# Cross-Language Reasoning Evaluation of Large Language Models

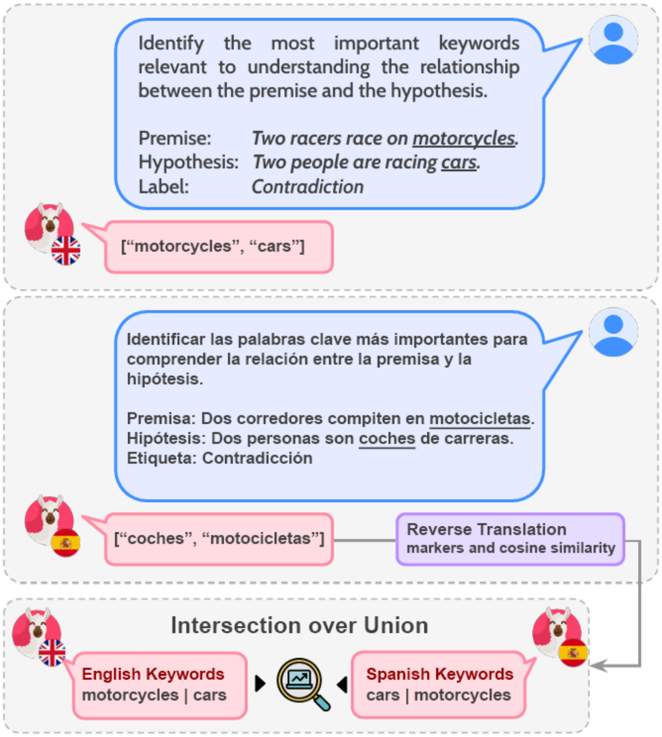
Gian Zignago
University of California, Los Angeles
grz@cs.ucla.edu

***Abstract*—Large Language Models (LLMs) have exhibited remarkable capabilities in generating coherent and contextually relevant rationales for diverse reasoning tasks. This capstone project evaluates the cross-language reasoning capabilities of 23 LLMs through a novel framework that combines translation-based prompting, reverse translation, and keyword marking techniques. This capstone introduces a pipeline to calculate reasoning performance across 68 languages using Intersection over Union (IoU) and rank correlation metrics (Spearman and Kendall Tau). The results demonstrate significant variations in performance across languages, providing insights into the influence of lexical similarity and speaker prevalence on LLM capabilities. These findings offer actionable guidance for the development of future LLMs.**

## I. INTRODUCTION

Large Language Models (LLMs) have demonstrated strong reasoning capabilities in multiple languages. However, evaluating their cross-language reasoning, particularly in comparison to English performance, remains an underexplored area. Fayyaz et al.'s prior work on human alignment and model faithfulness of LLM rationale underscored the challenges faced by LLMs when tasked with semantic reasoning tasks [1]. This paper builds on that foundation, incorporating multilingual translation, marking, and reverse translation into the evaluation pipeline to comprehensively analyze LLM reasoning in languages other than English. By introducing new methodologies and metrics, this project seeks to address limitations in prior evaluations.

Multilingual dialogue evaluation has historically been limited to a narrow set of languages, primarily due to the lack of linguistic diversity in dialogue corpora [2]. This constraint has hindered the development of multilingual chatbots and reduced the demand for multilingual evaluation metrics. However, recent advancements in LLMs offer a promising solution. These models, as highlighted by studies from Guo et al. and Bubeck et al. demonstrate fluency and adherence to linguistic conventions in widely studied languages [3][4].



Identify the most important keywords relevant to understanding the relationship between the premise and the hypothesis.

Premise: *Two racers race on motorcycles.*
Hypothesis: *Two people are racing cars.*
Label: *Contradiction*

["motorcycles", "cars"]

Identificar las palabras clave más importantes para comprender la relación entre la premisa y la hipótesis.

Premisa: Dos corredores compiten en motocicletas.
Hipótesis: Dos personas son coches de carreras.
Etiqueta: Contradicción

["coches", "motocicletas"]

Figure 1: An example of this study's analysis methodology on the MNLI dataset. The model is prompted for keywords in English and then in another language before the generated keywords are compared.

Despite this, inaccuracies and hallucinations persist, particularly in less-studied languages [5]. Furthermore, LLMs exhibit functional linguistic competence, such as discursive coherence, narrative structuring, and contextual understanding.

While these emergent capabilities have enabled significant advancements in multilingual chatbots, important questions remain. Can LLMs generate linguistically competent responses and evaluate dialogues based on formal and functional linguistic rules? Recent work, including Huynh et al., confirms the language understanding capabilities of instruction-based LLMs for dialogue evaluation [6]. Building on these findings, this capstone project presents a novel

| premise | hypothesis | label |
|---|---|---|
| The new rights are nice enough | Everyone really likes the newest benefits | 1 neutral |
| i don't know um do you do a lot of camping | I know exactly. | 2 contradiction |
| i'm not sure what the overnight low was | I don't know how cold it got last night. | 0 entailment |

Table 1: Example rows of manually annotated data from the GLUE MNLI dataset.

investigation into multilingual reasoning evaluation, with an emphasis on paraphrase robustness and multilinguality.

Existing literature highlights the advantages of multilingual pre training but reveals significant discrepancies in reasoning performance [7]. Current evaluations often rely on limited metrics, neglecting nuanced linguistic relationships and the inherent challenges of translation [8][9]. By leveraging LLMs' emerging multilingual capabilities, this project introduces a benchmark for multilingual reasoning, addressing gaps in the field and setting a foundation for future cross-linguistic evaluations.

**Key Findings of the Study**

- Languages lexically and syntactically similar to English, such as Dutch and Swedish, exhibited significantly higher IoU scores, highlighting the influence of linguistic proximity.
- Linguistically distant languages like Chinese, Vietnamese, and Kinyarwanda faced greater challenges, emphasizing the need for improved adaptations in diverse linguistic contexts.
- Gemini models demonstrated superior adaptability for non-Indo-European languages, while GPT models excelled in keyword ranking consistency across languages.
- Word limit analysis revealed that increasing word limits improves semantic reasoning but highlighted diminishing returns for morphologically complex or low-resource languages.
- Geographical disparities underscore the need for more inclusive and balanced training datasets to enhance LLM multilingual capabilities.

## II. Methodology and Data Collection

### *2.1 Dataset*

The GLUE MNLI dataset was selected as the foundational dataset for this project due to its established reputation and design tailored for premise-hypothesis reasoning tasks [10]. It consists of well-structured pairs of premises and hypotheses annotated with three possible relationships: entailment, contradiction, or neutral. Table 1 shows three example rows from the dataset. This structure makes it an ideal choice for evaluating reasoning consistency and contextual understanding across languages. By leveraging this dataset, the study ensures a balanced and representative set of language constructs, including complex sentence structures and nuanced semantic relationships. Additionally, the dataset's popularity in natural language processing (NLP) benchmarks provides a baseline for comparing multilingual reasoning capabilities across a diverse range of LLMs.

### *2.2 Data Preprocessing and Cleaning*

Data preprocessing played a crucial role in ensuring consistency across languages. Premises and hypotheses underwent normalization steps, including removing punctuation, converting text to lowercase, and handling special characters. These steps aimed to minimize inconsistencies arising from formatting differences and model-specific sensitivities. Furthermore, edge cases such as missing translations, improperly formatted sentences, or phrases misaligned during reverse translation were addressed during this stage.

### *2.3 Translation Corpus*

The translation corpus played a pivotal role in enabling the generation of multilingual prompts from the original English premises and hypotheses. After experimenting with multiple translation methods, including Google's Gemini, OpenAI's GPT-4, and Meta's LLaMA, it was determined that Helsinki NLP's Opus-MT models provided the most reliable English-to-target-language translations [11]. These models have been shown to consistently maintain the semantic integrity of complex sentences while balancing grammatical accuracy, making them the optimal choice for this study [12].

Helsinki NLP's Opus-MT excelled in preserving the meaning of nuanced and syntactically complex sentence structures, which was critical for ensuring consistency across multilingual prompts. While other models such as GPT-4 and Gemini demonstrated strengths in specific linguistic features—such as handling low-resource languages or achieving high semantic alignment in widely spoken languages—they often struggled with edge cases, such as morphologically rich languages or idiomatic expressions. In contrast, Helsinki NLP consistently delivered translations that accurately reflected the intent and meaning of the original English sentences, minimizing semantic drift.

This reliance on Helsinki NLP also simplified the workflow by providing a uniform translation standard across all target languages. Although alternative methods were explored to cross-validate translations, Helsinki's ability to reliably capture the complexities of reasoning tasks across a wide range of languages reduced the need for integrating multiple translation models. This decision ensured a consistent translation pipeline and highlighted the potential of

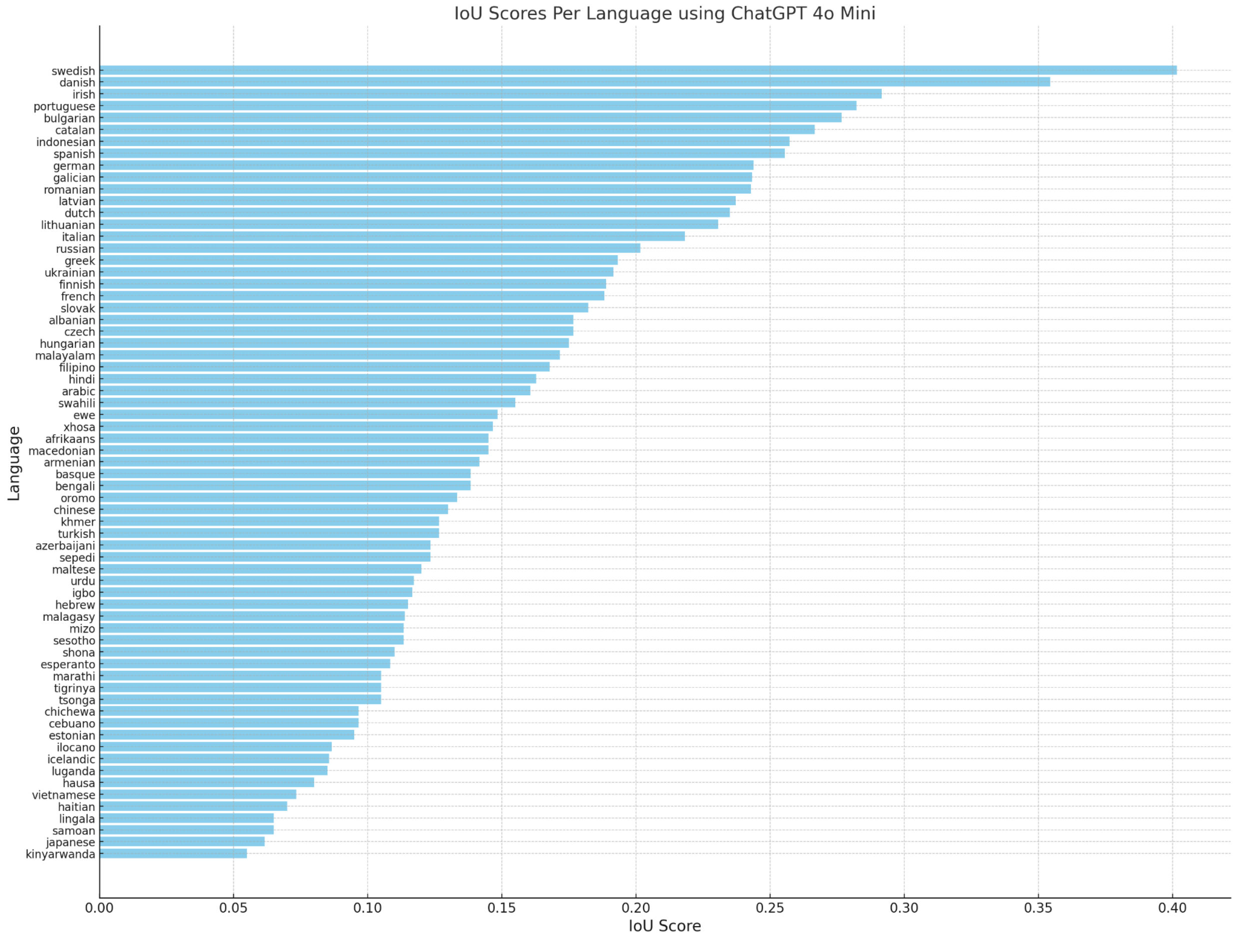


Figure 2: Sorted IoU scores for each of the 68 tested languages using ChatGPT 4o Mini with a subset of 30 items per language and a limit of 3 keyword responses.

Opus-MT models for multilingual natural language processing tasks. This decision meant that the study was limited to 68 languages, however, as there are only 68 supported Opus-MT models for direct translation from english to other languages.

By anchoring the evaluation on a single, high-performing translation system, the study eliminated variability caused by combining outputs from diverse models. This focus enabled a sharper analysis of reasoning performance across languages, ensuring that any observed discrepancies in model outputs were attributable to differences in reasoning capabilities rather than inconsistencies in translation quality.

### *2.4 Reverse Translation and Marking*

The process of translating extracted keywords from target languages back into English was essential for directly comparing multilingual reasoning with English, as it revealed how well semantic relationships were preserved during translation. To achieve this, the study employed the Deep Translate Google Translate module, a reliable tool for high-quality reverse translations. This module provided consistent performance across languages, handling both high-resource and low-resource linguistic contexts effectively. Unlike prior approaches that relied solely on basic translation outputs, this study incorporated advanced semantic alignment techniques, including cosine similarity, to refine reverse translations. Cosine similarity proved particularly effective in addressing lexical ambiguity, where a single translated word could have multiple English equivalents. By embedding keywords using transformer-based models and comparing them to embeddings of the original English premises and hypotheses, the study ensured that reverse-translated keywords closely matched their intended meanings. This approach significantly reduced errors caused by polysemy – words with multiple meanings – and enhanced the reliability of the evaluations.

$$\text{Cosine Similarity} = \frac{\vec{v_1} \cdot \vec{v_2}}{\|\vec{v_1}\| \|\vec{v_2}\|} \tag{1}$$

Additionally, the marking process incorporated a context-aware methodology to further refine keyword alignment. Keywords were evaluated for their semantic

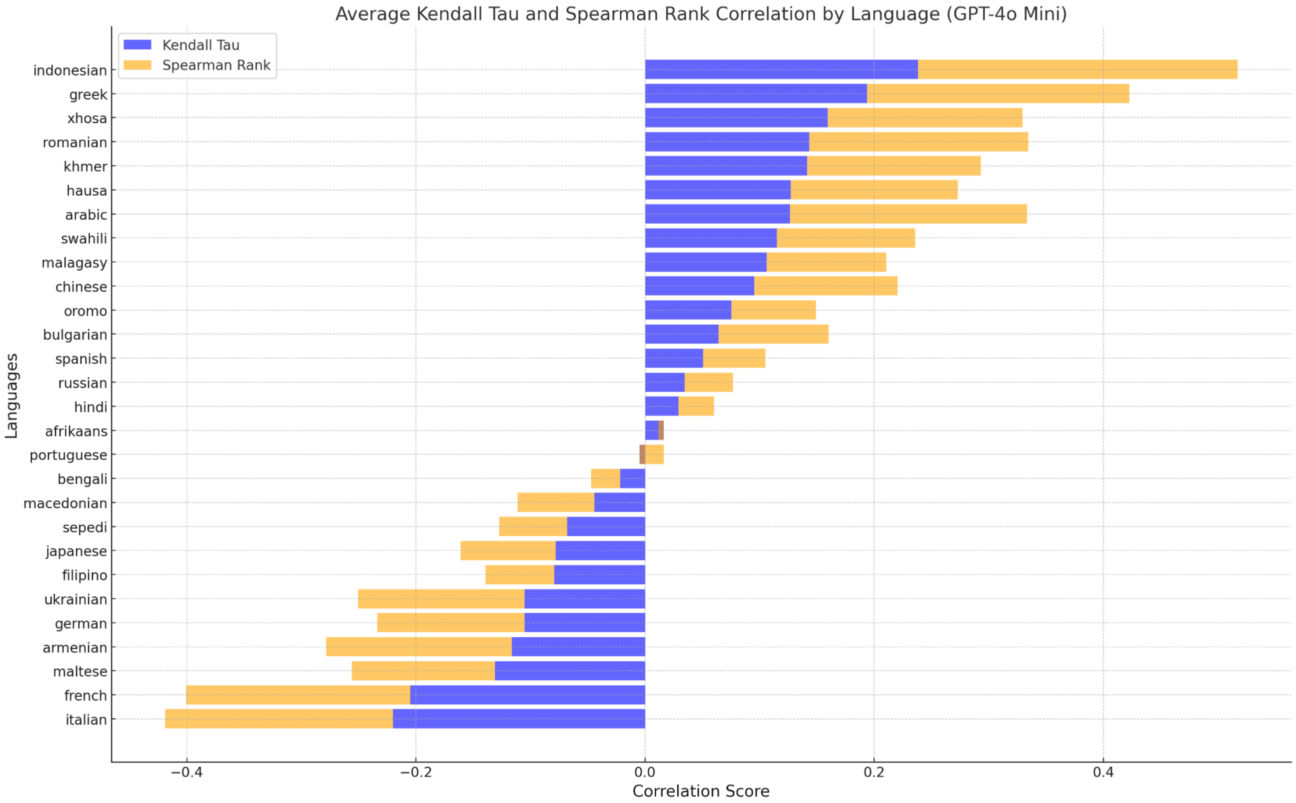


Figure 3: Kendall tau and spearman rank scores for a subset of 28 of the tested languages using ChatGPT 4o Mini with a subset of 30 items per language and a limit of 3 keyword responses, sorted in descending order by kendall tau score.

similarity and their relevance within the context of the original premise or hypothesis. If a reverse-translated keyword aligned with both the premise and hypothesis, its placement in the original target-language sentence was analyzed to disambiguate its intended context. For instance, positional alignment within sentences proved particularly effective for languages with flexible word orders, such as German and Chinese, ensuring robust and accurate keyword marking.

The Helsinki NLP's Opus-MT translation models were found to be insufficient for reverse translations from target languages to english. Opus-MT often struggled with low-resource languages and exhibited significant semantic drift during reverse translation, particularly for idiomatic expressions and syntactically complex sentences. Additionally, the models' tendency to prioritize fluency over fidelity led to inconsistencies when comparing reverse-translated keywords with their original English counterparts. By leveraging the Deep Translate module and combining it with semantic and contextual alignment techniques, the study created a robust reverse translation pipeline. This hybrid approach ensured that reverse translations retained both lexical fidelity and contextual accuracy, enabling a more nuanced analysis of multilingual reasoning.

## III. EXPERIMENT SETUP

### *3.1 Evaluation Metrics*

Evaluating multilingual reasoning required the selection of metrics that could capture both the semantic fidelity and contextual relevance of model outputs. Two primary metrics were employed: Intersection over Union (IoU) and rank correlation (using Spearman and Kendall Tau coefficients). These metrics were chosen for their ability to quantify different aspects of reasoning performance.

IoU was used to compare the overlap between the set of keywords extracted by the model in a given language and the corresponding generated English keywords. By focusing on overlap, IoU provided a direct measure of how well models preserved the semantic essence of reasoning tasks across languages. Equation 2 provides the method used for calculating IoU for each generated keyword batch. Unlike absolute reasoning benchmarks, IoU highlighted comparative performance, revealing how language affected reasoning consistency relative to English. This approach was critical for identifying gaps in multilingual capabilities, as English was used as the baseline for all reasoning comparisons.

$$IoU = \frac{|Union\ of\ Keywords|}{|Intersection\ of\ Keywords|} \qquad (2)$$

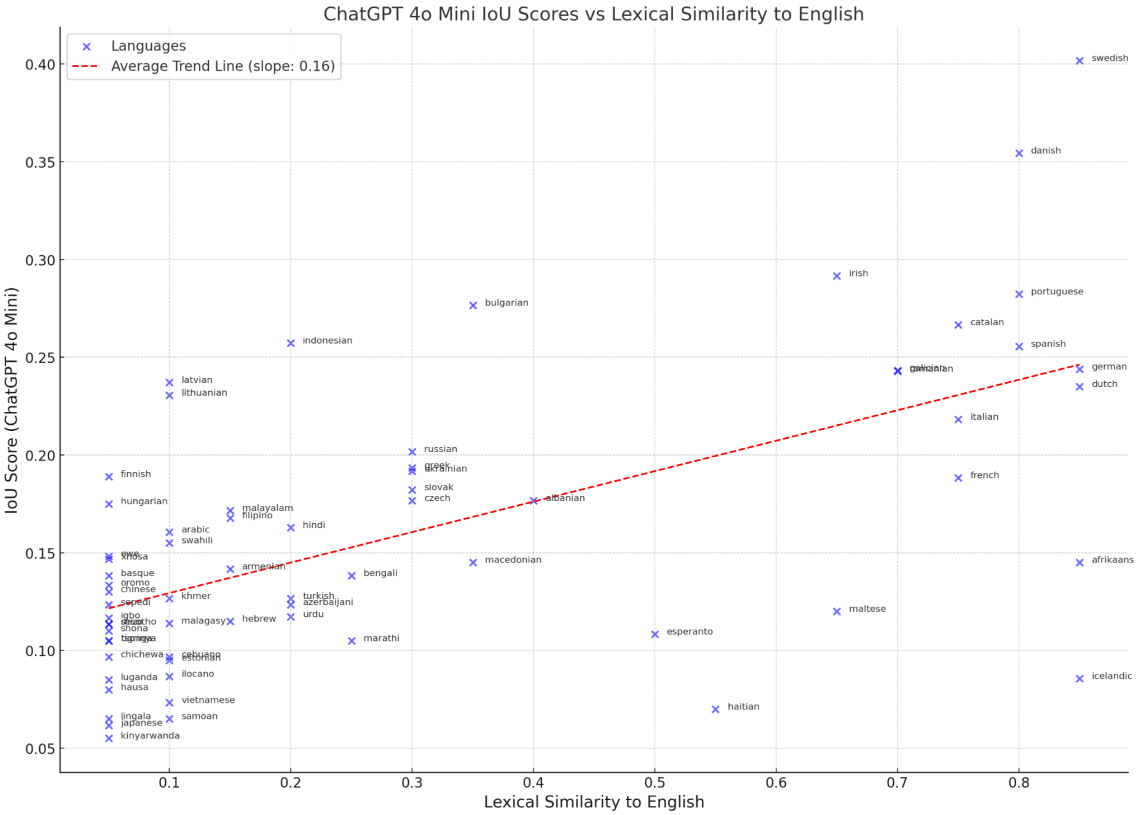


Figure 4: Each language's IoU score using ChatGPT 4o Mini compared to its quantified lexical similarity to English with an average trend line overlaid

While IoU measured keyword overlap, rank correlation metrics evaluated the order in which keywords were deemed relevant by the models. Spearman rank correlation assessed monotonic relationships, indicating whether higher-ranked keywords in one language corresponded to higher ranks in English. This metric provided a broader measure of correlation by evaluating whether the overall ranking order was preserved between languages, regardless of exact ranking matches. Kendall Tau, on the other hand, provided a more granular measure by examining pairwise ranking agreements. it captures pairwise rank consistency, offering insights into how well the relative importance of keywords was maintained across languages. Together, these metrics captured both global and local alignment of reasoning priorities, offering a comprehensive view of multilingual reasoning consistency.

Equation 3 shows the formal method for calculating Kendall Tau, where $(n¦2)$ is the total number of pairs of keywords, and $n$ is the total number of keywords in the set. Equation 4 shows the calculation for Spearman rank correlation, where dᵢ is the rank of a given keyword in English subtracted by the rank of the keyword in the translated language.

$$\tau = \frac{Number\ of\ Concordant\ Pairs - Number\ of\ Discordant\ Pairs}{(n¦2)} \quad (3)$$

$$\rho = 1 - \frac{6\, \Sigma\, d_i^2}{n(n^2-1)} \quad (4)$$

### *3.2 Metric Calculation and Implementation*

To calculate IoU, the system first extracted sets of keywords from both English and translated prompts. These sets were then compared using set intersection and union operations, with the IoU score computed as the ratio of the size of the intersection to the size of the union. This calculation allowed for intuitive comparisons across models.

Rank correlation metrics required more sophisticated computations. The system ranked keywords based on their importance as identified by the model, assigning numerical values to each keyword. These rankings were then compared between English and translated prompts to compute Spearman and Kendall Tau coefficients. By automating these calculations, the system ensured consistency and reproducibility across experiments, even when dealing with large datasets and multiple languages.

### *3.3 System Implementation for Reverse Translation*

Figure 1 shows the system's prompting and reverse translation pipeline, which was designed as the primary method of addressing the challenge of aligning multilingual

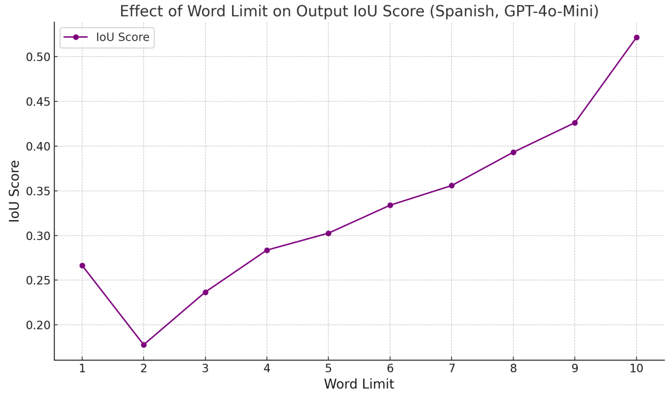


Figure 5: Effect of word limits on IoU scores for Spanish using GPT-4o-Mini ran on a subset of 30 items with a single iteration.

keywords with their English counterparts. Unlike traditional approaches that relied solely on translation accuracy, this pipeline incorporated semantic matching techniques to improve alignment. Extracted keywords were embedded using transformer models, such as BERT or Sentence-BERT, to capture their contextual meanings. These embeddings were then compared with English embeddings using cosine similarity, enabling the identification of the closest semantic match. Equation 1 describes the method for determining cosine similarity between two words, represented as vectors v1 and v2.

## IV. RESULTS

### 4.1 IoU Scores Across Models and Languages

The IoU scores provided valuable insights into how well each language and model preserved reasoning performance relative to English. As shown in Figure 2, Figure 8 and Figure 9, there was significant variation in IoU scores across languages. For instance, languages such as Dutch, Swedish, and German consistently achieved higher IoU scores across all models, reflecting their close linguistic and syntactic similarity to English. In contrast, languages such as Japanese, Vietnamese, and Kinyarwanda exhibited lower IoU scores, highlighting the challenges of adapting reasoning tasks to linguistically distant languages.

Figure 9 shows that the Gemini model outperformed others in maintaining high IoU scores across a diverse range of languages, with Dutch achieving the highest score of approximately 0.48. GPT and LLaMA models followed closely, as seen in Figure 2 and Figure 8 respectively, but the performance for each model varied significantly depending on the language. The differences in IoU scores between similar languages (e.g., Swedish and Danish) and distant ones (e.g., Chinese and Tamil) emphasize the importance of tailoring models to handle diverse linguistic structures effectively.

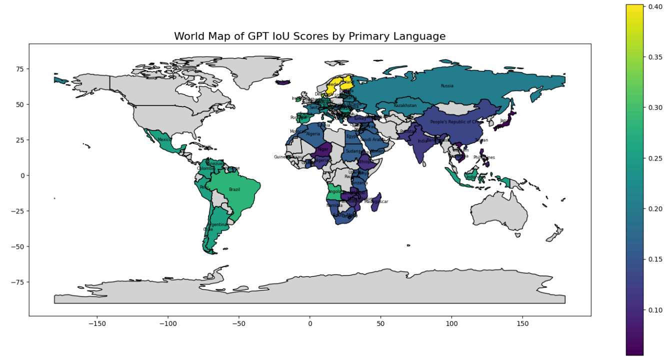


Figure 6: World map showing ChatGPT 4o mini IoU scores for each language mapped to the nations that speak it as their primary language.

### 4.2 Lexical Similarity vs. IoU Scores

Figure 4 shows a comparison of IoU scores for each language with its lexical similarity to English, according to Ethnologue [14]. A clear positive correlation emerged, with languages lexically closer to English, such as German and Dutch, achieving higher IoU scores. This trend aligns with expectations, as these languages share vocabulary and grammatical features with English, making translation and reasoning tasks less challenging.

However, there were notable outliers. For example, Bulgarian achieved a higher IoU score than expected based on its lexical similarity to English, while Afrikaans performed slightly worse than anticipated. These deviations suggest that factors beyond lexical similarity, such as training data quality and cultural context, also influence multilingual reasoning performance. This analysis underscores the complexity of developing LLMs that excel across diverse languages, highlighting the need for further research into language-specific optimizations.

### 4.3 IoU Scores vs. Number of Speakers

The analysis of IoU scores against the number of speakers of each tested language with more than 100 million global speakers, shown in Figure 11, revealed interesting trends. Languages with large speaker populations, such as Chinese, Hindi, and Spanish, generally performed well, benefiting from extensive training data in these languages. However, there were exceptions, such as Vietnamese and Japanese, which had relatively low IoU scores despite their large speaker bases. This finding suggests that while the quantity of training data is important, the quality and diversity of the data are equally critical in ensuring robust performance.

Conversely, smaller languages such as Dutch and Swedish performed exceptionally well, likely due to their syntactic similarity to English and strong representation in training datasets. This highlights the need for a balanced approach in dataset construction, ensuring that both widely spoken and syntactically diverse languages are adequately represented.

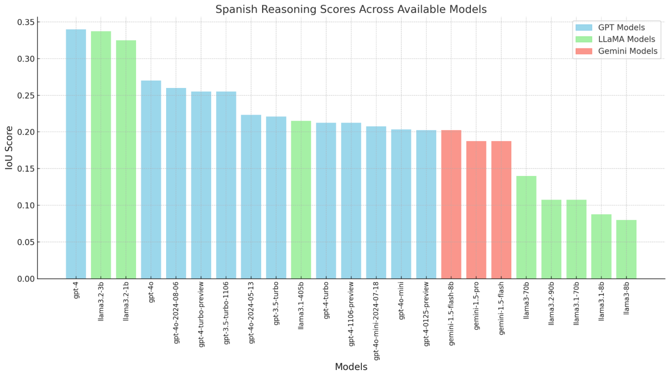


Figure 7: IoU scores across all tested models using Spanish with limit=3, subset=30 and iterations=1, color-coordinated by model family (among GPT, Llama and Gemini).

### 4.4 Rank Correlation Analysis

Rank correlation metrics, including Spearman and Kendall Tau coefficients, provided additional insights into multilingual reasoning consistency. Figures 3 and 10 illustrate that GPT models excelled in maintaining consistent keyword rankings across languages, followed closely by Gemini and LLaMA models. For example, languages such as Swedish, Danish, and German exhibited strong rank correlation scores, reflecting their syntactic similarity to English and effective handling by the models.

Interestingly, correlation scores for some non-Indo-European languages, such as Chinese and Arabic, were notably lower. This discrepancy likely stems from the challenges of aligning reasoning tasks across languages with vastly different grammatical structures and cultural contexts. The higher performance of Gemini in these cases highlights its potential to address such challenges due to its training corpus and architectural design, although further refinements are necessary.

It is also worth noting that near-zero Spearman rank correlation values, as seen for some languages, may indicate a lack of sufficient matching keywords to establish meaningful correlations—whether positive or negative. This scenario could suggest that the models struggled to generate coherent or semantically aligned keywords in those languages, which may be an even greater concern than observing a negative correlation. This observation underscores the importance of not only achieving high rank correlation values but also ensuring that models generate adequate and contextually relevant outputs for effective multilingual reasoning.

### 4.5 IoU and Rank Correlation Trends Across Word Limits

Figure 5 shows the effect of word limits on IoU scores, which reveals a clear trend: increasing the word limit improved IoU scores across all models and languages. For example, Spanish saw a significant jump in IoU scores as the word limit increased from 2 to 10, with scores nearly doubling in some cases. This result is likely due to the fact that, given more keyword options from relatively short inputs, the models are more likely to pick the same words even when selecting at random. The closer the keyword limit is to the number of words in the input, the higher the IoU score will inevitably be.

However, this improvement was not uniform across all languages. For low-resource or morphologically complex languages, such as Kinyarwanda and Tamil, the benefits of higher word limits were less pronounced. This indicates that while increasing the word limit can enhance performance, other factors, such as better training data and language-specific optimizations, are equally important.

### 4.6 World Map of IoU Scores by Primary Language

Figure 6 shows a world map visualizing countries according to the IoU score of Chat GPT 4o Mini using their official spoken language (sourced from the CIA factbook on world languages) [13]. English-speaking nations and countries with an official language that was not tested were left uncolored. Using this geographic perspective, clusters of languages with similar IoU scores can be observed based on world region. Western European languages, such as Dutch, Swedish, and German, dominated the higher IoU ranges, reflecting their proximity to English in both geography and linguistic structure. In contrast, languages from East Asia, Sub-Saharan Africa, and Southeast Asia exhibited lower IoU scores, highlighting the challenges of adapting models to handle linguistic diversity effectively. Two notable exceptions are Angola, which is a Portuguese-speaking country, and Indonesia, as Indonesian scored much higher across all tested models than any other Asian language.

### 4.7 Performance Across Models

The comparison of IoU and rank correlation scores across models revealed significant differences in multilingual reasoning capabilities. GPT models consistently outperformed LLaMA and Gemini in IoU and rank correlation metrics, particularly for Indo-European languages. However, Gemini demonstrated stronger performance for non-Indo-European languages, such as Chinese and Arabic, indicating its suitability for handling diverse linguistic structures.

Interestingly, the gap between GPT and Gemini narrowed as the complexity of the reasoning task increased, suggesting that Gemini's architecture is better equipped to handle challenging scenarios. LLaMA, while competitive for some languages, generally lagged behind GPT and Gemini, highlighting the need for further refinement of its multilingual capabilities.

Figure 7 shows the average IoU score for spanish keywords across all currently available LLM models within the ChatGPT, LLaMA and Gemini families. While the top 3 best-performing models are ChatGPT 4, LLaMA 3.2 3b and LLaMA 3.2 1b, most of the high performing models came from the ChatGPT family, with each available Gemini model presenting similar results to the worst-performing GPT model (GPT 4 0125 preview). This suggests that, on average, the GPT models are better equipped to handle complex multilingual reasoning than every Gemini model and most LLaMA models. The LLaMA 3.2 models, however, are remarkably capable at handling reasoning in non-English languages, far eclipsing the capabilities of previous LLaMA models.

## V. CONCLUSION

The results of this study provide insights into the performance of LLMs across a diverse range of languages and reasoning tasks. By analyzing metrics such as IoU scores and rank correlation coefficients (Spearman and Kendall Tau), this research highlights the strengths and weaknesses of models within the GPT, Gemini, and LLaMA families in maintaining reasoning performance in multilingual contexts.

One key takeaway is the significant variation in IoU scores across languages. Languages that are lexically and syntactically similar to English, such as Dutch and Swedish, consistently achieved higher scores, benefiting from shared grammatical structures and vocabulary. Conversely, linguistically distant languages like Chinese, Vietnamese, and Kinyarwanda exhibited lower scores, reflecting the challenges of adapting reasoning tasks to languages with different structures and cultural contexts. These results underscore the importance of linguistic similarity, but also highlight the pivotal roles of training corpus quality, model design, and the availability of high-quality parallel data.

Rank correlation analysis provided additional nuances in model performance. GPT models excelled in maintaining consistent keyword rankings across languages, while Gemini demonstrated greater adaptability for non-Indo-European languages. This suggests that different architectures have distinct strengths depending on the linguistic context. For instance, Gemini's ability to handle complex grammatical structures and cultural nuances in languages like Chinese and Arabic highlights the potential of hybrid approaches that combine strengths of multiple architectures.

In summary, this research contributes to understanding multilingual reasoning performance in LLMs, offering actionable insights for developing more inclusive and robust models. Prioritizing the inclusion of diverse languages in training datasets, refining metrics for evaluating multilingual reasoning, and exploring fine-tuning strategies for linguistically complex languages will pave the way for the next generation of global LLMs. By bridging the performance gap between high-resource and low-resource languages, future LLMs can ensure equitable benefits across linguistic and cultural boundaries.

## VI. PROJECT CODE

github.com/zignago/multilingual-llm

## REFERENCES

[1] Fayyaz, M., Yin, F., Sun, J., & Peng, N. (2024). Evaluating Human Alignment and Model Faithfulness of LLM Rationale. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2407.00219

[2] Mendonça, J., Pereira, P., Moniz, H., Carvalho, J. P., Lavie, A., & Trancoso, I. (2023). Simple LLM prompting is state-of-the-art for robust and multilingual dialogue evaluation. arXiv preprint arXiv:2308.16797.

[3] Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., … Wu, Y. (2023). How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2301.07597

[4] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., … Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2303.12712

[5] Guerreiro, N. M., Alves, D., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., & Martins, A. F. T. (2023). Hallucinations in Large Multilingual Translation Models. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2303.16104

[6] Huynh, J., Jiao, C., Gupta, P., Mehri, S., Bajaj, P., Chaudhary, V., & Eskenazi, M. (2023). Understanding the Effectiveness of Very Large Language Models on Dialog Evaluation. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2301.12004

[7] Yuan, F., Yuan, S., Wu, Z., & Li, L. (2024). How Vocabulary Sharing Facilitates Multilingualism in LLaMA? arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2311.09071

[8] Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., … Li, L. (2024). Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2304.04675

[9] Treviso, M., Guerreiro, N. M., Agrawal, S., Rei, R., Pombal, J., Vaz, T., … Martins, A. F. T. (2024). xTower: A Multilingual LLM for Explaining and Correcting Translation Errors. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2406.19482

[10] Wang, B., Xu, C., Wang, S., Gan, Z., Cheng, Y., Gao, J., … Li, B. (2022). Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2111.02840

[11] Vázquez, R., Raganato, A., Tiedemann, J., & Creutz, M. (2019). Multilingual NMT with a Language-Independent Attention Bridge. Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). doi:10.18653/v1/w19-4305

[12] Smirnov, A. V., Teslya, N., Shilov, N., Frank, D., Minina, E., & Kovacs, M. (2022). Comparative Analysis of Neural Translation Models based on Transformers Architecture. In ICEIS (1) (pp. 586-593).

[13] Central Intelligence Agency. (n.d.). *Field Listing - Languages*. Central Intelligence Agency.

[14] Collin, R. O. (2010). Ethnologue. Ethnopolitics, 9(3-4), 425-432.

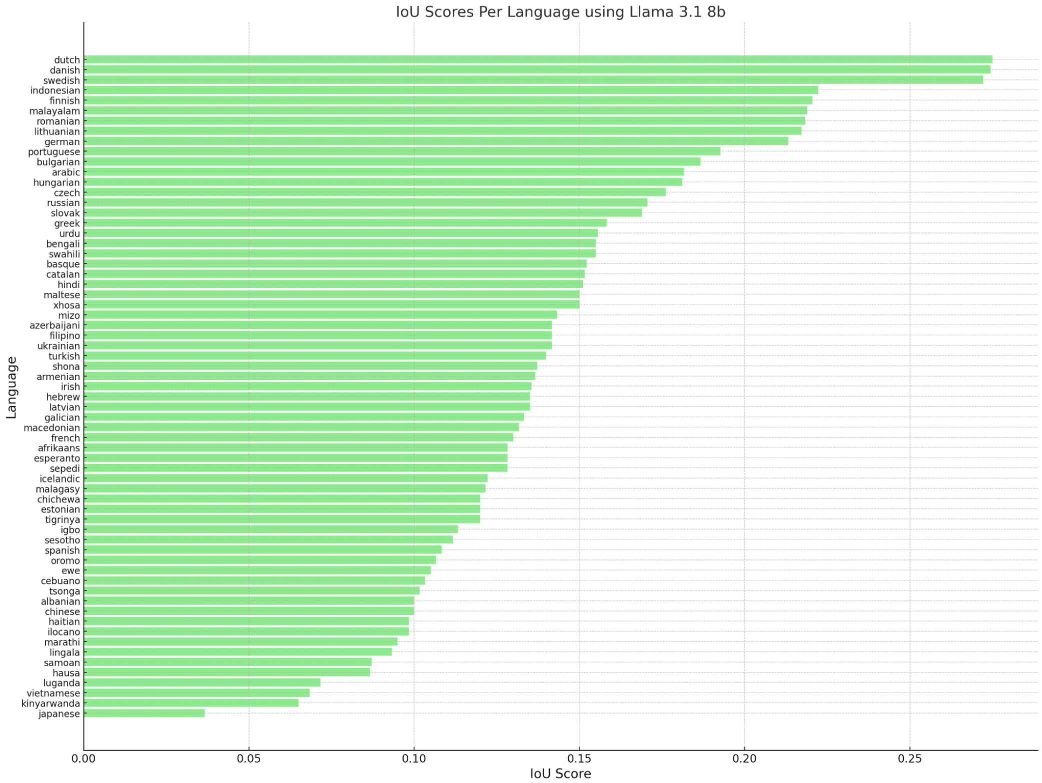


Figure 8: IoU scores for all tested languages using Llama 3.1 8b

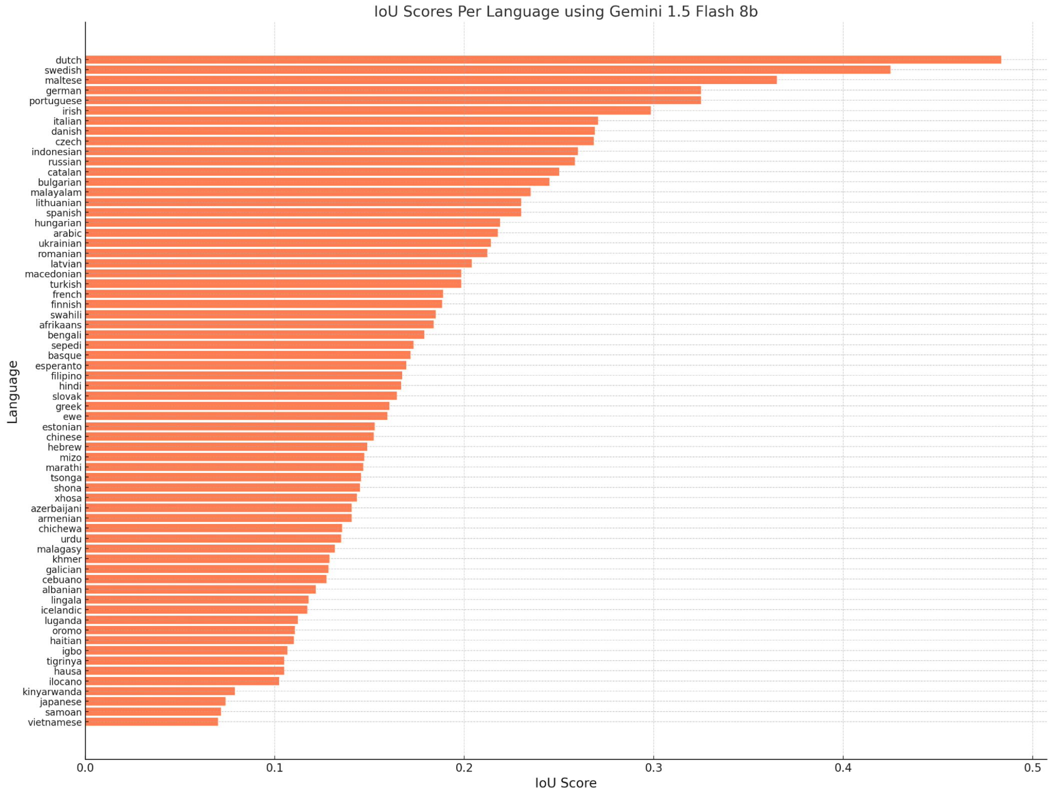


Figure 9: IoU scores for all tested languages using Gemini 1.5 Flash 8b

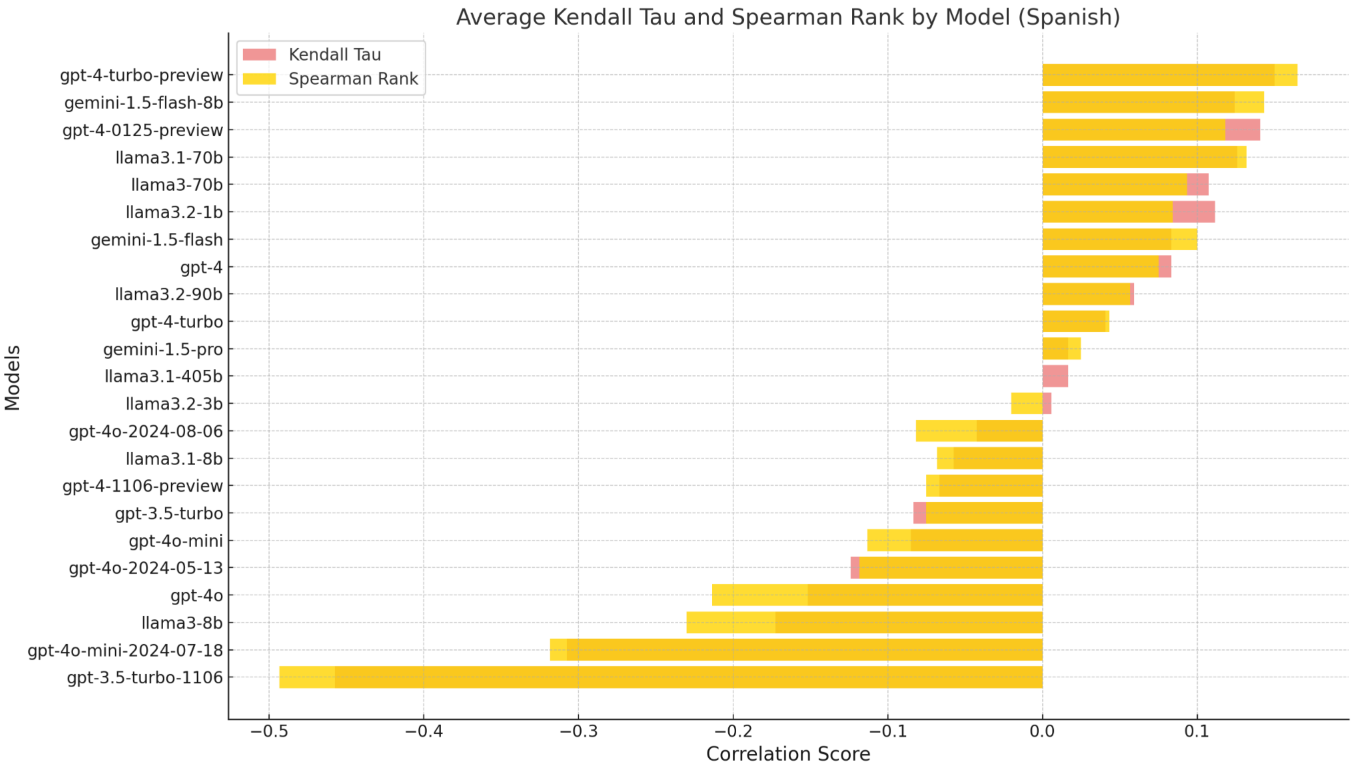


Figure 10: Average Kendall Tau and Spearman Rank scores for each tested model using Spanish

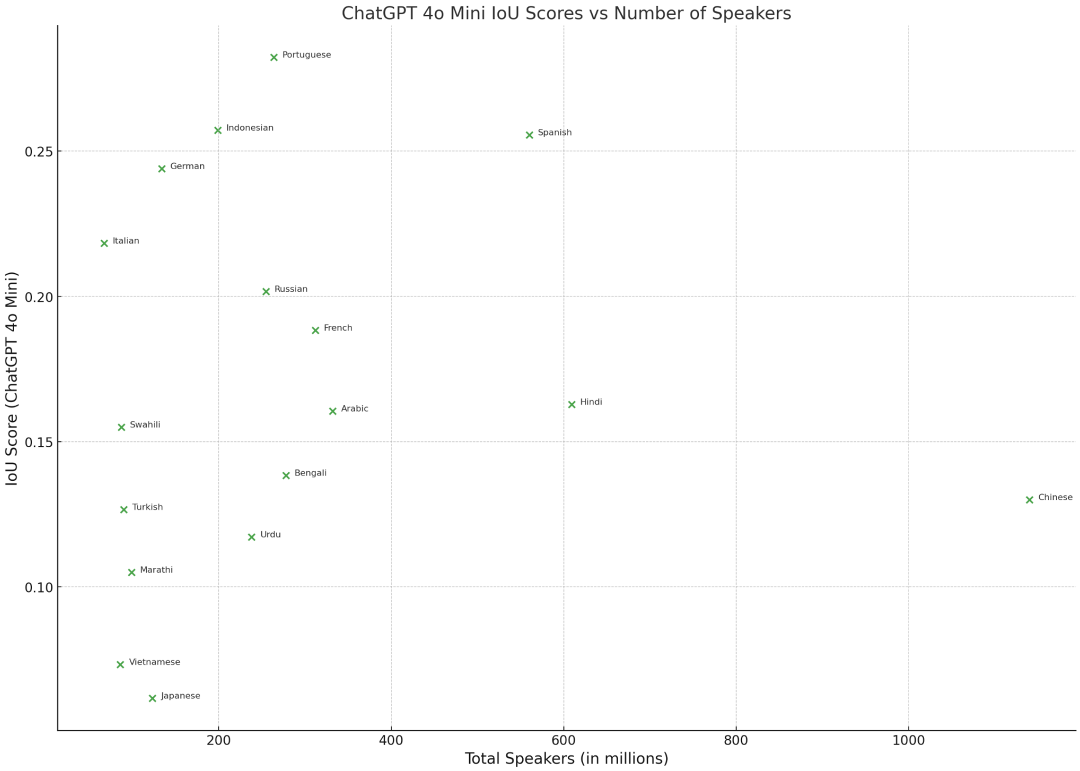


Figure 11: ChatGPT 4o Mini IoU Scores for each tested language with >=100 million worldwide speakers compared to the estimated number of total speakers.