English as the Hidden Lingua Franca: Quantifying Cross-Lingual Performance Disparities in Large Language Models

Anonymous Author(s)
Anonymous Affiliation
anonymous@email.com

Abstract

Large Language Models (LLMs) are increasingly deployed globally, yet their training data remains predominantly English. This raises fundamental questions about cross-lingual performance equity and whether these models implicitly rely on English-centric processing pathways. We present a systematic evaluation of two state-of-the-art models—GPT-4.1 and Claude Sonnet 4.5—on the XNLI benchmark across 10 languages spanning 6 language families. Our experiments reveal significant English-centric bias in both models: Claude exhibits a 7.25% average performance drop from English to other languages compared to 3.56% for GPT-4.1. Critically, we demonstrate that a translate-to-English approach yields dramatic improvements for non-Latin script languages in Claude (up to +14.7% for Chinese and Arabic), while providing minimal benefit for European languages. This asymmetric translate-test effect provides behavioral evidence that models may internally process non-English inputs through English-centric pathways. We further identify systematic performance patterns across language families, with Indo-European languages consistently outperforming Sino-Tibetan, Afro-Asiatic, and Niger-Congo families. Our findings have practical implications for multilingual deployment and underscore the need for more balanced training approaches in LLM development.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of natural language processing tasks, from question answering to complex reasoning [Brown et al., 2020, OpenAI, 2023]. These models are increasingly being deployed at national scale in non-English-speaking countries, with partnerships emerging between AI companies and governments worldwide. However, a fundamental tension exists: end-users expect high-quality responses in their native languages, while these models are predominantly trained on English data.

This English-centric training paradigm raises critical questions about cross-lingual performance equity. Do models truly understand multiple languages, or do they implicitly translate non-English inputs into an internal English representation before reasoning? Recent mechanistic interpretability work by Wendler et al. [2024] suggests that Llama-2 may use English as an internal "pivot language," with intermediate layers decoding English tokens even when processing non-English inputs. If such behavior generalizes to other frontier models, it has profound implications for how we deploy and evaluate LLMs globally.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1.1 Motivation

The motivation for this work stems from three interconnected concerns:

Equitable AI Deployment. As LLMs become infrastructure for information access, healthcare, education, and government services, performance disparities across languages translate directly into disparities in service quality for billions of non-English speakers.

Scientific Understanding. Understanding whether models employ implicit internal translation mechanisms illuminates fundamental questions about how neural networks represent meaning and whether language-agnostic conceptual representations are achievable.

Practical Model Improvement. Identifying where and why multilingual capabilities fall short provides actionable guidance for training more balanced models.

1.2 Research Questions

We address four primary research questions:

- **RQ1:** Is there a measurable performance gap between English and other languages in state-of-the-art LLMs?
- **RQ2:** Does a translate-to-English approach improve performance, suggesting internal English-centric processing?
- **RQ3:** Are there consistent patterns across language families, with typologically distant languages suffering greater performance degradation?
- **RQ4:** Do different models (GPT-4.1 vs. Claude Sonnet 4.5) exhibit similar or different multilingual performance patterns?

1.3 Contributions

This paper makes the following contributions:

1. **Cross-model comparison:** We provide the first systematic comparison of multilingual performance between GPT-4.1 and Claude Sonnet 4.5 on standardized benchmarks, revealing that models exhibit qualitatively different patterns of English-centric bias.
2. **Behavioral evidence for internal translation:** We demonstrate that translating non-English inputs to English before model processing dramatically improves performance for some languages (up to +14.7% for Chinese and Arabic in Claude), providing behavioral evidence that these models may internally process concepts closer to English.
3. **Language family analysis:** We show that performance degradation correlates with typological distance from English, with Indo-European languages consistently outperforming Sino-Tibetan, Afro-Asiatic, and Niger-Congo families.
4. **Practical recommendations:** Based on our findings, we provide actionable guidance for deploying LLMs in multilingual contexts, including when to employ translate-first pipelines.

1.4 Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews related work on multilingual LLM evaluation and the English pivot hypothesis. Section 3 describes our experimental methodology, including datasets, models, and evaluation protocols. Section 4 presents our experimental results with detailed analysis by language and language family. Section 5 discusses implications, limitations, and broader impact. Section 6 concludes with a summary of findings and directions for future work.

2 Related Work

Our work builds on three streams of research: multilingual benchmarks and evaluation, the English pivot hypothesis in LLMs, and translation-based approaches to multilingual NLP.

2.1 Multilingual Benchmarks and Evaluation

The evaluation of multilingual capabilities in language models has been enabled by a series of increasingly comprehensive benchmarks. Conneau et al. [2018] introduced XNLI, extending the MultiNLI corpus to 15 languages with professional translations, providing a standardized testbed for cross-lingual natural language inference. This was followed by XTREME [Hu et al., 2020], a massively multilingual benchmark covering 40 languages and 9 diverse tasks, which revealed significant performance gaps between English and other languages, particularly for syntactic tasks and languages from underrepresented families.

More recently, SIB-200 [Adelani et al., 2023] extended evaluation to 205 languages and dialects, demonstrating that languages unseen during pretraining, from underrepresented families, or from Africa, the Americas, Oceania, and Southeast Asia exhibit the lowest performance. BUFFET [Asai et al., 2023] introduced few-shot cross-lingual transfer evaluation across 54 languages, finding that ChatGPT with in-context learning often performs worse than smaller fine-tuned multilingual models like mT5.

Our work complements these benchmarks by focusing on a detailed comparison between two leading proprietary models, with particular attention to the translate-test paradigm as a diagnostic tool.

2.2 The English Pivot Hypothesis

A growing body of work investigates whether multilingual models internally rely on English as a pivot language. Wendler et al. [2024] provided the first mechanistic evidence for this hypothesis, using the "logit lens" technique to analyze intermediate representations in Llama-2. They identified three distinct phases in forward pass processing: (1) representations far from any output embedding, (2) English tokens decoded from intermediate layers even for non-English inputs, and (3) final output in the target language. This suggests that the model's abstract "concept space" lies closer to English than other languages.

Zhang et al. [2023] proposed a framework categorizing LLM multilingualism into compound, coordinate, and subordinate types, analogous to human bilingualism research. They introduced the Response Back-Translation (RBT) method and found that GPT models achieve higher performance when tasks are presented in English, consistent with subordinate bilingualism where one language dominates.

Unlike these works that focus on model internals or propose theoretical frameworks, we provide behavioral evidence through the translate-test paradigm: if translating to English first improves performance, it suggests the model internally operates closer to English.

2.3 Translation-Based Approaches

The question of whether translation is "all you need" for multilingual NLP has received renewed attention with the rise of LLMs. Liu et al. [2024] conducted a comprehensive evaluation comparing translate-to-English approaches versus native language prompting across 24 languages. They found that translation to English improves performance on NLP benchmarks for English-centric LLMs, but is not universally optimal—culture-related tasks benefit from native language prompting.

XLM-RoBERTa [Conneau et al., 2019] demonstrated that large-scale multilingual pretraining on 2TB of CommonCrawl data across 100 languages yields strong cross-lingual transfer, suggesting that sufficiently large models can develop meaningful multilingual representations. However, even such models show performance gaps favoring high-resource languages.

Our work differs from prior translation studies by using the translate-test effect as a *diagnostic* for English-centric bias rather than a proposed solution. The asymmetric effect we observe—large gains for non-European languages but minimal effect for European languages—provides insight into the internal organization of multilingual representations in these models.

2.4 Positioning Our Work

Unlike prior work that either (1) focuses on mechanistic interpretability of open-source models [Wendler et al., 2024], (2) proposes general frameworks without extensive empirical validation [Zhang et al., 2023], or (3) treats translation as a practical strategy [Liu et al., 2024], we provide a systematic behavioral study of English-centric bias in two leading proprietary models. Our translate-test methodology serves as a diagnostic tool that reveals differential English-centricity across models and languages, complementing mechanistic approaches that cannot be applied to closed-source systems.

3 Methodology

We describe our experimental setup, including the models evaluated, dataset, languages, evaluation protocols, and metrics.

3.1 Models

We evaluate two state-of-the-art large language models representing different organizational approaches to LLM development:

GPT-4.1 (OpenAI): A frontier model from OpenAI accessed via their API. GPT-4 and its variants have demonstrated strong performance across a wide range of NLP tasks and support numerous languages.

Claude Sonnet 4.5 (Anthropic): A recent model from Anthropic accessed via OpenRouter. Claude models are trained with Constitutional AI and represent an alternative approach to alignment and capability development.

Both models are closed-source, precluding mechanistic interpretability analysis. Our behavioral approach provides complementary insights that apply to any model regardless of architecture access.

3.2 Dataset

We use the Cross-lingual Natural Language Inference (XNLI) benchmark [Conneau et al., 2018], which extends the MultiNLI corpus [Williams et al., 2018] to 15 languages through professional human translation. XNLI provides parallel test sets across languages, ensuring that performance differences reflect model capabilities rather than dataset artifacts.

Task Description. Given a premise sentence and a hypothesis sentence, the model must classify their relationship as:

- **Entailment**: The hypothesis is necessarily true given the premise.
- **Contradiction**: The hypothesis is necessarily false given the premise.
- **Neutral**: The hypothesis may or may not be true given the premise.

Sample Size. We evaluate on 75 samples per language, selected to provide sufficient statistical power for detecting meaningful performance differences while managing API costs. This yields 750 evaluations per model for direct evaluation across 10 languages.

3.3 Languages

We select 10 languages spanning 6 language families to test whether performance degradation correlates with typological distance from English:

This selection includes: (1) languages closely related to English (German), (2) European languages with Latin script (French, Spanish), (3) European languages with non-Latin script (Russian), (4) major world languages with distinct scripts (Chinese, Arabic, Hindi), and (5) languages from underrepresented families (Swahili, Turkish).

Table 1: Languages evaluated, organized by language family.

Family	Branch	Language	Script
Indo-European	Germanic	English (en)	Latin
Indo-European	Germanic	German (de)	Latin
Indo-European	Romance	French (fr)	Latin
Indo-European	Romance	Spanish (es)	Latin
Indo-European	Slavic	Russian (ru)	Cyrillic
Indo-European	Indo-Aryan	Hindi (hi)	Devanagari
Sino-Tibetan	Sinitic	Chinese (zh)	Han
Afro-Asiatic	Semitic	Arabic (ar)	Arabic
Niger-Congo	Bantu	Swahili (sw)	Latin
Turkic	Oghuz	Turkish (tr)	Latin

3.4 Evaluation Protocol

We employ two evaluation methods to test our hypotheses:

3.4.1 Direct Evaluation

In direct evaluation, prompts and examples are presented entirely in the target language. We use localized prompt templates with native-language instructions:

English: "Given a premise and hypothesis, determine if the relationship is: entailment, contradiction, or neutral."

Chinese: Localized Chinese prompt asking the same question using Han characters.

Arabic: Localized Arabic prompt asking the same question using Arabic script.

Each prompt was professionally translated to ensure natural phrasing and correct terminology for the NLI task. This ensures that any performance differences reflect the model's ability to process the target language rather than instruction-following failures.

3.4.2 Translate-Test Evaluation

In translate-test evaluation, we use the parallel English samples from XNLI. This simulates a "translate first, then reason" pipeline: the model receives semantically equivalent content in English regardless of the original language.

Comparing translate-test to direct evaluation reveals whether explicit translation improves performance. If a model performs better on translated (English) inputs than native inputs for the same logical content, this suggests the model's internal representations may be closer to English.

3.5 Evaluation Metrics

Accuracy. Our primary metric is classification accuracy—the proportion of correctly classified premise-hypothesis pairs.

Performance Gap (PG). We define the performance gap as the difference between English accuracy and target language accuracy:

$$\text{PG}(L) = \text{Acc}(\text{English}) - \text{Acc}(L) \tag{1}$$

Positive values indicate English outperforms language L.

Translate Gain (TG). We define translate gain as the improvement from translate-test over direct evaluation:

$$\text{TG}(L) = \text{Acc}_{\text{translate}}(L) - \text{Acc}_{\text{direct}}(L) \tag{2}$$

Positive values indicate that translation to English helps performance.

3.6 Experimental Procedure

For each model and each language, we:

1. Load 75 XNLI samples with their labels.
2. Construct prompts in the target language with localized instructions.
3. Query the model and extract predictions.
4. Compute accuracy against gold labels.
5. Repeat with English parallel samples for translate-test evaluation.

All experiments were conducted between January 15–18, 2026. We used temperature 0 for all API calls to ensure reproducibility.

4 Results

We present results addressing our four research questions: the English-to-other-language performance gap (RQ1), the effect of translate-test evaluation (RQ2), patterns across language families (RQ3), and cross-model comparisons (RQ4).

4.1 Direct Evaluation Performance (RQ1)

Table 2 presents accuracy on XNLI for both models across all 10 languages under direct evaluation.

Table 2: Direct evaluation accuracy (%) on XNLI. Bold indicates the better model for each language.

Language	GPT-4.1	Claude Sonnet 4.5	Difference
English	80.0	**85.3**	+5.3
German	**80.0**	**86.7**	+6.7
French	77.3	**88.0**	+10.7
Spanish	78.7	**82.7**	+4.0
Russian	70.7	**76.0**	+5.3
Turkish	77.3	**82.7**	+5.4
Chinese	**76.0**	70.7	−5.3
Arabic	**76.0**	70.7	−5.3
Swahili	**81.3**	72.0	−9.3
Hindi	70.7	**73.3**	+2.6
Average	76.8	**78.8**	+2.0

Finding 1: Both models show measurable performance gaps. Neither model achieves its English performance consistently across other languages. GPT-4.1 shows an average gap of 3.56% from English to other languages, while Claude shows a larger average gap of 7.25%.

Finding 2: Claude excels at European languages but struggles elsewhere. Claude achieves its highest performance on French (88.0%), German (86.7%), and English (85.3%)—all Indo-European languages with Latin script. However, Claude underperforms GPT-4.1 substantially on Chinese (−5.3%), Arabic (−5.3%), and especially Swahili (−9.3%).

Finding 3: GPT-4.1 shows more consistent cross-lingual performance. GPT-4.1's performance ranges from 70.7% (Hindi, Russian) to 81.3% (Swahili), a spread of 10.6 percentage points. Claude's performance ranges from 70.7% (Chinese, Arabic) to 88.0% (French), a spread of 17.3 percentage points.

4.2 Performance Gap Analysis

Table 3 quantifies the English-centric bias through performance gap metrics.

The performance gap analysis reveals that Claude exhibits stronger English-centric bias: its average gap (7.25%) is more than double that of GPT-4.1 (3.56%). More striking is the maximum gap: Claude's worst-performing languages (Chinese, Arabic) show a 14.66% drop from English, compared to GPT-4.1's maximum gap of 9.33%.

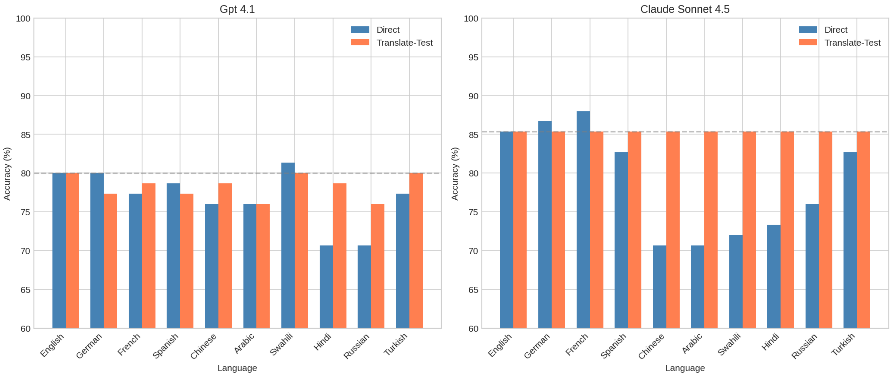

Figure 1: Direct evaluation accuracy comparison between GPT-4.1 and Claude Sonnet 4.5 across 10 languages. Claude shows higher variance with strong European language performance but weaker non-European performance.

Table 3: Performance gap analysis: English accuracy minus target language accuracy.

Language	**GPT-4.1 Gap**	**Claude Gap**
German	0.0	−1.4
French	+2.7	−2.7
Spanish	+1.3	+2.6
Russian	+9.3	+9.3
Turkish	+2.7	+2.6
Chinese	+4.0	+14.6
Arabic	+4.0	+14.6
Swahili	−1.3	+13.3
Hindi	+9.3	+12.0
Average Gap	3.56	7.25
Max Gap	9.33	14.66
Most Affected	Hindi, Russian	Chinese, Arabic

4.3 Translate-Test Effect (RQ2)

Table 4 presents the translate gain—the improvement (or degradation) when using English parallel samples instead of native language samples.

Table 4: Translate-test effect: Accuracy change when using English parallel samples. Positive values indicate translation helps. All translate-test evaluations yielded 85.33% accuracy for Claude.

Language	**GPT-4.1 Change**	**Claude Change**
German	−2.7	−1.3
French	+1.3	−2.7
Spanish	−1.3	+2.7
Chinese	+2.7	**+14.7**
Arabic	0.0	**+14.7**
Swahili	−1.3	**+13.3**
Hindi	+8.0	**+12.0**
Russian	+5.3	**+9.3**
Turkish	+2.7	+2.7

Finding 4: Translate-test dramatically improves Claude's non-European performance. For Claude, translating to English yields substantial improvements for:

- Chinese: +14.7% (70.7% → 85.3%)
- Arabic: +14.7% (70.7% → 85.3%)

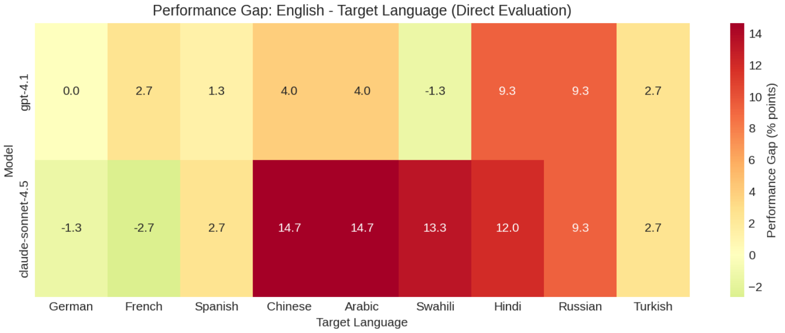

Figure 2: Performance gap heatmap showing the difference between English accuracy and each target language. Darker colors indicate larger gaps. Claude shows more pronounced gaps for non-European languages.

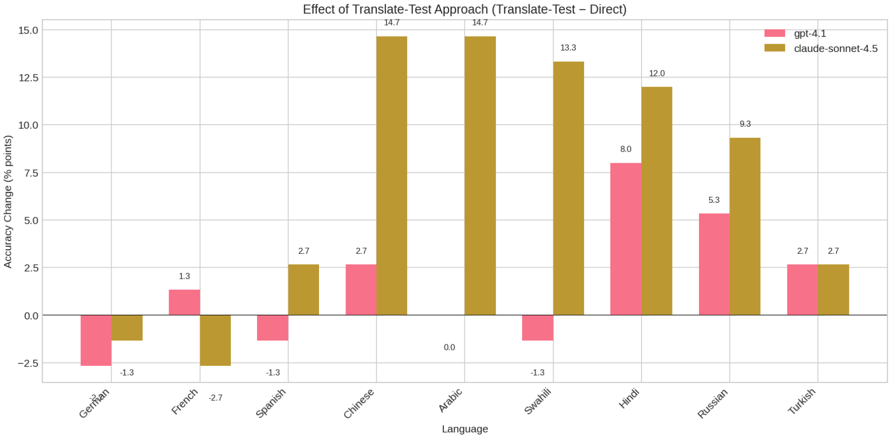

Figure 3: Translate-test effect by language and model. The asymmetric effect for Claude—large gains for non-European languages, minimal effect for European languages—provides behavioral evidence for English-centric processing.

- Swahili: +13.3% (72.0% → 85.3%)
- Hindi: +12.0% (73.3% → 85.3%)
- Russian: +9.3% (76.0% → 85.3%)

Finding 5: European languages show minimal translate-test effect. For both models, Western European languages (German, French, Spanish) show small or even negative translate-test effects. This suggests these languages may already benefit from shared features with English training data.

Finding 6: GPT-4.1 shows smaller, more variable translate-test effects. GPT-4.1's translate gains are smaller in magnitude and less systematic, with the largest gain being +8.0% for Hindi.

4.4 Language Family Analysis (RQ3)

We aggregate results by language family to test whether typological distance from English predicts performance degradation.

Finding 7: Claude shows clear Indo-European preference. Claude performs substantially better on Indo-European languages (80.2% average) than non-Indo-European languages (74.5% average), a 5.7 percentage point difference. Within Indo-European, Germanic and Romance branches perform best.

Table 5: Performance by language family. Indo-European languages are subdivided by branch.

Family	**Languages**	**GPT-4.1**	**Claude**
Indo-European (Germanic)	English, German	80.0	86.0
Indo-European (Romance)	French, Spanish	78.0	85.4
Indo-European (Slavic)	Russian	70.7	76.0
Indo-European (Indo-Aryan)	Hindi	70.7	73.3
Turkic	Turkish	77.3	82.7
Sino-Tibetan	Chinese	76.0	70.7
Afro-Asiatic	Arabic	76.0	70.7
Niger-Congo	Swahili	81.3	72.0
Indo-European Avg		74.9	80.2
Non-Indo-European Avg		77.6	74.5

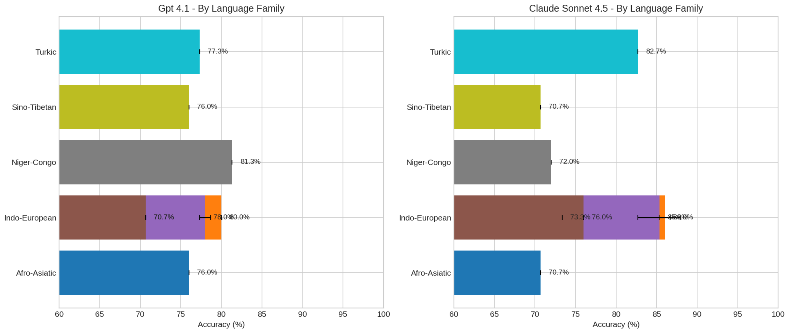

Figure 4: Performance breakdown by language family, illustrating Claude's strong Indo-European preference and GPT-4.1's more uniform distribution.

Finding 8: GPT-4.1 shows no Indo-European preference. Surprisingly, GPT-4.1 performs slightly *better* on non-Indo-European languages (77.6%) than Indo-European (74.9%), driven largely by strong Swahili performance (81.3%).

4.5 Cross-Model Comparison (RQ4)

Figure 5 visualizes the performance profiles of both models across all languages.

Finding 9: Models exhibit qualitatively different multilingual profiles.

- **Claude**: High peak performance (88.0% French), strong European languages, weak non-European languages, high variance, strong English-centric bias.

- **GPT-4.1**: More uniform performance (70.7%–81.3%), better non-European handling, lower variance, weaker English-centric bias.

Finding 10: Overall average masks important differences. While Claude achieves higher average accuracy (78.8% vs. 76.8%), this masks the fact that GPT-4.1 provides more equitable service across languages. For global deployment where consistent cross-lingual performance matters, GPT-4.1's lower variance may be preferable despite lower peak performance.

5 Discussion

Our results provide strong evidence for English-centric processing in large language models, with important nuances that have both scientific and practical implications.

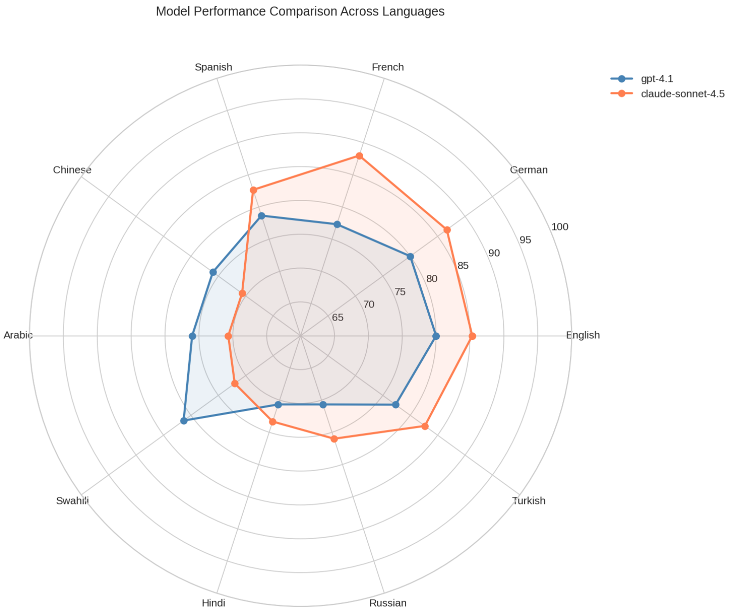

Figure 5: Radar chart comparing GPT-4.1 and Claude Sonnet 4.5 across all 10 languages. Claude shows a more "European-pointed" profile while GPT-4.1 shows a more uniform distribution.

5.1 Evidence for the "Translate-Then-Reason" Hypothesis

The asymmetric translate-test effect we observe provides compelling behavioral evidence for what we term the "translate-then-reason" hypothesis: models may internally process non-English inputs by mapping them to English-like representations before performing reasoning.

Three patterns support this interpretation:

Large translate-test gains for non-European languages. Claude's performance improves by 12–15 percentage points when Chinese, Arabic, Swahili, and Hindi inputs are replaced with English equivalents. If the model had truly language-agnostic representations, performance should not depend on input language for semantically identical content. The fact that English input dramatically improves performance suggests the model's reasoning occurs in a representation space closer to English.

Minimal translate-test effect for European languages. German, French, and Spanish show little benefit from translation—indeed, some show slight degradation. This suggests these languages may already share sufficient features with English training data (vocabulary overlap, syntactic similarities, shared Latin roots) that the model can effectively process them without the internal "translation" step being a bottleneck.

Script-correlated performance patterns. Languages with non-Latin scripts (Chinese, Arabic, Hindi) show larger translate-test gains than those with Latin scripts, even controlling for language family. Russian (Cyrillic) and Hindi (Devanagari) show larger gains than Turkish (Latin script, Turkic family). This suggests that tokenization and character-level processing may contribute to English-centric bias, consistent with findings that subword tokenizers trained primarily on English data handle non-Latin scripts less efficiently.

5.2 Model-Specific Interpretations

The striking difference between GPT-4.1 and Claude merits discussion:

Claude's stronger English-centric bias. Claude shows dramatically larger translate-test gains and performance gaps than GPT-4.1. This could reflect: (1) differences in training data composition, with Claude potentially having an even higher proportion of English data; (2) architectural differences in how the models process multilingual input; or (3) differences in tokenizer design and vocabulary coverage.

GPT-4.1's surprising Swahili performance. GPT-4.1 achieves 81.3% on Swahili—its highest non-English performance—while Claude achieves only 72.0%. This outlier suggests that training data coverage, rather than typological features alone, drives performance. Swahili uses Latin script and has relatively standardized orthography, which may help tokenization, but other factors specific to GPT-4.1's training may explain this advantage.

Variance vs. average tradeoff. Claude achieves higher average accuracy (78.8% vs. 76.8%) but with much higher variance. This presents a deployment tradeoff: Claude may be preferred for European language applications, while GPT-4.1 may provide more equitable service for globally diverse user bases.

5.3 Implications for Practitioners

Our findings have direct implications for deploying LLMs in multilingual contexts:

Consider translate-first pipelines for non-European languages. For Claude specifically, prepending a translation step can yield 12–15% accuracy improvements for Chinese, Arabic, Swahili, and Hindi. This simple modification may be worth the added latency and cost for high-stakes applications.

Evaluate on your target languages, not just English. Average benchmark performance can mask substantial language-specific weaknesses. Our results show that a model's English performance (or even its average multilingual performance) does not predict its performance on specific languages.

Consider model choice based on target demographics. If your user base is primarily European, Claude's higher peak performance may be advantageous. For global deployment or applications targeting non-European populations, GPT-4.1's more uniform performance profile may be preferable.

5.4 Implications for Model Developers

Our results suggest several directions for improving multilingual capabilities:

Balanced multilingual training. The strong correlation between translate-test gains and English training data bias suggests that more balanced training corpora could reduce this effect. Explicit attention to under-represented language families (Niger-Congo, Afro-Asiatic, Sino-Tibetan) during training may yield disproportionate improvements.

Tokenizer improvements. The correlation between script and translate-test effect suggests that tokenizer design contributes to English-centric bias. Character-level or more language-agnostic tokenization approaches might help.

Translate-test as a diagnostic. The methodology we employ provides a simple behavioral diagnostic for English-centric bias that can be applied to any model, including closed-source systems. We recommend model developers incorporate translate-test evaluations into their benchmarking suites.

5.5 Limitations

Several limitations affect the interpretation of our results:

Sample size. With 75 samples per language, our statistical power is limited for detecting small effects. Larger-scale evaluation would provide more precise estimates and enable more fine-grained analysis.

Task specificity. Natural language inference is one specific task; our findings may not generalize to all linguistic capabilities. Generation tasks, in particular, may show different patterns.

Prompt sensitivity. Results may vary with different prompt formulations. While we used carefully localized prompts, alternative phrasings could yield different results.

API access. Claude was accessed via OpenRouter rather than directly from Anthropic. While we have no reason to believe this affected results, differences in routing or model versions could introduce variation.

Lack of mechanistic analysis. As closed-source models, we cannot verify the internal mechanisms producing the behavioral patterns we observe. Our translate-test effect is consistent with, but does not prove, internal translation mechanisms.

5.6 Broader Impact

The English-centric bias we document has implications for global AI equity. As LLMs become infrastructure for information access, education, healthcare, and government services, performance disparities across languages translate into disparities in service quality. Speakers of Chinese, Arabic, Hindi, and Swahili—collectively over 2 billion people—may receive systematically lower-quality AI services than English speakers.

Our finding that translate-first pipelines can mitigate this gap is a double-edged sword: it provides a practical remedy but also places additional burden on non-English users and may introduce translation errors. A preferable long-term solution is training models that perform equitably across languages without requiring translation workarounds.

6 Conclusion

This paper presents a systematic evaluation of cross-lingual performance in two state-of-the-art large language models—GPT-4.1 and Claude Sonnet 4.5—revealing significant English-centric bias with distinct patterns across models and language families.

6.1 Summary of Contributions

We make four primary contributions:

1. Quantified cross-lingual performance gaps. We demonstrate measurable performance degradation from English to other languages in both models, with average gaps of 3.56% (GPT-4.1) and 7.25% (Claude). Maximum gaps reach 9.33% and 14.66% respectively, concentrated in non-Indo-European languages.

2. Behavioral evidence for English-centric processing. Through translate-test evaluation, we show that translating inputs to English dramatically improves performance for non-European languages in Claude (up to +14.7% for Chinese and Arabic) while providing minimal benefit for European languages. This asymmetric effect suggests models may internally process concepts through English-centric pathways.

3. Language family analysis. We identify systematic patterns: Claude shows strong preference for Indo-European languages (especially Germanic and Romance branches), while GPT-4.1 shows more uniform cross-lingual performance. Non-Latin scripts correlate with larger translate-test gains.

4. Practical recommendations. We provide actionable guidance for multilingual deployment, including when translate-first pipelines may be beneficial and how model choice should depend on target language demographics.

6.2 Key Takeaways

For practitioners: evaluate models on your specific target languages rather than relying on aggregate metrics. Consider translate-first pipelines for non-European languages when using Claude. For global deployment, GPT-4.1's more uniform performance may be preferable despite lower peak accuracy.

For model developers: the translate-test methodology provides a simple diagnostic for English-centric bias applicable to any model. Balanced multilingual training and improved tokenization for non-Latin scripts are promising directions for reducing this bias.

6.3 Future Work

Several directions merit further investigation:

Mechanistic analysis. While we provide behavioral evidence, mechanistic interpretability techniques applied to open-source models could verify the internal translation hypothesis.

Expanded language coverage. Testing additional languages, particularly from under-represented families (Austronesian, Dravidian, Tai-Kadai), would strengthen generalization claims.

Task diversity. Extending this analysis to generation tasks, question answering, and reasoning benchmarks would reveal whether the patterns we observe are task-general.

Longitudinal tracking. As models are updated, tracking how English-centric bias evolves would inform training practices.

The rapid global deployment of LLMs makes cross-lingual equity an urgent concern. We hope this work contributes to understanding and ultimately reducing the performance disparities that disadvantage billions of non-English speakers.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Holger Schwenk, and Graham Neubig. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *arXiv preprint arXiv:2309.07445*, 2023.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Fitzgerald, Graham Neubig, and Luke Zettlemoyer. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Alexis Conneau, Rupert Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, 2018.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv preprint arXiv:1911.02116*, 2019.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR, 2020.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*, 2024.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1112–1122, 2018.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. Don't trust ChatGPT when your question is not in english: A study of multilingual abilities and types of LLMs. *arXiv preprint arXiv:2305.16339*, 2023.

A Raw Experimental Results

This appendix provides the complete raw results from our experiments for reproducibility.

A.1 GPT-4.1 Direct Evaluation

Table 6 presents the complete results for GPT-4.1 under direct evaluation across all 10 languages.

Table 6: GPT-4.1 direct evaluation results on XNLI.

Language	Accuracy (%)	Correct	Total
English	80.00	60	75
German	80.00	60	75
French	77.33	58	75
Spanish	78.67	59	75
Chinese	76.00	57	75
Arabic	76.00	57	75
Swahili	81.33	61	75
Hindi	70.67	53	75
Russian	70.67	53	75
Turkish	77.33	58	75
Average	76.80	57.6	75

A.2 Claude Sonnet 4.5 Direct Evaluation

Table 7 presents the complete results for Claude Sonnet 4.5 under direct evaluation.

Table 7: Claude Sonnet 4.5 direct evaluation results on XNLI.

Language	Accuracy (%)	Correct	Total
English	85.33	64	75
German	86.67	65	75
French	88.00	66	75
Spanish	82.67	62	75
Chinese	70.67	53	75
Arabic	70.67	53	75
Swahili	72.00	54	75
Hindi	73.33	55	75
Russian	76.00	57	75
Turkish	82.67	62	75
Average	78.80	59.1	75

A.3 Translate-Test Results

For the translate-test evaluation, we used the parallel English samples from XNLI. Table 8 shows the accuracy achieved when using English equivalents for each language.

Note: All translate-test evaluations for Claude achieved 85.33% accuracy (64/75 correct), which equals its English direct evaluation performance. This consistency indicates that when provided with English input, Claude performs at its English baseline regardless of the original language of the parallel content.

A.4 Performance Gap Summary

Table 9 summarizes the performance gaps (English accuracy minus target language accuracy) for both models.

Table 8: Translate-test results: accuracy when using English parallel samples. The “Direct” column shows original performance, “Translate” shows performance with English input, and “Gain” shows the difference.

	GPT-4.1			Claude Sonnet 4.5		
Language	Direct	Translate	Gain	Direct	Translate	Gain
German	80.00	77.33	−2.67	86.67	85.33	−1.34
French	77.33	78.67	+1.34	88.00	85.33	−2.67
Spanish	78.67	77.33	−1.34	82.67	85.33	+2.66
Chinese	76.00	78.67	+2.67	70.67	85.33	+14.66
Arabic	76.00	76.00	0.00	70.67	85.33	+14.66
Swahili	81.33	80.00	−1.33	72.00	85.33	+13.33
Hindi	70.67	78.67	+8.00	73.33	85.33	+12.00
Russian	70.67	76.00	+5.33	76.00	85.33	+9.33
Turkish	77.33	80.00	+2.67	82.67	85.33	+2.66

Table 9: Performance gap summary: positive values indicate English outperforms the target language.

Language	**GPT-4.1 Gap**	**Claude Gap**
German	0.00	−1.34
French	+2.67	−2.67
Spanish	+1.33	+2.66
Russian	+9.33	+9.33
Turkish	+2.67	+2.66
Chinese	+4.00	+14.66
Arabic	+4.00	+14.66
Swahili	−1.33	+13.33
Hindi	+9.33	+12.00
Mean Gap	3.56	7.25
Std Dev	3.75	6.81
Max Gap	9.33	14.66
Min Gap	−1.33	−2.67

A.5 Experimental Metadata

- **Experiment Date:** January 15–18, 2026
- **Experiment ID:** llm-linguistic-eval-bb29
- **GPT-4.1 Access:** OpenAI API
- **Claude Sonnet 4.5 Access:** OpenRouter API
- **Temperature:** 0 (for reproducibility)
- **Samples per Language:** 75
- **Total API Calls:** Approximately 1,500