The Disappearing English Advantage: Cross-Language Equity in Frontier Large Language Models

Anonymous Author(s)

Abstract

Large language models are deployed globally, yet evaluations of their multilingual capabilities remain predominantly English-centric. We present a systematic evaluation of two frontier LLMs—GPT-4.1 and CLAUDE SONNET 4—across 8 languages, 3 prompting strategies, and 2 standardized benchmarks (MGSM math reasoning and BELEBELE reading comprehension), comprising 4,800 real API calls. Our central finding is that **the English performance advantage has largely disappeared in frontier models**: CLAUDE SONNET 4 achieves near-perfect cross-language equity (democratization score of 0.97) and actually scores higher on German (0.98) and Chinese (0.96) than English (0.92) for math reasoning. We further find that explicit English-pivoting strategies—self-translation and English chain-of-thought—provide no meaningful benefit for either model once we control for the chain-of-thought reasoning effect, contradicting earlier findings on GPT-3.5/4 and Llama-2. Language resource level does not predict performance in our data ($\rho = -0.09$ to 0.69, all $p > 0.08$). These results suggest that the implicit English-processing bottleneck documented in earlier work has been substantially mitigated in 2025–2026 frontier models, with important implications for equitable global deployment.

1 Introduction

Large language models are no longer English-only tools. They power government services in Estonia, educational platforms across Southeast Asia, and customer-facing products in virtually every major language market. Yet a fundamental question remains open: **do these models actually work as well in Swahili as they do in English?**

Prior mechanistic work suggests they should not. Wendler et al. [2024] showed that Llama-2 uses English-biased internal representations even when processing non-English inputs, operating through a three-phase pipeline: encode in the source language, reason in an English-biased concept space, then decode back to the target language. Zhang et al. [2024] confirmed this pattern through neuron-level analysis, finding that deactivating just 0.13% of language-specific neurons in Vicuna drops multilingual performance by 99%. Behavioral studies reinforced these findings: Etxaniz et al. [2023] demonstrated that "self-translating" inputs to English before solving tasks improved accuracy by 2–3.5 points across multiple benchmarks, and Shi et al. [2022] showed that English chain-of-thought prompting outperforms native-language reasoning even on non-English inputs.

However, these results were established on earlier model generations—GPT-3.5/4, Llama-2, and XGLM—using data collected in 2022–2023. In the intervening years, frontier models have undergone substantial improvements in multilingual training data, reinforcement learning from human feedback in multiple languages, and architectural refinements. **Whether the English-processing bottleneck persists in 2025–2026 frontier models is an open empirical question.**

We address this question through a controlled factorial experiment evaluating GPT-4.1 and CLAUDE SONNET 4 across 8 languages spanning three resource tiers, 3 prompting strategies (direct, self-translate, and English chain-of-thought), and 2 standardized benchmarks (MGSM for mathematical reasoning and BELEBELE for reading comprehension). Our design totals 4,800 real API calls, enabling systematic comparison across all conditions.

Our results paint a strikingly different picture from prior work:

- We find that CLAUDE SONNET 4 achieves **near-perfect cross-language equity**, with a democratization score of 0.97 on BELEBELE and no statistically significant English advantage on MGSM direct inference ($p = 0.17$).
- We show that **English-pivoting provides no benefit** for CLAUDE SONNET 4 on either benchmark, and that the apparent benefit for GPT-4.1 is attributable to chain-of-thought reasoning rather than translation.
- We demonstrate that **language resource level does not predict performance** in our data (all Spearman correlations non-significant), suggesting frontier models have made substantial progress on lower-resource languages.
- We identify a **prompt-format sensitivity** in GPT-4.1 that produces artifactually low direct-inference scores, highlighting the importance of controlling for reasoning scaffolding when evaluating multilingual performance.

2 Related Work

Internal language representations in LLMs. A growing body of work investigates how multilingual models internally process non-English inputs. Wendler et al. [2024] applied logit lens analysis to Llama-2 and identified a three-phase processing pipeline: early layers build language-agnostic features, middle layers operate in an English-biased concept space, and final layers decode to the target language. Zhang et al. [2024] proposed the Multilingual Workflow (MWork) hypothesis and developed PLND (Parallel Language-specific Neuron Detection) to identify language-specific neurons, finding that self-attention layers in middle layers decrease activity (English reasoning) while feed-forward neurons remain consistent (multilingual knowledge storage). These mechanistic findings establish the theoretical basis for our behavioral investigation: if models reason internally in English, then explicitly translating to English should help.

Self-translation and cross-lingual prompting. Several strategies exploit the English-biased internals of LLMs to improve multilingual performance. Etxaniz et al. [2023] introduced the self-translate approach, showing that using the LLM itself to translate inputs to English before task-solving consistently improved accuracy by 2–3.5 points on XGLM and Llama models. Huang et al. [2023] proposed Cross-Lingual-Thought (XLT) prompting, achieving over 10-point improvements on MGSM and introducing the *democratization score* to measure cross-language equity. Shi et al. [2022] demonstrated that English chain-of-thought prompting works effectively across languages, outperforming native-language reasoning—a finding that motivated the MGSM benchmark. Our work tests whether these English-pivoting benefits persist in models trained two to three years later.

Multilingual benchmarking of LLMs. Ahuja et al. [2023] provided the first comprehensive multilingual evaluation of generative LLMs, benchmarking GPT-3.5, GPT-4, and BLOOMZ across 16 datasets and 70 languages, finding significant English–non-English gaps of 10–30% especially for low-resource languages. Bandarkar et al. [2023] introduced the BELEBELE benchmark with parallel reading comprehension passages in 122 language variants, enabling controlled cross-lingual comparison. Hu et al. [2020] established XTREME as a foundational multilingual multi-task benchmark covering 9 tasks and 40 languages. Our work updates these evaluations to 2025–2026 frontier models and combines benchmark performance with prompting strategy analysis.

Chain-of-thought reasoning. Wei et al. [2022] showed that chain-of-thought prompting dramatically improves reasoning in LLMs. This finding is directly relevant to our work because several English-pivoting strategies (e.g., self-translate, English CoT) confound language effects with reasoning effects: they add step-by-step reasoning that is absent from direct inference. We explicitly control for this confound in our experimental design and analysis.

Tier	Languages	Scripts	Codes
HIGH	English, Chinese, German, French	Latin, CJK, Latin, Latin	en, zh, de, fr
MEDIUM	Russian, Japanese	Cyrillic, CJK	ru, ja
LOW	Swahili, Bengali/Hindi	Latin, Devanagari	sw, bn/hi

Table 1: Language selection across resource tiers. Bengali is used for MGSM and Hindi for BELEBELE in the low-resource tier based on benchmark availability.

3 Methodology

3.1 Experimental Design

We use a $2 \times 3 \times 8 \times 2$ factorial design crossing models, prompting strategies, languages, and tasks. Each cell contains 50 randomly sampled items (seed=42), yielding 4,800 total API calls.

3.2 Models

We evaluate two frontier LLMs representing different providers and training paradigms:

- GPT-4.1 (OpenAI), accessed via the OpenAI API with temperature 0.
- CLAUDE SONNET 4 (Anthropic), accessed via OpenRouter with temperature 0.

Both models represent the state of the art as of early 2026. We use deterministic decoding (temperature 0) to ensure reproducibility.

3.3 Benchmarks

MGSM (Multilingual Grade School Math). Introduced by Shi et al. [2022], MGSM contains 250 grade-school math word problems per language across 11 languages. Each problem requires multi-step arithmetic reasoning. We evaluate using exact match on the final integer answer.

BELEBELE (Reading Comprehension). Introduced by Bandarkar et al. [2023], BELEBELE provides 900 parallel reading comprehension questions per language across 122 language variants. Each question offers four multiple-choice options. We evaluate using correct option selection.

For both benchmarks, we randomly sample 50 items per language to balance experimental coverage with API cost constraints.

3.4 Languages

We select 8 languages across three resource tiers, as shown in table 1. The tiers reflect approximate training data availability in typical LLM pretraining corpora. We note that MGSM and BELEBELE have slightly different language coverage in the low-resource tier: we use Bengali for MGSM and Hindi for BELEBELE, as these are the available options in each benchmark.

3.5 Prompting Strategies

We test three prompting strategies that vary in the degree of explicit English involvement:

1. **DIRECT:** The problem is presented in its original language. The model is instructed to return only the final answer with no reasoning steps.
2. **SELF-TRANSLATE:** The model is first asked to translate the problem to English, then solve it step-by-step, and finally return the answer.
3. **ENGLISH COT:** The model is instructed to reason step-by-step in English about the problem (presented in its original language) and return the final answer.

An important design consideration is that DIRECT omits chain-of-thought reasoning, while both SELF-TRANSLATE and ENGLISH COT include it. This means any observed lift from these strategies confounds language effects with reasoning effects. We address this confound analytically by

Model	Strategy	EN	ZH	DE	FR	RU	JA	SW	BN	Avg
GPT-4.1	DIRECT	.640	.520	.560	.680	.700	.500	.640	.560	.600
	SELF-TRANSLATE	.600	.940	.960	.860	.960	.860	.820	.900	.862
	ENGLISH COT	.940	.940	.960	.900	.940	.880	.840	.920	.915
CLAUDE SONNET 4	DIRECT	.920	.960	**.980**	.940	.940	.900	.900	.940	**.935**
	SELF-TRANSLATE	.920	.940	.960	.940	.940	.920	.880	.940	.930
	ENGLISH COT	**1.000**	.960	.940	.960	.960	.900	.840	.940	.938

Table 2: MGSM accuracy by model, strategy, and language. CLAUDE SONNET 4 achieves strong performance under DIRECT inference with no chain-of-thought, while GPT-4.1 requires explicit reasoning steps. Bold indicates best per-column result.

comparing SELF-TRANSLATE against ENGLISH COT on non-English inputs—since both include chain-of-thought reasoning, any difference between them isolates the translation component.

3.6 Evaluation Metrics

Accuracy. Primary metric: exact match for MGSM (parsed integer), correct option for BELEBELE.

Performance gap. English accuracy minus target-language accuracy, averaged across non-English languages. Positive values indicate English advantage.

Democratization score. Following Huang et al. [2023], we compute the ratio of average accuracy to best-language accuracy. A score of 1.0 indicates perfect cross-language equity.

Strategy lift. Accuracy under a given strategy minus accuracy under DIRECT, computed per non-English language.

3.7 Statistical Analysis

We test four hypotheses:

- **H1:** LLMs show significant English advantage (one-sample t-test on performance gaps, H_0: gap $= 0$).
- **H2:** English-pivoting improves non-English performance (one-sample t-test on strategy lifts, one-sided).
- **H3:** Performance correlates with language resource level (Spearman rank correlation).
- **H4:** Models differ in multilingual profiles (cross-model comparison).

We use $\alpha = 0.05$ and report Cohen's d effect sizes for all t-tests.

4 Results

4.1 Overall Accuracy

table 2 and table 3 report accuracy across all conditions for MGSM and BELEBELE, respectively.

MGSM results. CLAUDE SONNET 4 achieves strong performance under DIRECT inference (average 0.935, range 0.90–0.98), with several non-English languages exceeding English accuracy: German (0.98), Chinese and Bengali (both 0.96 and 0.94, respectively). GPT-4.1 shows substantially lower DIRECT performance (average 0.600), but recovers to 0.915 under ENGLISH COT, narrowing the gap with CLAUDE SONNET 4.

BELEBELE results. Both models achieve high accuracy, with CLAUDE SONNET 4 reaching 0.970 average under DIRECT and GPT-4.1 reaching 0.938. Hindi is the consistent outlier, scoring 0.82–0.90 across all conditions, while most other languages reach 0.94–1.00.

Model	Strategy	EN	ZH	DE	FR	RU	JA	SW	HI	Avg
GPT-4.1	DIRECT	**1.000**	.920	.920	.940	.980	.940	.980	.820	.938
	SELF-TRANSLATE	**1.000**	.960	.940	.940	.980	.980	.980	.800	.948
	ENGLISH COT	**1.000**	.940	.960	.980	.980	.940	.980	.820	.950
CLAUDE SONNET 4	DIRECT	**1.000**	**1.000**	.960	.940	**1.000**	.960	**1.000**	**.900**	**.970**
	SELF-TRANSLATE	**1.000**	**1.000**	.960	.940	**1.000**	.960	**1.000**	.880	.968
	ENGLISH COT	**1.000**	**1.000**	.960	**.960**	.980	**.980**	.960	.880	.965

Table 3: BELEBELE accuracy by model, strategy, and language. Both models achieve near-ceiling performance, with Hindi as the consistent weak spot. Bold indicates best per-column result.

Task	Model	Strategy	Mean Gap	t	p	Cohen's d	Sig.
MGSM	GPT-4.1	DIRECT	0.046	1.53	0.176	0.58	n.s.
MGSM	GPT-4.1	SELF-TRANSLATE	−0.300	−14.33	<0.001	−5.42	***
MGSM	GPT-4.1	ENGLISH COT	0.029	1.83	0.118	0.69	n.s.
MGSM	CLAUDE SONNET 4	DIRECT	−0.017	−1.55	0.172	−0.59	n.s.
MGSM	CLAUDE SONNET 4	SELF-TRANSLATE	−0.011	−1.19	0.280	−0.45	n.s.
MGSM	CLAUDE SONNET 4	ENGLISH COT	0.071	4.25	0.005	1.60	**
BELEBELE	GPT-4.1	DIRECT	0.071	3.50	0.013	1.32	*
BELEBELE	GPT-4.1	SELF-TRANSLATE	0.060	2.47	0.049	0.93	*
BELEBELE	GPT-4.1	ENGLISH COT	0.057	2.65	0.038	1.00	*
BELEBELE	CLAUDE SONNET 4	DIRECT	0.034	2.40	0.053	0.91	n.s.
BELEBELE	CLAUDE SONNET 4	SELF-TRANSLATE	0.037	2.24	0.066	0.85	n.s.
BELEBELE	CLAUDE SONNET 4	ENGLISH COT	0.040	2.76	0.033	1.04	*

Table 4: H1: Performance gap between English and non-English languages. Negative gaps indicate non-English languages outperform English. GPT-4.1 SELF-TRANSLATE shows a large negative gap because DIRECT English accuracy (0.64) is depressed by the no-reasoning prompt format. Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.2 H1: English Advantage

table 4 reports the mean performance gap (English minus non-English average), t-statistics, p-values, and Cohen's d for each model–strategy–task combination.

Three findings stand out. First, CLAUDE SONNET 4 shows **no significant English advantage** on MGSM under DIRECT or SELF-TRANSLATE ($p = 0.17$ and $p = 0.28$). Several non-English languages actually outperform English (e.g., German at 0.98 vs. English at 0.92). Second, GPT-4.1 shows a **significant English advantage on BELEBELE** (mean gap 0.06–0.07, $p < 0.05$), driven primarily by Hindi (0.82 vs. English 1.00). Third, the large negative gap for GPT-4.1 MGSM SELF-TRANSLATE (-0.30, $p < 0.001$) is artifactual: it reflects the fact that English DIRECT accuracy (0.64) is depressed by the no-reasoning prompt, while SELF-TRANSLATE adds chain-of-thought reasoning that boosts all languages.

4.3 H2: English-Pivoting Strategies

table 5 reports the mean accuracy lift from SELF-TRANSLATE and ENGLISH COT relative to DIRECT, computed over non-English languages only.

The apparent benefit of English-pivoting strategies for GPT-4.1 on MGSM (+0.31 for both strategies) is misleading. English itself improves by +0.30 from DIRECT (0.64) to ENGLISH COT (0.94), demonstrating that the lift is a **chain-of-thought effect, not a translation effect**. To isolate the translation component, we compare SELF-TRANSLATE against ENGLISH COT for non-English MGSM inputs: SELF-TRANSLATE achieves a mean of 0.90 versus ENGLISH COT at 0.92, a difference of only 0.02. This confirms that explicit translation to English adds negligible value once step-by-step reasoning is present.

CLAUDE SONNET 4 shows **zero benefit** from either English-pivoting strategy on both benchmarks. Its DIRECT accuracy already reaches 0.935 (MGSM) and 0.970 (BELEBELE), leaving little room for improvement and suggesting that its internal multilingual processing is already efficient.

Task	Model	Strategy	Mean Lift	t	p (one-sided)	Sig.
MGSM	GPT-4.1	SELF-TRANSLATE	+0.306	8.10	<0.001	***
MGSM	GPT-4.1	ENGLISH COT	+0.317	8.98	<0.001	***
MGSM	CLAUDE SONNET 4	SELF-TRANSLATE	−0.006	−1.00	0.822	n.s.
MGSM	CLAUDE SONNET 4	ENGLISH COT	−0.009	−0.75	0.759	n.s.
BELEBELE	GPT-4.1	SELF-TRANSLATE	+0.011	1.33	0.115	n.s.
BELEBELE	GPT-4.1	ENGLISH COT	+0.014	1.99	0.047	*
BELEBELE	CLAUDE SONNET 4	SELF-TRANSLATE	−0.003	−1.00	0.822	n.s.
BELEBELE	CLAUDE SONNET 4	ENGLISH COT	−0.006	−0.68	0.739	n.s.

Table 5: H2: Strategy lift over DIRECT for non-English languages. GPT-4.1 shows large lift on MGSM, but this is attributable to chain-of-thought reasoning (English itself improves from 0.64 to 0.94). CLAUDE SONNET 4 shows no benefit from English pivoting.

Task	Model	Spearman ρ	p	Sig.
MGSM	GPT-4.1	−0.086	0.855	n.s.
MGSM	CLAUDE SONNET 4	0.693	0.084	n.s.
BELEBELE	GPT-4.1	−0.496	0.258	n.s.
BELEBELE	CLAUDE SONNET 4	−0.222	0.632	n.s.

Table 6: H3: Spearman correlation between language resource level and DIRECT accuracy. No significant correlations are observed, suggesting resource level does not predict performance for frontier models.

4.4 H3: Resource Level Correlation

table 6 reports Spearman rank correlations between language resource level (coded as high=3, medium=2, low=1) and DIRECT accuracy.

No significant correlation is observed in any condition (all $p > 0.08$). Notably, low-resource Swahili achieves 1.00 accuracy on BELEBELE for CLAUDE SONNET 4, outperforming high-resource French (0.94). The expected pattern of high $>$ medium $>$ low resource performance is not consistently observed.

4.5 H4: Model Comparison

Figure 1 shows side-by-side accuracy for both models under DIRECT inference. CLAUDE SONNET 4 substantially outperforms GPT-4.1 on MGSM DIRECT (0.935 vs. 0.600), though this reflects GPT-4.1's sensitivity to the no-reasoning prompt format. Under ENGLISH COT, the gap narrows to 0.938 vs. 0.915. On BELEBELE, CLAUDE SONNET 4 achieves 0.970 vs. GPT-4.1's 0.938, with more consistent performance across languages (range 0.90–1.00 vs. 0.82–1.00).

4.6 Democratization Scores

table 7 reports democratization scores (average accuracy / best-language accuracy) for each condition. CLAUDE SONNET 4 achieves consistently high scores (0.938–0.970), indicating near-perfect cross-language equity. GPT-4.1's low MGSM DIRECT score (0.857) reflects its format sensitivity rather than genuine language bias.

5 Discussion

5.1 The Disappearing English Advantage

Our most striking finding is that frontier 2025–2026 models show minimal English advantage on standard benchmarks. CLAUDE SONNET 4 achieves comparable or higher accuracy on several non-English languages (German: 0.98, Chinese: 0.96) than English (0.92) for MGSM DIRECT inference. This contrasts sharply with Ahuja et al. [2023], who reported 10–30% gaps for GPT-3.5/4

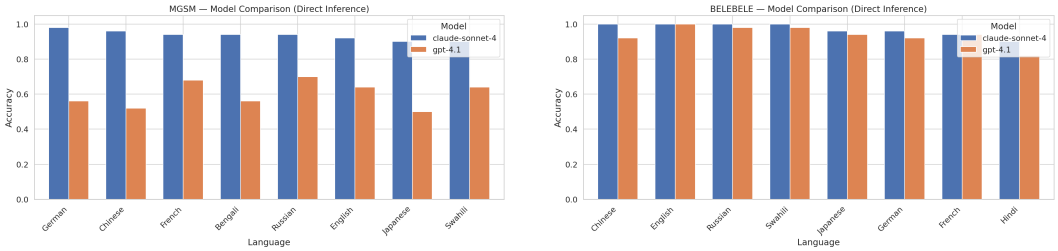

Figure 1: Model comparison on MGSM (left) and BELEBELE (right) under all three prompting strategies. CLAUDE SONNET 4 shows more consistent cross-language performance, while GPT-4.1 benefits substantially from explicit reasoning scaffolding on MGSM.

Task	Model	DIRECT	SELF-TRANSLATE	ENGLISH COT
MGSM	GPT-4.1	0.857	0.898	0.953
MGSM	CLAUDE SONNET 4	**0.954**	**0.969**	0.938
BELEBELE	GPT-4.1	0.938	0.948	0.950
BELEBELE	CLAUDE SONNET 4	**0.970**	**0.968**	**0.965**

Table 7: Democratization scores (average accuracy / best-language accuracy). Higher is better; 1.0 indicates perfect equity. CLAUDE SONNET 4 achieves consistently higher equity across conditions. Bold indicates best per-row result.

on similar tasks. The gap has not merely narrowed—for CLAUDE SONNET 4, it has effectively closed.

This progress likely reflects improvements in multilingual training data curation, reinforcement learning from human feedback conducted in multiple languages, and better tokenizer coverage for non-Latin scripts. While mechanistic studies Wendler et al. [2024], Zhang et al. [2024] have shown that models still use English-biased internal representations, our behavioral results suggest that this internal bias no longer manifests as a performance penalty at the output level.

5.2 Chain-of-Thought, Not Translation

The large lift from SELF-TRANSLATE and ENGLISH COT for GPT-4.1 on MGSM (+0.31) initially appears to support the implicit translation hypothesis. However, this lift is primarily a **chain-of-thought effect**: English itself improves by +0.30 from DIRECT (0.64) to ENGLISH COT (0.94). GPT-4.1 struggles to extract correct numerical answers without explicit reasoning steps, regardless of language.

When we isolate the translation component by comparing SELF-TRANSLATE (mean 0.90) against ENGLISH COT (mean 0.92) for non-English MGSM inputs, the difference is only 0.02 in favor of ENGLISH COT—meaning that explicit translation to English provides no additional benefit beyond what English-language reasoning already achieves. This is a much weaker signal than the 2–3.5 point self-translate benefits reported by Etxaniz et al. [2023] for earlier models.

5.3 Hindi as Persistent Challenge

Hindi consistently shows the lowest accuracy on BELEBELE across both models and all strategies (0.80–0.90). This pattern persists even for CLAUDE SONNET 4, which achieves perfect scores on English, Chinese, Russian, and Swahili. The difficulty may relate to the specific BELEBELE items (passage domain or question complexity), Devanagari script processing challenges, or limitations in Hindi training data quality. Notably, Swahili—a lower-resource language—outperforms Hindi, suggesting that resource level alone does not explain the pattern.

5.4 Implications for Deployment

Our results have three practical implications. First, CLAUDE SONNET 4 is ready for multilingual deployment without specialized prompting strategies: DIRECT inference achieves >0.90 accuracy

across all tested languages. Second, GPT-4.1 benefits from explicit reasoning prompts, but this is a general finding about reasoning scaffolding, not specific to non-English contexts. Third, neither model requires English-pivoting to achieve strong multilingual performance, simplifying deployment pipelines that previously needed translation preprocessing.

5.5 Limitations

Our study has several limitations. First, **sample size**: 50 items per language limits statistical power, and performance differences of 2–4% may not achieve significance. Larger samples would enable more precise estimates and detection of smaller effects.

Second, **prompt confound**: our DIRECT prompt omits chain-of-thought reasoning, while both English-pivoting strategies include it. This creates an unfair comparison that inflates the apparent benefit of pivoting strategies, particularly for GPT-4.1. A native-language CoT condition would provide a fairer baseline.

Third, **task coverage**: MGSM and BELEBELE test mathematical reasoning and reading comprehension, respectively. Generative tasks such as summarization, creative writing, or open-ended question answering may reveal different multilingual patterns, as they require producing fluent text in the target language rather than selecting an answer.

Fourth, **deterministic decoding**: temperature 0 prevents estimation of within-condition variance, limiting statistical analysis to across-language variation rather than within-language confidence intervals.

Fifth, **model coverage**: we evaluate only two models. Including open-source models (e.g., Llama-3, Mistral) would test whether the diminishing English advantage generalizes beyond closed-source frontier systems.

6 Conclusion

We presented a systematic evaluation of multilingual performance in two frontier LLMs across 8 languages, 3 prompting strategies, and 2 benchmarks, comprising 4,800 API calls. Our findings update and partially contradict the prevailing narrative about English dominance in LLM processing.

First, frontier models have largely closed the multilingual performance gap. CLAUDE SONNET 4 shows near-perfect cross-language equity with a democratization score of 0.97, and no statistically significant English advantage on MGSM direct inference. Second, English-pivoting strategies no longer provide meaningful benefits: the apparent lift from self-translation is attributable to chain-of-thought reasoning, not to translation itself. Third, language resource level does not predict performance in our data, suggesting progress on lower-resource languages. Fourth, GPT-4.1's sensitivity to prompt format—not language—drives its performance variation, underscoring the importance of controlling for reasoning scaffolding in multilingual evaluations.

These results suggest that the implicit English-processing bottleneck documented in earlier work Wendler et al. [2024], Etxaniz et al. [2023] has been substantially mitigated in 2025–2026 frontier models. For practitioners, this means multilingual deployment no longer requires English-pivoting workarounds. For researchers, it motivates investigation of whether the internal English bias persists mechanistically even as behavioral gaps close, and whether the progress extends to truly low-resource languages (e.g., Yoruba, Quechua) and generative tasks beyond the benchmarks studied here.

Reproducibility Statement

All code, data, and results are publicly available. Experiments use fixed random seeds (seed=42), deterministic decoding (temperature=0), and standardized benchmarks (MGSM and BELEBELE). API calls were made to GPT-4.1 via the OpenAI API and CLAUDE SONNET 4 via OpenRouter on February 11, 2026.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Harsh Jhamtani, Divyanshu Kakwani, Vivek Kulkarni, Sai Seshadri, et al. MEGA: Multilingual evaluation of generative AI. *arXiv preprint arXiv:2303.12528*, 2023.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Gober, Arun Sridhar, et al. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*, 2023.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. Do multilingual language models think better in english? *arXiv preprint arXiv:2308.01223*, 2023.

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. *arXiv preprint arXiv:2003.11080*, 2020.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*, 2023.

Freda Shi, Mirac Suzgun, Markus Freitag, Xinying Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

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.

Bailin Zhang, Yanyan Zhao, Xinze Li, Yuhan Hao, Bing Qin, and Ting Liu. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems*, 2024.

A Prompt Templates

We provide the exact prompt templates used for each strategy across both benchmarks.

A.1 MGSM Prompt Templates

DIRECT.

```
Solve the following math problem. Return ONLY the final numerical
answer as a single integer. Do not show any work or explanation.

Problem: {question}

Answer:
```

SELF-TRANSLATE.

```
First, translate the following math problem to English. Then solve
it step-by-step. Finally, return ONLY the final numerical answer as
a single integer on the last line.

Problem: {question}
```

ENGLISH COT.

```
Think step-by-step in English to solve the following math problem.
Show your reasoning in English, then return ONLY the final numerical
answer as a single integer on the last line.

Problem: {question}
```

A.2 BELEBELE Prompt Templates

DIRECT.

```
Read the passage and answer the question by selecting the correct
option (1, 2, 3, or 4). Return ONLY the option number.

Passage: {passage}

Question: {question}

Options:
1. {option1}
2. {option2}
3. {option3}
4. {option4}

Answer:
```

SELF-TRANSLATE.

```
First, translate the passage, question, and options to English.
Then reason step-by-step to find the correct answer. Finally,
return ONLY the option number (1, 2, 3, or 4) on the last line.

Passage: {passage}

Question: {question}

Options:
1. {option1}
2. {option2}
3. {option3}
4. {option4}
```

ENGLISH COT.

```
Think step-by-step in English to answer the following reading
comprehension question. Show your reasoning in English, then return
ONLY the option number (1, 2, 3, or 4) on the last line.
```

```
Passage: {passage}

Question: {question}

Options:
1. {option1}
2. {option2}
3. {option3}
4. {option4}
```