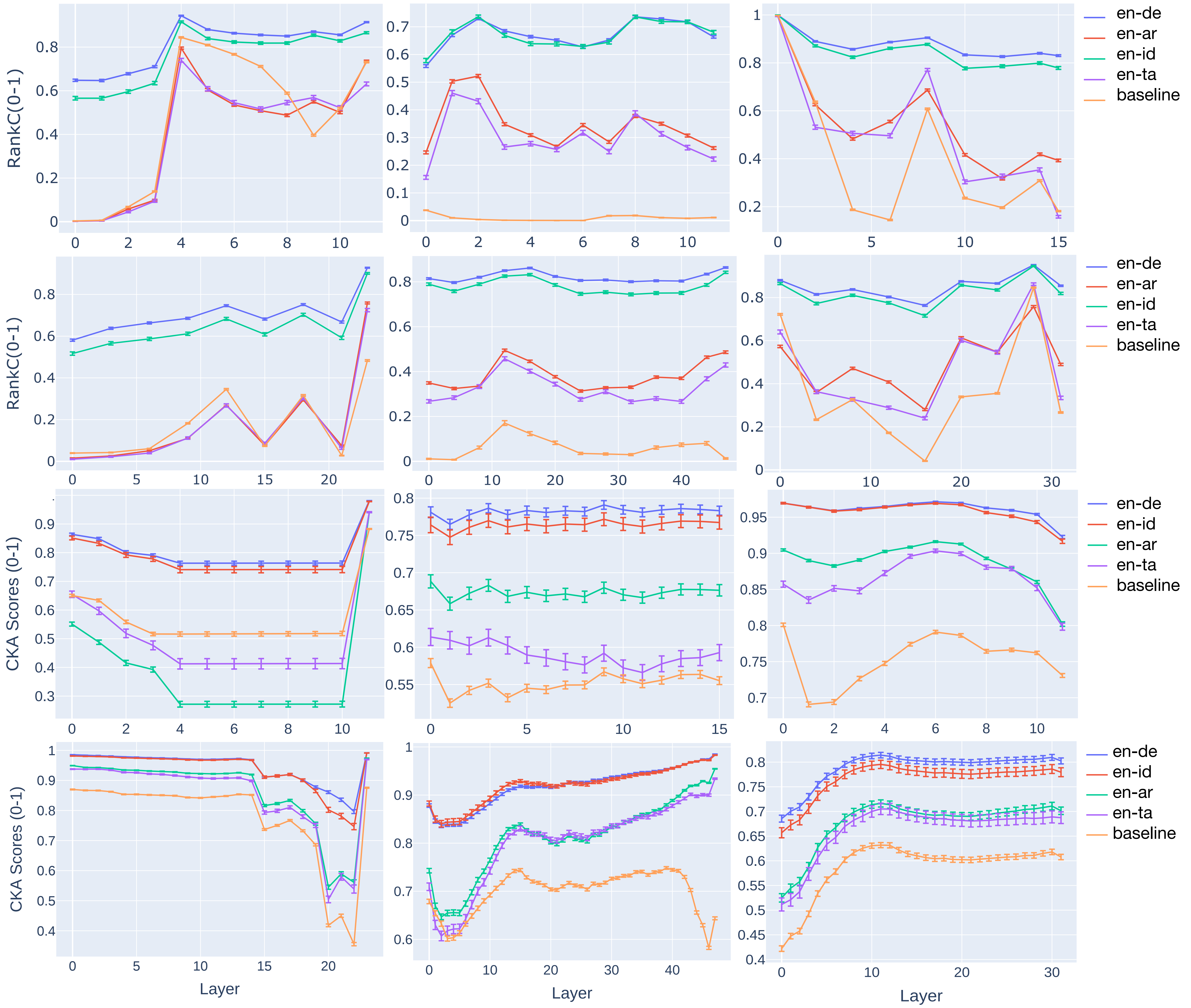
Consistency bottlenecks and issues tied to language characteristics, scripts, and training biases through layer-wise analyses and interpretability approaches, which potentially prevent cross-lingual consistency improvements and gains from scaling.

Method

We compute the consistency metrics for each layer's representation and organize the results from the final output into 6 bins.



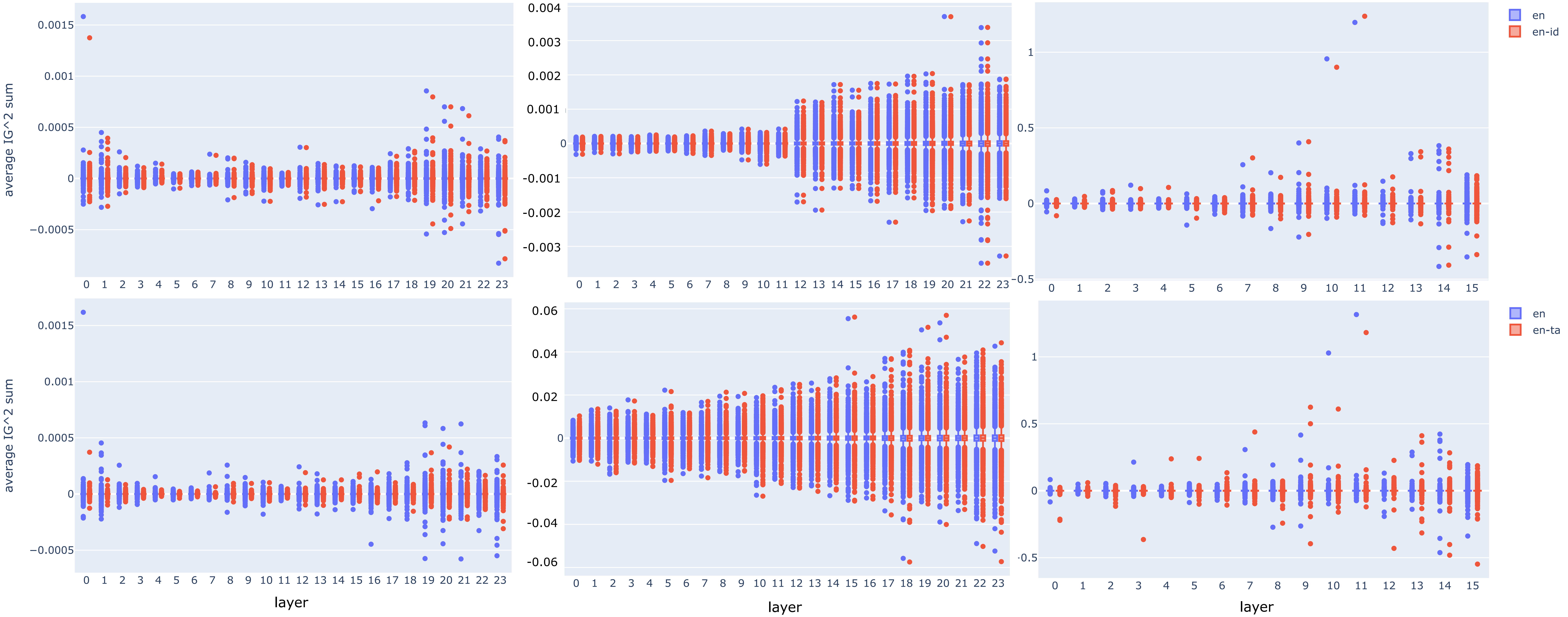
Consistency vs Language characteristics and scripts



Consistency evolution over layers(L: Encoder, M: Encoder-Decoder,R: Decoder)

Findings 2/5

Consistency moderately correlates with sharing feed-forward neurons at all the layers statistically.

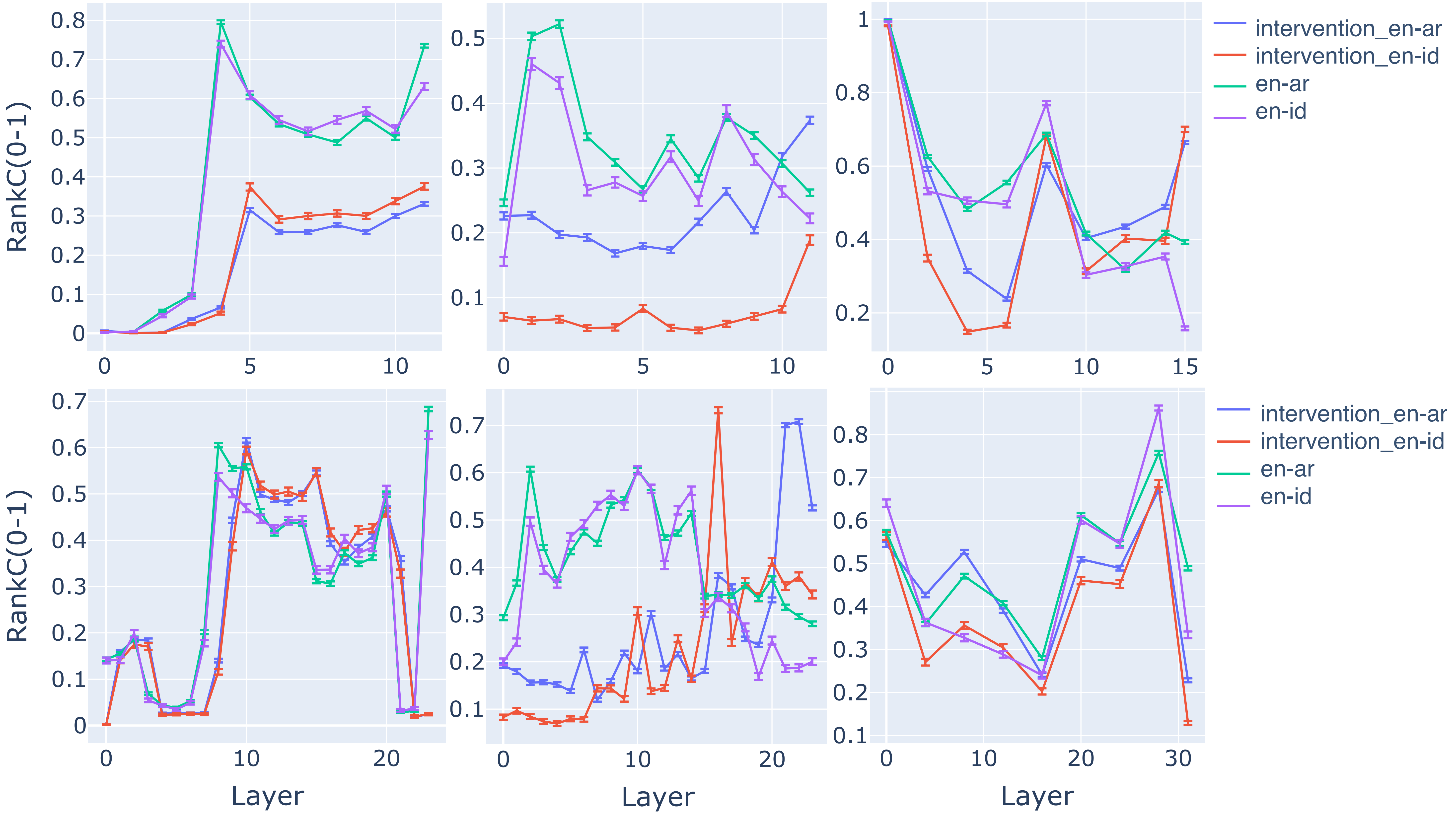


Method

We take the activated neurons (IG2 scores) into account.

Findings 3/5

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.



Method

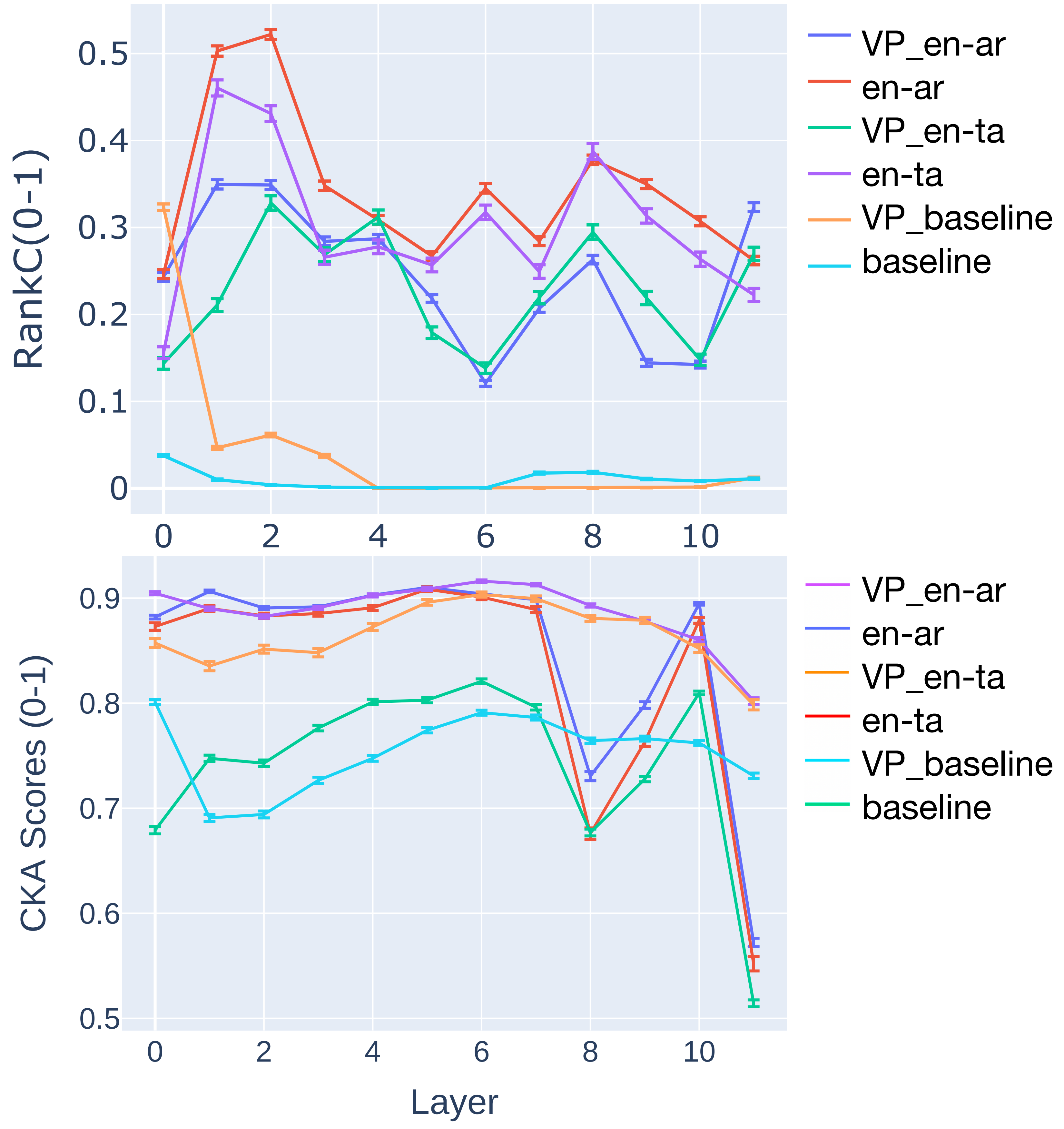
We patch English-biased activation to other languages.

Findings 4/5

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.

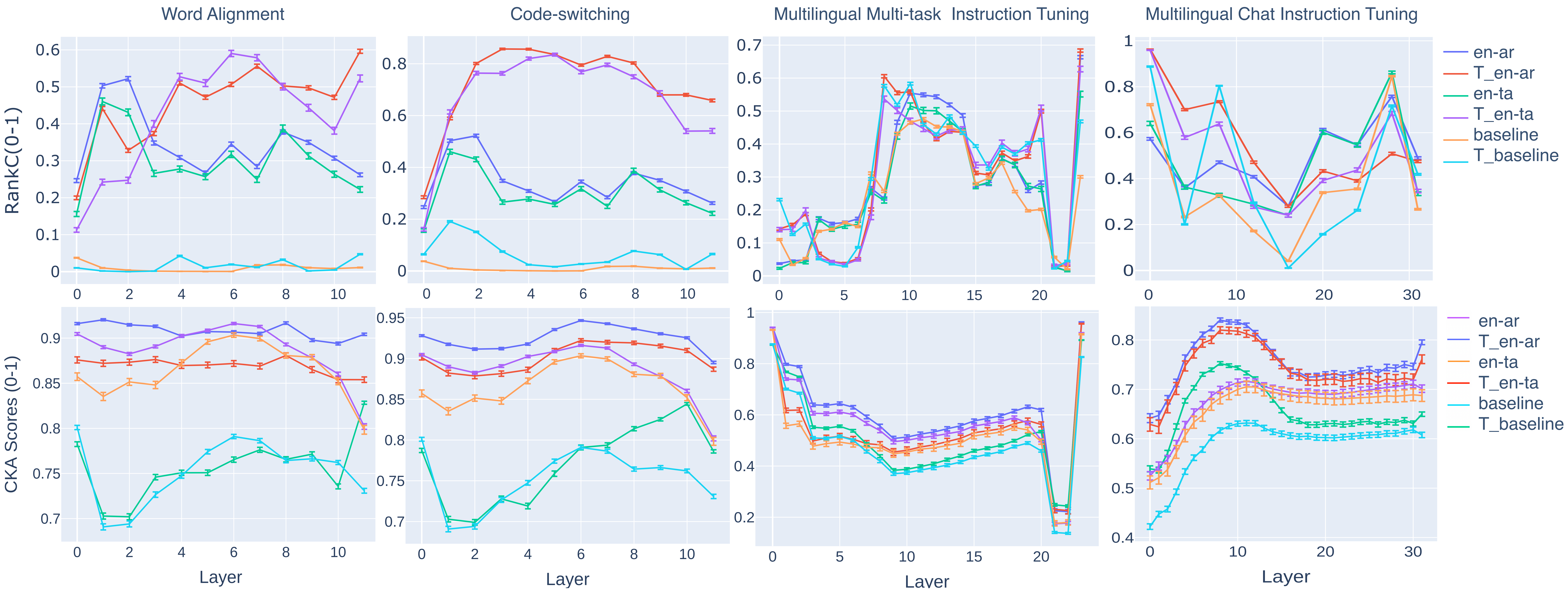
Method

We compare a base model with a vocabulary-expended model (VP-x) to isolate the vocabulary factor or the shared-script factor.



Findings 5/5

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.



Method

We examine tuned models that show promising performance in multilinguality.

Conclusion

Multilingual models uncovers a consistency bottleneck whereby the consistency does not grow monotonically on each layer.

Key Factors

- model architectures
- training strategies
- deep semantic alignments

Promising directions

- test-time calibration
- training with cross-lingual alignment objectives

Necessary but not sufficient conditions

- Cross-lingual representations
- shared scripts
- parallel samples