# What if I ask in *alia lingua*? Measuring Functional Similarity Across Languages

**Debangan Mishra**[*1] **Arihant Rastogi**[*1] **Agyeya Negi**[1]
**Shashwat Goel**[2] **Ponnurangam Kumaraguru**[1]
[1]IIIT Hyderabad [2]ELLIS Institute Tübingen

## Abstract

How similar are model outputs across languages? In this work, we study this question using a recently proposed model similarity metric—$\kappa_p$—applied to 20 languages and 47 subjects in GlobalMMLU. Our analysis reveals that a model's responses become increasingly consistent across languages as its size and capability grow. Interestingly, models exhibit greater cross-lingual consistency within themselves than agreement with other models prompted in the same language. These results highlight not only the value of $\kappa_p$ as a practical tool for evaluating multilingual reliability, but also its potential to guide the development of more consistent multilingual systems.

## 1 Introduction

Users interact with large language models (LLMs) in a variety of languages across families and resource availabilities (Nicholas and Bhatia, 2023). As such, there is a need for LLMs to perform well across languages. These models should provide consistent responses—if switching languages results in incorrect answers to the same question, it could potentially mislead users, especially in critical areas like medical advice or legal interpretation. However, current evaluations primarily focus on per-language accuracy, with little attention to consistency across languages (Koto et al., 2024; Romanou et al., 2024; Singh et al., 2024).

To quantify this consistency, we study the functional similarity of model outputs. We use Chance Adjusted Probabilistic Agreement (CAPA or $\kappa_p$), a metric recently proposed by (Goel et al., 2025), which incorporates model accuracy on a given benchmark. We extend it to measure how similar the mistakes are across different languages, giving a view of multilingual functional similarity.

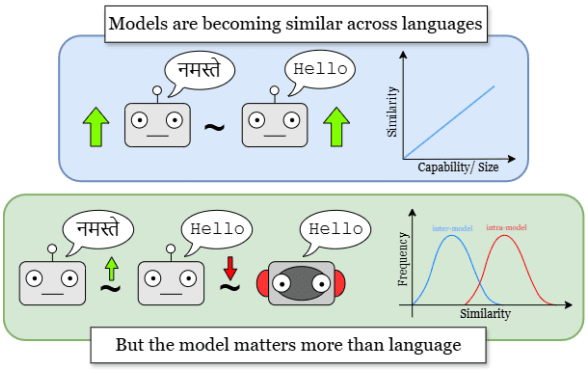



Figure 1: **Our Main Findings:** We use functional similarity to measure the consistency of model outputs across different languages. We find: (1) as language models get bigger and more capable, their outputs become more similar across languages; (2) models tend to be more self-consistent across languages than when comparing different models in a common language.

We use GlobalMMLU (Singh et al., 2024) - a carefully translated version of MMLU across multiple languages - as our benchmark. It tests the factual QA capabilities of models across a variety of subjects, ranging from mathematics to philosophy, in a multiple-choice format. Our choice of this benchmark is motivated by its parallel nature, which allows us to test whether models behave consistently across languages on factual tasks.

Our study encompasses two dimensions of functional similarity: intra-model (consistency across languages for a given model) and inter-model (consistency across models for a given language). When considering intra-model similarity, we find that with increasing size and accuracy, models are becoming more functionally similar across languages. Notably, we observe that all models are more consistent with themselves across languages than they are with other LLMs for the same language, indicating that intra-model similarity exceeds inter-model similarity for our task. Interestingly, multilingual similarity further varies by

* *These authors contributed equally*
Please find our code here: GitHub



*Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 496–506
November 8-9, 2025 

domain and resource levels of the languages.

Primarily, we show that $\kappa_p$, a chance-adjusted functional similarity metric, provides a powerful lens for analyzing multilingual consistency of LLMs. We explore cross-lingual patterns that accuracy and representational similarity alone cannot capture, by combining the output behavior and performance of the LLM. We find interesting patterns about multilingual model behavior, including effects of scale, domain, and resources.

## 2 Related Work

**Similarity Metrics:** Prior work on model similarity falls broadly into two classes: *representational similarity* and *functional similarity*. Representational similarity metrics (Huh et al., 2024; Klabunde et al., 2025) focus on the internal states of models such as weights and activations, whereas functional similarity metrics (Goel et al., 2025) evaluate models based on their input–output behavior, making them applicable across architectures. Importantly, functional similarity better reflects the user experience, since what ultimately matters is whether models behave consistently across inputs, rather than how their internal representations align.

**Multilingual Evaluations:** In representational studies, researchers have identified language-specific neurons (Tang et al., 2024) and language-agnostic "semantic hubs" (Wu et al., 2024), and even used steering interventions to demonstrate their causal effects. While such work sheds light on cross-lingual representations, it does not establish quantitative trends in cross-lingual *output* consistency as models scale. On the functional side, prior work on multilingual factual consistency (Qi et al., 2023), as well as classical agreement metrics (Scott, 1955; Cohen, 1960), do not account for model accuracy and can overestimate similarity. This leaves a gap for metrics such as $\kappa_p$, which explicitly account for error consistency with agreement to provide a more realistic view of multilinguality.

## 3 Methodology

The accuracy of LLMs differ greatly across languages and their performance is particularly in low-resource languages (Li et al., 2025). This can artificially inflate similarity scores for some languages as high performance leaves little room for disagreement (as explained further in Appendix A). Given that $\kappa_p$ addresses these issues, we use it to compare similarity of model outputs in light of variable performance across languages. Our work complements studies on representational similarity across languages such as (Wu et al., 2024).

$\kappa_p$ computes observed agreement $c^p_{\text{obs}}$ as the proportion with which the same option is selected across samples. To account for agreement by chance, $\kappa_p$ introduces an expected agreement $c^p_{\text{exp}}$, derived from the marginal distribution of each set of predictions. The $\kappa_p$ score is given by:

$$\kappa_p = \frac{c^p_{\text{obs}} - c^p_{\text{exp}}}{1 - c^p_{\text{exp}}},$$

We use the discrete variant of $\kappa_p$ as described in (Goel et al., 2025). As $\kappa_p$ increases, models make more similar mistakes, and their errors become more correlated, making them functionally more similar. Henceforth, we compute the average $\kappa_p$ using micro-averaging by concatenating all datasets in the group and then computing the $\kappa_p$ across the combined set. Since $\kappa_p$ is non-linear, the technique of micro-averaging is preferred as it smooths out extremes and operates directly at the per-sample level to better understand $\kappa_p$ across a dataset.

We use Gemma-3 (1B, 4B and 12B variants) (Team et al., 2025) and Qwen-3 (1.7B, 4B, 8B and 14B variants) (Yang et al., 2025) in our experiments, as they are some of the latest models as of August 2025 which have undergone multilingual pretraining. We also use the older Gemma-7B (Team et al., 2024) as a sanity check. We evaluate these models on a subset of 20 languages of the GlobalMMLU dataset (Singh et al., 2024) with our choice of languages justified in Appendix B. Building on our evaluation methodology, we leverage the LM Evaluation Harness (Gao et al., 2024), a unified framework for testing generative language models on a wide variety of benchmarks known for its reproducibility and extensive adoption.

## 4 Experimentation

### 4.1 Intra-Model Multilingual Similarity

**RQ1: Are LLMs becoming similar across languages?** Motivated by the findings of (Huh et al., 2024) which shows that model representations tend to converge with an increase in size and performance of models, we investigate whether a similar convergence occurs in the output space across languages. A clear trend is observed — as the model size increases, the average $\kappa_p$ score across languages also increases. $\kappa_p$ also positively correlates with model accuracy. These findings suggest

that outputs become more consistent across languages for larger and more accurate LLMs. The statistically significant results are illustrated in Figure 2. A possible reason for this could be that bigger models are trained on a greater volume of data including from low resource languages allowing for greater similarity. But it is not possible to confirm this hypothesis as we do not have access to their exact training data.

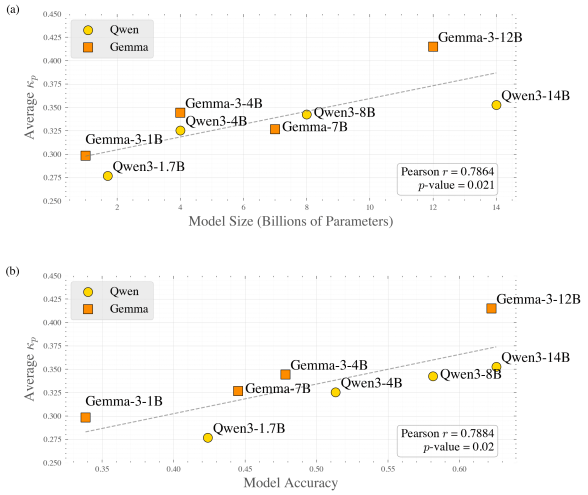



Figure 2: $\kappa_p$ correlates positively with model size and accuracy. (a) $\kappa_p$ averaged over languages positively correlates with model size (b) Similarly, $\kappa_p$ averaged over languages positively correlates with model performance. This indicates that models grow similar across languages with their capability and size.

**RQ2: Does the domain of questions asked matter?** Prior work shows that the language of prompting shapes LLM outputs, influencing both cultural preferences and ethical judgments (Vida et al., 2024; Agarwal et al., 2024; Aksoy, 2025). We thus hypothesize that models will be more inconsistent for subjects like ethics, morality, and sociology, which tend to be heavily influenced by sociocultural norms, as opposed to topics with relatively fewer cultural priors, such as mathematics and computer science. The questions in GlobalMMLU are divided into four domains- *STEM, Humanities, Social Sciences* and *Other*. We further subdivide these categories to provide a more detailed analysis. $\kappa_p$ tends to be greater for *STEM* in all the models as opposed to the other subjects (see Figure 3). This affirms our hypothesis about language sensitivity for culturally sensitive domains. Looking at the fine-grained categories (refer Table 6) in Figure 4 we continue to see a substantial difference between $\kappa_p$ of the subjects.

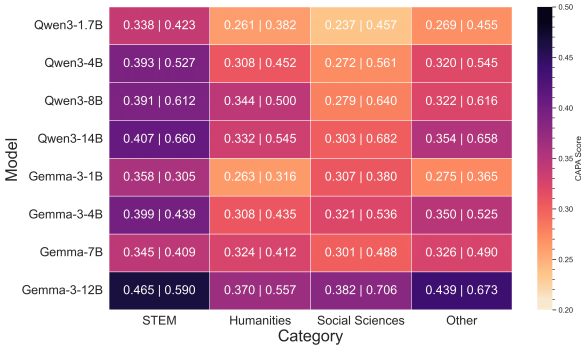



Figure 3: Models answer more similarly across languages for STEM than other domains. Each heatmap cell represents the $\kappa_p$ and accuracy averaged over languages. For example, a cell value of (0.3 | 0.4) for a given model and category would represent an average $\kappa_p$ of 0.3 and an average accuracy of 40%, both averaged over all the languages.

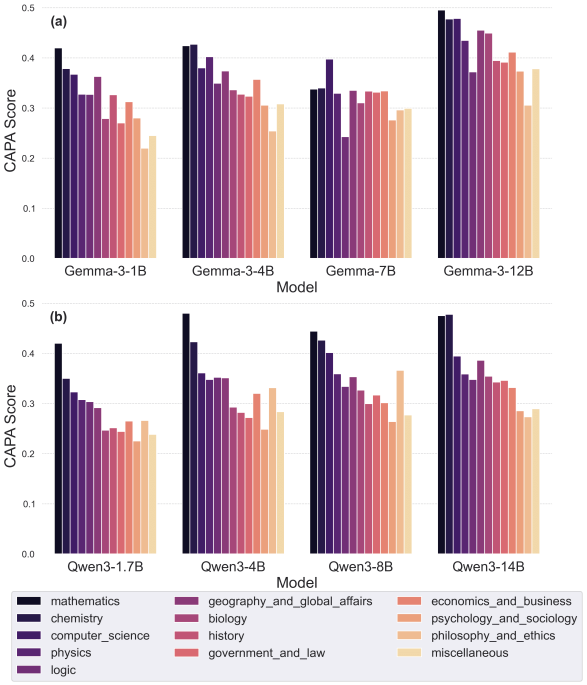



Figure 4: Intra-model $\kappa_p$ scores are higher for categories belonging to STEM (Mathematics, Physics, Computer Science) than the Humanities (Philosophy, Psychology, Sociology). (a) Family of Gemma models (b) Family of Qwen Models.

## 4.2 Inter-Model Multilingual Similarity

**RQ3: Do models agree more on high-resource languages?** When we average the $\kappa_p$ scores for a given language across all unique model pairs - a clear trend emerges - high-resource languages tend to have greater inter-model functional similarity, implying that the results are more consistent for languages like English than Amharic across all the

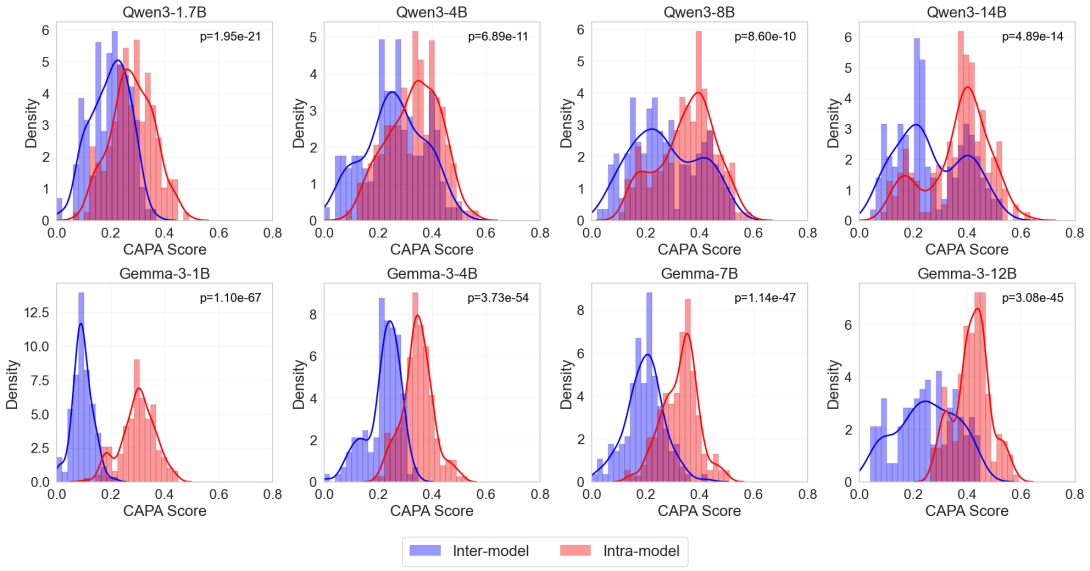



Figure 5: Frequency density distribution of the intra-model (across 20 language pairs) and inter-model (1 model vs remaining 7) $\kappa_p$ scores along with the p-values of the Mann-Whitney U Test. Intra-Model similarity is greater for all models than Inter-Model similarity with high significance.

models. We confirm this by using the number of Wikipedia articles for a given language as a proxy for their resource availability. Figure 6 indicates a significant positive correlation between the count of Wikipedia articles and inter-model functional similarity $\kappa_p$ score.

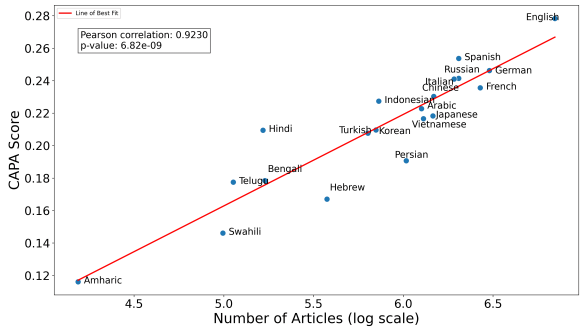



Figure 6: Higher-resource languages exhibit more model agreement. We observe a high correlation between $\kappa_p$ and number of wiki articles (*Pearson correlation* = 0.923).

**RQ4: Is cross-lingual similarity within the same model stronger than cross-model similarity in the same language?** For each model, we find the distribution of the $\kappa_p$ scores for two cases- Intra-Model (across all unique language pairs) and Inter-Model (across all models for each language). For the most part, models tend to be more similar to themselves for different languages than other models for the same language (see Figure 5). We employ the Mann-Whitney U test (Nachar et al., 2008) - a non-parametric statistical test commonly used to compare two independent samples - for this purpose. The null hypothesis of this test is that randomly selected values from two populations have the same distribution. The p-values ($< 0.001$) indicate that all tests are statistically significant, confirming that the intra-model and inter-model similarity distributions are significantly different, with intra-model scores tending to be higher. We further conduct an ablation using English, the highest-resource language, as a pivot. The results (see Appendix D) remain consistent: intra-model similarity scores are higher than inter-model similarity scores, reinforcing our main findings. Additionally, we find that the functional and representational similarity correlate to a certain degree in Appendix E.

## 5 Conclusion

We introduced $\kappa_p$ as a functional similarity metric for evaluating multilingual consistency in LLMs. Across GlobalMMLU, we found that larger and more capable models are more consistent across languages, with intra-model similarity exceeding inter-model similarity. Consistency also varies by domain — being higher in STEM than in culturally sensitive subjects — and by resource availability, with high-resource languages showing stronger inter-model agreement. Together, these results establish $\kappa_p$ as a practical tool for analyzing multilingual functional behavior beyond accuracy alone.

## 6 Future Work

We advocate for $\kappa_p$ to be used as a tool for analyzing multilinguality. We find interesting observations on the GlobalMMLU dataset, and feel that using this approach would be beneficial to the field of multilingual NLP in addition to the substantial work already being done in the representational space. There is also a great scope to explore if the two notions of similarity have any fundamental connection.

Although we hypothesize that having more data could help in improving multilingual consistency, it is also possible that it is inherently easier to learn one language from a greater capacity in another language if their underlying structures are similar. Is the cause of high functional similarity between two languages a function of their training (multilingual or parallel corpus), a natural alignment or a common syntactic structure of the two languages, or something different altogether? Establishing causality to our observations using interpretability techniques would be challenging but worthwhile.

Besides our current use case, we can see it being valuable in several applications. Higher functional similarity between two languages can have consequences on downstream tasks. For example, if a model with high $\kappa_p$ between Hindi and English exists, it might become easier to translate between the two languages. Furthermore, it might allow such models to interpret Hindi-English code mixed text samples more easily than another pair with a lower score.

## 7 Limitations

Although our findings establish statistically significant correlations across languages and models, we cannot establish causality for the observed phenomena as this would require extensive mechanistic interventions. $\kappa_p$ is limited to multiple-choice benchmarks, and there is a lack of free-form functional similarity metrics that take error consistency into account. This restricts our study to multilingual MCQ benchmarks. Additionally, there is also a lack of parallel multilingual MCQ benchmarks, and most existing ones, such as (Xuan et al., 2025), are variants of MMLU. Hence, we limit our analysis to the largest of these, GlobalMMLU.

## Acknowledgments

The authors of the paper would like to thank Srija Mukhopadhyay, Hemang Jain, Sweta Jena, Monish Singhal and other members of IIITH's Precog lab for their valuable feedback and support. We thank Eleuther AI as their tool, lm-evaluation-harness (Gao et al., 2024), was instrumental in our experimentation.

## References

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of llms depend on the language we prompt them in. *arXiv preprint arXiv:2404.18460*.

Meltem Aksoy. 2025. Whose morality do they speak? unraveling cultural bias in multilingual language models. *Natural Language Processing Journal*, page 100172.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. 2025. Great models think alike and this undermines ai oversight. *arXiv preprint arXiv:2502.04313*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. 2025. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, and 1 others. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *arXiv preprint arXiv:2402.12840*.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: A metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28186–28194.

Nadim Nachar and 1 others. 2008. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1):13–20.

Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: large language models in non-english content analysis. *arXiv preprint arXiv:2306.07377*.

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

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, and 1 others. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.

William A Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

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

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Karina Vida, Fabian Damken, and Anne Lauscher. 2024. Decoding multilingual moral preferences: Unveiling llm's biases through the moral machine experiment. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1490–1501.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2024. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. *arXiv preprint arXiv:2411.04986*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, and 1 others. 2025. Mmluprox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

## A $\kappa_p$ vs Other Metrics

We choose $\kappa_p$ as it has clear advantages over other metrics which have been theoretically and empirically validated in (Goel et al., 2025). $\kappa_p$ metric is chance-adjusted, meaning it is not inflated when model accuracy is high. An example to help understand this is A model with 95% accuracy in English and Spanish answers 95/100 questions correctly in both. Raw agreement appears high, but this is trivial—it reflects correctness. $\kappa_p$ downweighs such expected agreement. In contrast, with 50% accuracy in two low-resource languages, if the model makes similar mistakes, $\kappa_p$ captures this meaningful functional similarity as agreement beyond chance. When we compare it to other metrics Cohen's $\kappa$ and Scott's $\pi$, we observe the difference in inflation due to accuracy.

Note that in our analysis we are using the discrete variant of $\kappa_p$ which converts probability logits to their softmax labels. Consider two raters with predictions [0, 0, 0, 1, 2, 1] and [0, 0, 0, 1, 2, 0] respectively with ground truths [0, 0, 0, 1, 2, 2].

- **Cohen's $\kappa$**

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{\frac{5}{6} - \frac{15}{36}}{1 - \frac{15}{36}}$$

$$= \frac{0.833 - 0.417}{0.583} \approx 0.714$$

- **Scott's $\pi$**

$$\pi = \frac{P_o - P_e}{1 - P_e}$$

where $P_o = \frac{5}{6} = 0.833$,

$$P_e = (0.583)^2 + (0.250)^2 + (0.167)^2$$

$$= 0.431$$

thus

$$\pi = \frac{0.833 - 0.431}{1 - 0.431} \approx 0.707$$

- $\kappa_p$

$$\kappa_p = \frac{c^{E,M}_{\text{obs}} - c^{E,M}_{\text{exp}}}{1 - c^{E,M}_{\text{exp}}}$$

where $c^{E,M}_{\text{obs}} = \frac{5}{6}$,

$c^{E,M}_{\text{exp}} = acc_1 \times acc_2 = \frac{5}{6} \times \frac{5}{6} = \frac{25}{36}$

thus

$$\kappa_p \approx 0.45$$

Since both models are highly accurate (83.3%), the similarity scores as measured by traditional metrics are inflated. This is not the case with $\kappa_p$ as it takes model accuracy into account.

All the results we have presented for remains consistent for other metrics. These results substantiate our findings, indicating their robustness and generalizability beyond the confines of the $\kappa_p$ metric. Computed values are in tables 2, 3 and 4.

Here we showcase a numerical example of the advantage of probabilistic $\kappa_p$ over RankC (Qi et al., 2023). Consider two raters with probabilistic predictions

$$R_1 = \begin{bmatrix} 0.50 & 0.45 & 0.05 \\ 0.50 & 0.05 & 0.45 \\ 0.05 & 0.45 & 0.50 \end{bmatrix}$$

$$R_2 = \begin{bmatrix} 0.50 & 0.05 & 0.45 \\ 0.50 & 0.45 & 0.05 \\ 0.45 & 0.05 & 0.50 \end{bmatrix}.$$

Finding the maximum probabilities from $R_1$ and $R_2$, the hard labels are

$$r_1 = [0, 0, 2], \quad r_2 = [0, 0, 2].$$

Thus $P_o = 1$.

- **Cohen's $\kappa$**

Rater marginals: $p^{(1)} = p^{(2)} = \left[\frac{2}{3},\ 0,\ \frac{1}{3}\right]$,

$$P_e = \sum_i p_i^{(1)} p_i^{(2)} = \left(\tfrac{2}{3}\right)^2 + 0^2 + \left(\tfrac{1}{3}\right)^2 = \tfrac{5}{9},$$

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{1 - \frac{5}{9}}{1 - \frac{5}{9}} = 1.0.$$

- **Scott's $\pi$**

Pooled counts over both raters: $[4,\ 0,\ 2]$

$p = \left[\frac{2}{3},\ 0,\ \frac{1}{3}\right]$,

$$P_e = \sum_i p_i^2 = \left(\tfrac{2}{3}\right)^2 + 0^2 + \left(\tfrac{1}{3}\right)^2 = \tfrac{5}{9},$$

$$\pi = \frac{P_o - P_e}{1 - P_e} = \frac{1 - \frac{5}{9}}{1 - \frac{5}{9}} = 1.0.$$

- **RankC**

For each item, let $r^{(1)}, r^{(2)}$ be class rankings from $R_1, R_2$.For $j = 1, 2, 3$

$$\text{P@}j \ = \ \frac{\left|\text{Top-}j(r^{(1)}) \cap \text{Top-}j(r^{(2)})\right|}{j}$$

Weights: $w_j = \dfrac{e^{3-j}}{\sum_{\ell=1}^{3} e^{3-\ell}}$

$\Rightarrow \quad (w_1, w_2, w_3) \approx (0.665,\ 0.245,\ 0.090).$

From the matrices:

$(\text{P@1}, \text{P@2}, \text{P@3}) = (1,\ 0.5,\ 1).$

$$\Rightarrow \text{item score} = \sum_{j=1}^{3} w_j \cdot \text{P@}j$$

$$= 0.665 \cdot 1 + 0.245 \cdot 0.5 + 0.090 \cdot 1$$

$$\approx 0.878.$$

Averaging over all three items (identical here) gives

$$\text{RankC} \approx 0.878.$$

- $\kappa_p$

$$\kappa_p = \frac{c^{p}_{\text{obs}} - c^{p}_{\text{exp}}}{1 - c^{p}_{\text{exp}}}$$

where $c^{p}_{\text{obs}} = \ \dfrac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} p_{i,c}^{(1)}\, p_{i,c}^{(2)} = 0.295$

$$c^{p}_{\text{exp}} = \ \bar{p}^{(1)}\, \bar{p}^{(2)} \ + \ \frac{\left(1 - \bar{p}^{(1)}\right)\left(1 - \bar{p}^{(2)}\right)}{C - 1}$$

$$= 0.375$$

thus

$$\kappa_p = -0.128.$$

Collapsing to hard labels yields perfect agreement ($\kappa = \pi = 1.0$). RankC, which compares top-$j$ sets from the probability rankings, shows high but non-perfect agreement ($\approx 0.878$). $\kappa_p$, which directly

evaluates the full probability distributions, detects conflicting uncertainty allocations across classes and therefore yields a *negative* chance-corrected agreement ($-0.128$). This is intuitive, as when the models are incorrect, they give very different (and in fact, opposite) predictions which is not captured by the other metrics.

## B Choice of Languages Used

We choose to do our analysis over twenty languages as listed in Table 1. The languages chosen belong to a wide range of groups, including the Afro-Asiatic (Amharic, Arabic, Hebrew), Dravidian (Telugu), Germanic (English, German), and Indo-Iranian (Persian, Hindi, Bengali) language families/branches, among others. The subset of GlobalMMLU was curated to represent a spectrum of resource availability, where high-resource languages refer to those with abundant linguistic data, such as large corpora, annotated datasets, and digital tools (e.g., English, Spanish), while low-resource languages lack such resources and infrastructure (e.g., Amharic, Telugu). This selection allows us to assess model behavior across typologically and resource-diverse settings. All the languages have an equal number of questions, and we have chosen the subset among these which have consistent answers among all the languages leading to a total of 13844 questions in each language.

## C Sub-Categorization of GlobalMMLU

We sub-categorized the existing categories of GlobalMMLU to make better and fine-grained inferences. We follow the standard GlobalMMLU setup in lm-evaluation-harness (Gao et al., 2024) to conduct the evaluations. The tables 5 and 6 show the categorization based on the four domains and further split 14 categories, respectively. The tables also show the distribution of the samples for each category. Each numerical value in the Samples columns of the table corresponds to the number of resulting samples for a given model for a given language.

## D Ablations for Inter-model vs Intra-model Similarity

We explore an alternate way to plot inter-model similarity by removing potential confounders from cross-size comparisons. Initially, the computation for the inter-model similarity was plotting the distribution of the computed $\kappa_p$ values for each model with the remaining seven models across 20 languages. For intra-model similarity, we compute, for each model, the distribution of $\kappa_p$ values across 20 unique language pairs. For this ablation, we compute the $\kappa_p$ values for inter-model similarity to be a single $\kappa_p$ value for each model with the model of the other family with the closest number of parameters (model size). We then plot two distributions for intra-model similarity. In Figure 8a, the intra-model similarity computation remains the same, calculation $\kappa_p$ across 20 unique language pairs. In Figure 8b, the intramodel similarity distribution has been revised to include only pairs of English-non-English languages ($en - \{lang\}$). The results remain consistent with previous results, showing that intra-model similarity is still greater than inter-model similarity.

## E Some Correlation Between Functional and Representational Similarity

Following the procedure in (Wu et al., 2024), we compute the representation cosine similarity and use the last token position as the sentence representation over a subset of the translation dataset, FLORES-101 (Goyal et al., 2022). We subtract these scores by a baseline of non-matching sentences and find that when two languages have a greater $\kappa_p$ score, i.e. they have high functional similarity, they also tend to have a greater representational similarity as measured by the increase over the baseline. We do it over limited layers of the Qwen model (Qwen3-4B and Qwen3-8B) due to compute constraints. This experiment is carried out to establish some degree of correlation between the two notions of similarity, the existence of which has been debated before in (Klabunde et al., 2025).

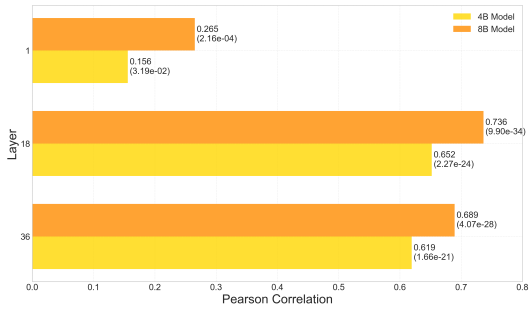
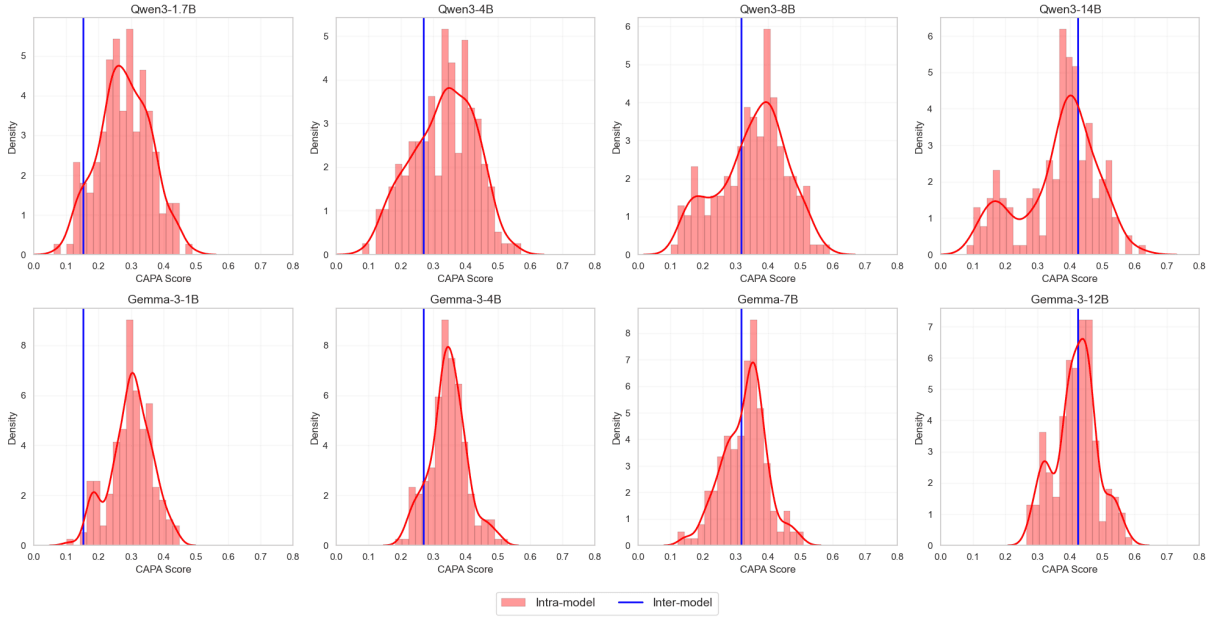


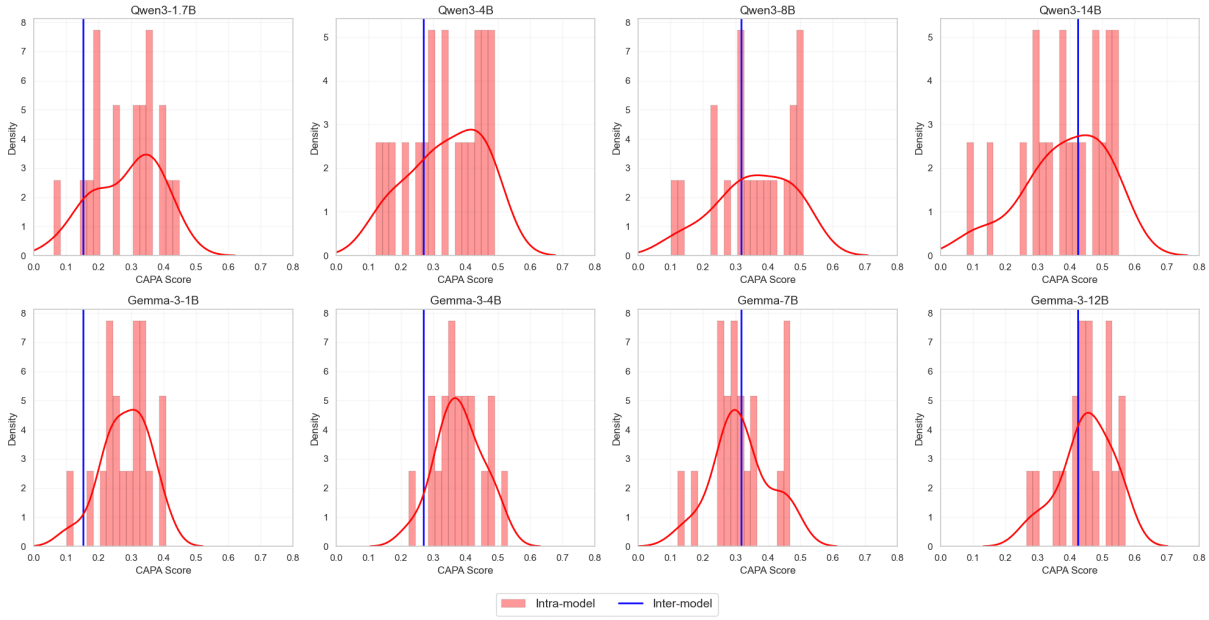

Figure 7: Languages with higher functional similarity ($\kappa_p$) also exhibit greater representational similarity. Representation cosine similarity is computed using the last token position from FLORES-101 sentence pairs. Scores are baseline-adjusted using non-matching sentence pairs.

| Code | Language | Code | Language | Code | Language | Code | Language |
|---|---|---|---|---|---|---|---|
| am | Amharic | fr | French | it | Italian | es | Spanish |
| ar | Arabic | de | German | ja | Japanese | sw | Swahili |
| bn | Bengali | he | Hebrew | ko | Korean | te | Telugu |
| zh | Chinese | hi | Hindi | fa | Persian | tr | Turkish |
| en | English | id | Indonesian | ru | Russian | vi | Vietnamese |

Table 1: Language codes and their corresponding language names used in our experiments.




(a) Frequency distribution of the intra-model (across 20 language pairs) and inter-model (1 model vs closest family model).




(b) Frequency distribution of the intra-model (across English-non-English pairs) and inter-model (1 model vs closest family model).

| Metric | Pearson correlation for Size | Pearson correlation for Accuracy | Pearson correlation for Resource (log no. of Articles) |
|---|---|---|---|
| $\kappa_p$ | 0.7864<br>(0.02062) | 0.7884<br>(0.02009) | 0.9230<br>(6.82e-09) |
| Cohen's $\kappa$ | 0.8862<br>(0.003376) | 0.9714<br>(5.694e-05) | 0.9321<br>(2.28e-09) |
| Scott's $\pi$ | 0.8861<br>(0.003385) | 0.9714<br>(5.728e-05) | 0.9313<br>(2.53e-09) |

Table 2: Pearson correlation coefficients (top) with p-values in parentheses (bottom).

| Metric | Qwen3-1.7B | Qwen3-4B | Qwen3-8B | Qwen3-14B |
|---|---|---|---|---|
| $\kappa_p$ | 12150<br>(1.95e-21) | 16686<br>(6.89e-11) | 17278<br>(8.60e-10) | 15148<br>(4.89e-14) |
| Cohen's $\kappa$ | 23750<br>(6.08e-02) | 21710<br>(1.29e-03) | 17992<br>(1.48e-08) | 14766<br>(6.90e-15) |
| Scott's $\pi$ | 22358<br>(5.26e-03) | 21644<br>(1.11e-03) | 17542<br>(2.53e-09) | 14460<br>(1.38e-15) |

Table 3: Mann–Whitney U statistics for Qwen models (p-values in parentheses).

| Metric | gemma-3-1b-it | gemma-3-4b-it | gemma-7b | gemma-3-12b-it |
|---|---|---|---|---|
| $\kappa_p$ | 180<br>(1.10e-67) | 3050<br>(3.73e-54) | 4556<br>(1.14e-47) | 5148<br>(3.08e-45) |
| Cohen's $\kappa$ | 5660<br>(3.46e-43) | 11650<br>(7.83e-23) | 16394<br>(1.88e-11) | 7316<br>(6.87e-37) |
| Scott's $\pi$ | 4364<br>(1.79e-48) | 11176<br>(3.37e-24) | 15598<br>(4.53e-13) | 7228<br>(3.27e-37) |

Table 4: Mann–Whitney U statistics for Gemma models (p-values in parentheses).

| Domain | Subjects | # Samples |
|---|---|---|
| STEM | College Chemistry, High School Computer Science, College Biology, Abstract Algebra, High School Mathematics, Computer Security, Machine Learning, College Physics, Conceptual Physics, Astronomy, High School Biology, High School Physics, Anatomy, College Mathematics, Electrical Engineering, College Computer Science, High School Chemistry, High School Statistics, Elementary Mathematics | 3153 |
| Humanities | Philosophy, World Religions, Professional Law, Moral Scenarios, High School European History, Moral Disputes, Jurisprudence, Formal Logic, High School US History, Prehistory, High School World History, International Law, Logical Fallacies | 4511 |
| Social Sciences | High School Microeconomics, High School Geography, US Foreign Policy, Professional Psychology, Security Studies, High School Government and Politics, High School Psychology, Econometrics, Sociology, High School Macroeconomics, Public Relations, Human Sexuality | 3076 |
| Other | Professional Accounting, Professional Medicine, College Medicine, Marketing, Nutrition, Global Facts, Clinical Knowledge, Human Aging, Virology, Miscellaneous, Business Ethics, Management, Medical Genetics | 3104 |

Table 5: Original Grouping of GlobalMMLU subjects into 4 domains with corresponding sample counts.

| Category | Subjects | # Samples |
|---|---|---|
| Mathematics | Abstract Algebra, College Mathematics, Elementary Mathematics<br>High School Mathematics, High School Statistics, Formal Logic<br>Logical Fallacies | 1064 |
| Logic | Formal Logic, Logical Fallacies | 289 |
| Physics | College Physics, Conceptual Physics, High School Physics, Astronomy | 640 |
| Biology | College Biology, High School Biology, Human Aging<br>Human Sexuality, Virology | 971 |
| Chemistry | College Chemistry, High School Chemistry | 303 |
| Medicine | Anatomy, Clinical Knowledge, College Medicine<br>Medical Genetics, Nutrition, Professional Medicine | 1251 |
| Computer Science | College Computer Science, High School Computer Science<br>Computer Security, Machine Learning | 412 |
| Economics and Business | Econometrics, High School Macroeconomics, High School Microeconomics<br>Business Ethics, Management, Marketing<br>Professional Accounting | 1461 |
| Psychology and Sociology | High School Psychology, Professional Psychology, Sociology | 1358 |
| Geography and Global Affairs | Global Facts, High School Geography, US Foreign Policy<br>Security Studies | 643 |
| History | High School US History, High School European History<br>High School World History, Prehistory | 741 |
| Government and Law | High School Government and Politics, International Law, Jurisprudence<br>Professional Law | 1951 |
| Philosophy and Ethics | Philosophy, Moral Disputes, Moral Scenarios | 1552 |
| Miscellaneous | World Religions, Public Relations, Electrical Engineering, Miscellaneous | 1208 |

Table 6: Fine-grained categorization of GlobalMMLU subjects used in our ablation.