ROUND-TRIP TRANSLATION REVEALS WHAT FRONTIER MULTILINGUAL BENCHMARKS MISS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual benchmarks guide the development of frontier models. Yet multilingual evaluations reported by frontier models are structured similar to popular reasoning and knowledge benchmarks, but across many languages. We show such benchmarks, and consequently multilingual evaluations, measure mathematical reasoning and factual recall, not multilingual proficiency. For example, thinking variants dramatically outperform instruct variants on these benchmarks, yet often perform worse on real-world multilingual tasks, such as LMArena. We propose a simple alternative: evaluate multilingual capability via round-trip translation. Given text in a source language, translate it to a target language and back; semantic gaps between the original and result expose failures in multilingual *generation* capabilities. Round-trip translation correlates almost perfectly ($\rho = 0.94$) with user ratings on LMArena with our benchmark, requires no human reference translations, and does not require a more capable multilingual judge than tested models. Lastly, we introduce Lost in Translation (LiT), a challenging round-trip translation benchmark spanning widely spoken languages worldwide, for realistic evaluation of multilingual frontier models.

1 INTRODUCTION

Multilingual benchmarks (Son et al., 2025; Wang et al., 2025; Romanou et al., 2025; Singh et al., 2025) shape how frontier models are built. Yet a fundamental question remains open: Does progress on current multilingual benchmarks truly reflect progress in multilingual proficiency?

We find it does not. The evaluation of frontier multilingual models is currently dominated by two predominant paradigms: mathematical reasoning tasks, such as MT-AIME24 (Son et al., 2025) and PolyMath (Wang et al., 2025), and general knowledge multiple-choice question answering (MCQA), such as INCLUDE (Romanou et al., 2025) and Global-MMLU (Singh et al., 2025). We discover, as shown in Figures 1 and 2, that performance gaps on such benchmarks primarily reflect differences in mathematical problem-solving ($\rho = 0.94$) or factual recall ($\rho = 0.83$) respectively – not multilingual generation capability ($\rho = -0.09$ and -0.26 respectively). Previous works (Wu et al., 2025) provide evidence of the failure of multilingual benchmarks to align with human preferences.

We propose a simple solution: evaluate multilingual capability through round-trip translation (Brislin, 1970; Sennrich et al., 2016). Given a passage in a source language, round-trip translation involves translating it to a target (sequence of) language(s) and back – comparing the result to the original passage. Semantic gaps between the original and back-translated text show whether a model can preserve meaning across languages, a critical capability multilingual benchmarks should capture (Wu et al., 2025). Unlike MT-AIME or Global-MMLU, multilingual understanding and generation are the challenging tasks in a round-trip translation benchmark.

We describe its comparative advantages in Table 1 when compared to current machine translation and current multilingual benchmarks. LiT combines recent public paper abstracts (20%) with texts we authored (80%) to prevent data contamination. Round-trip translation has two advantages over classical translation evaluation: scalable sample creation – it requires no human-written reference translations; and ease of judging – the judge needs to evaluate two English paragraphs, not compare them in the low-resource language. This implies we do not need a stronger model to evaluate multilingual translation quality, which does not exist when testing frontier models. Current translation benchmarks are too easy for evaluating frontier models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: **Benchmark comparison across five evaluation criteria.** We compare nine multilingual benchmarks across five evaluation criteria including: contamination-free, challenging (not saturated yet, difficult for frontier models), linguistic diversity (covers diverse language families and scripts), NLG (tests natural language generation), and ground truth-free (without per-sample human annotation).

Benchmark	Contamination-Free	Challenging	Linguistic Diversity	NLG	Ground Truth-Free
General Knowledge Multiple-Choice					
Include	✓	✓	✓	✗	✗
MMMLU	✗	✗	✗	✗	✗
Global MMLU	✗	✗	✗	✗	✗
Mathematical Reasoning					
MT-AIME24	✗	✓	✗	✓	✗
MGSM	✗	✗	✗	✓	✗
PolyMath	✓	✓	✗	✓	✗
Machine Translation					
FLORES-200	✗	✗	✓	✓	✗
WMT24++	✓	✗	✓	✓	✗
LiT (Ours)	✓	✓	✓	✓	✓

To enable systematic evaluation, we introduce Lost in Translation (LiT), a round-trip translation benchmark of 1600 samples across 8 language sequences spanning high-, medium-, and low-resource languages. LiT covers technical (Taguchi et al., 2025), pragmatic (Park et al., 2024), and informal language (Yao et al., 2024) using MQM-based automated judging. First, we show that round-trip translation correlates almost perfectly ($\rho = 0.94$) with user ratings on LMArena (Chiang et al., 2024)[1]. Second, it exposes failure modes that reasoning and multiple-choice benchmarks miss. We document systematic hallucinations, content omissions, and semantic drift in multilingual *generation* using MQM scores (Lommel et al., 2014; Kocmi & Federmann, 2023), not faithfully captured by current multiple-choice (MCQ) or mathematical reasoning evaluations.

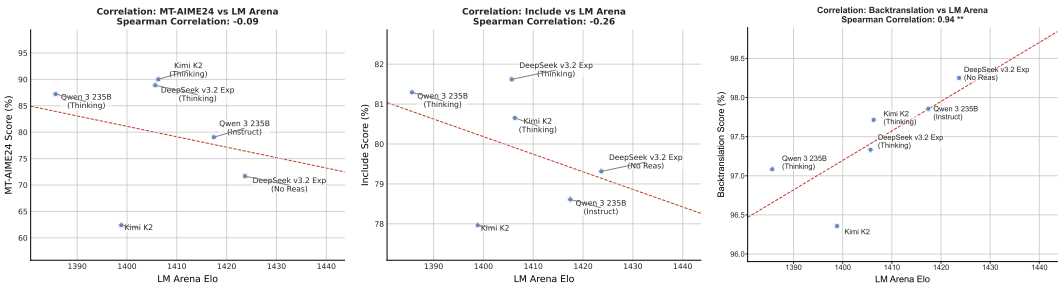
Our results reveal a stark capability gap: open-source frontier models achieve above 88 MQM scores on high-resource sequences but collapse to below 50 on low-resource languages, indicating unusable output[2]. Furthermore, we test model performance on challenging constructions for round-trip translation (Somers, 2005; Zhuo et al., 2023). Model rankings on these challenging cases match general performance closely, validating the robustness of the LiT benchmark. When instructed to translate faithfully, frontier models prioritize instruction-following over fluency correction. Rather than self-correcting or masking intermediate errors, they reliably reproduce the awkward phrasing. Consequently, this demonstrates that round-trip evaluation is not artificially derailed by complex linguistic phenomena, but instead serves as a stable, highly correlated proxy for genuine cross-lingual generation.

2 Related Works

Multilingual Benchmarks. The standard approach to multilingual LLM evaluation translates English reasoning and knowledge tasks into other languages (Hendrycks et al., 2021; Son et al., 2025). However, such translated benchmarks often correlate poorly with human evaluations, performing significantly worse than natively localized benchmarks Wu et al. (2025). Alternatives like INCLUDE Romanou et al. (2025) and Global-MMLU Singh et al. (2025) address this by incorporating native content or cultural context, mitigating both "translationese" artifacts and Western-centric bias Wu et al. (2025). However, these localized benchmarks inherit a deeper structural flaw: they evaluate knowledge retrieval via multiple-choice formats rather than actual natural language generation, suffering from selection bias (Zheng et al., 2024; Balepur et al., 2025). Furthermore, natively sourced benchmarks like INCLUDE (Romanou et al., 2025) inherently provide imbalanced cross-

[1]LMArena is considered an expensive but gold standard set, as it collects a vast array of continuously updated real-world user queries and uses a wisdom of the crowd effect (see Ni et al. (2024) for details).

[2]Translations achieving an MQM score above 80 are generally considered faithful and fit-for-purpose, while a score of <80 should trigger mandatory human review (Lommel et al., 2014; 2024).



(a) MT-AIME24 vs. LMArena (b) INCLUDE vs. LMArena (c) LiT vs. LMArena

Figure 1: **Multilingual benchmarks correlate poorly with human preferences.** We show benchmark scores against LMArena Elo ratings for six frontier open-source models, each in Thinking and Non-Thinking variants. (a) MT-AIME24 (Son et al., 2025) shows near-zero correlation ($\rho = -0.09$), with thinking variants dramatically outperforming on the benchmark but performing no better on LMArena; (b) INCLUDE (Romanou et al., 2025) shows moderate negative correlation ($\rho = -0.26$), with a similar Thinking vs. Non-Thinking disconnect; (c) Round-trip translation using LiT with percentage of MQM scores at least 80 correlates almost perfectly with LMArena ($\rho = 0.94$, with one-sided permutation test p=0.008), with minimal gap between thinking and instruct variants of a model.

lingual comparisons. Evaluating German on the Driver's License topic while testing French on other topics confounds language proficiency with task difficulty. Ultimately, real-world utility depends on coherent, fluent, and semantically faithful generation. We provide empirical evidence demonstrating how current benchmarks fail to capture this, motivating our standardized generative approach.

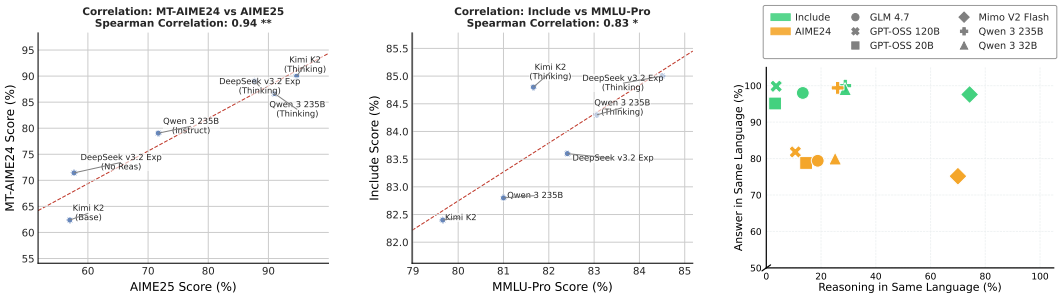
Natural Language Generation Foundational benchmarks such as MMLU (Hendrycks et al., 2021; Sai et al., 2022) acknowledge that the future of model evaluation lies in Natural Language Generation (NLG), but use multiple-choice formats because NLG remains difficult to evaluate. Consequently, assessing multilingual generation quality remains an open problem.

Traditional machine translation benchmarks like WMT (Kocmi et al., 2025) and FLORES (Team et al., 2022) directly evaluate translation quality, but they depend on reference translations or per-example human ratings. These benchmarks also draw from Wikipedia and other widely-used sources in training datasets, raising concerns about data contamination into model training (Sainz et al., 2024; Karpinska & Iyyer, 2023; Vilar et al., 2023). Furthermore, most evaluations are sentence-level, lacking the complexity of real-world translation tasks.

Recent NLG benchmarks such as PolyMath (Wang et al., 2025), MGSM (Shi et al., 2023), and MT-AIME24 (Son et al., 2025) attempt to test multilingual generation by machine-translating English-sourced mathematical questions. However, as shown in Fig 2, these approaches primarily evaluate mathematical proficiency rather than multilingual capability.

To address these gaps, we curate a benchmark of complex, multi-sentence passages reflecting real-world use cases. We adapt the MQM framework (Lommel et al., 2014), used in WMT Shared Tasks (Freitag et al., 2021; Lavie et al., 2025) and replace human translators with LLM judges for scalability (Lu et al., 2024; Kocmi & Federmann, 2023; Kim, 2025). Separately, LMArena Chiang et al. (2024) provides a large-scale measure of real-world user preferences. This approach has become the standard for tracking real-world model performance and has inspired work on efficient proxies that maintain high correlation with full arena rankings (Li et al., 2025; Dubois et al., 2024; Spangher et al., 2025).

Round-Trip Translation. Round-trip translation has a long history as an evaluation tool. The foundational Brislin (1970) paper established its use for validating translation quality. It later became a widely used data augmentation approach in neural machine translation (Sennrich et al., 2016). The core insight is straightforward: if meaning degrades through a round-trip (English $\rightarrow$ target $\rightarrow$ English), the forward translation is usually unreliable. Early statistical models undermined this approach by simply utilizing a "copy mechanism" (Somers, 2005). However, modern neural machine translation (NMT) architectures lack this flaw, confirming that round-trip translation is an effective



(a) MT-AIME24 vs AIME25 (b) INCLUDE vs MMLU-Pro (c) Probing reasoning language

Figure 2: **Multilingual benchmarks track English reasoning performance, and models reason in English regardless of input language.** (a) MT-AIME24 scores correlate strongly with English AIME25 performance ($\rho = 0.94$, one-sided permutation test p=0.008), indicating the benchmark primarily measures mathematical reasoning ability. (b) INCLUDE scores correlate strongly with English MMLU-Pro ($\rho = 0.83$, one-sided permutation test p=0.029), indicating the benchmark primarily measures factual knowledge. (c) Models overwhelmingly default to English for internal reasoning (y-axis) even when answering in the target language (x-axis). This rules out the possibility that benchmark errors reflect multilingual reasoning failures.

method for reference-free evaluations (Moon et al., 2020; Zhuo et al., 2023). Where prior work mainly uses round-trip translation for MT data augmentation (Sennrich et al., 2016), we use it to evaluate frontier LLMs (Allamanis et al., 2024).

3 Current Benchmarks Do Not Measure Multilingual Capability

Frontier model reports (Yang et al., 2025) evaluate multilingual capabilities using popular multilingual reasoning benchmarks like MT-AIME24 (Son et al., 2025), MGSM (Shi et al., 2023) and PolyMath (Wang et al., 2025), as well as general knowledge MCQA benchmarks such as MMMLU (Hendrycks et al., 2021) and INCLUDE (Romanou et al., 2025). In this section, we show that these benchmarks correlate poorly with human preferences because they measure reasoning or factual recall, not multilingual comprehension.

3.1 Benchmark Scores Diverge from User Preferences

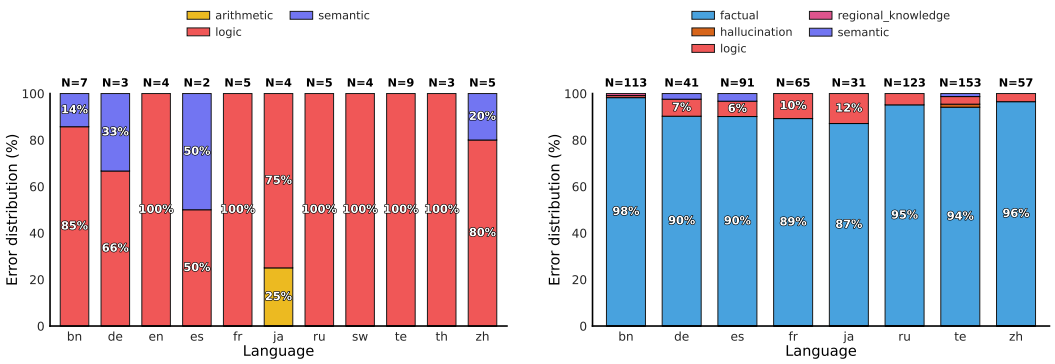
Higher scores on a good benchmark should imply better real-world utility. We test whether multilingual benchmarks satisfy this criterion by correlating their scores with user experience in-the-wild.

Setup. We evaluate six frontier open-source models: Kimi K2[3] (Team et al., 2025), DeepSeek-V3.2-Exp (Liu et al., 2025), and Qwen3-235B-2507 (Yang et al., 2025), each in Thinking and Non-Thinking variants. We constrain the comparison to same-tier models to avoid spurious correlations driven by model size (Kaplan et al., 2020; Ghorbani et al., 2022); Appendix A.2 discusses the statistical interpretation. We measure zero-shot accuracy on MT-AIME24 (Son et al., 2025), a multilingual reasoning benchmark, and INCLUDE (Romanou et al., 2025), a culturally-curated, knowledge-intensive benchmark. We then compute Spearman rank-order correlations between benchmark scores and LMArena (Chiang et al., 2024) Elo ratings. For comparison, we also evaluate backtranslation quality on LiT (English → Language → English) across seven LMArena languages[4] and report $MQM_{\geq 80}$: the percentage of backtranslations whose MQM score is at least 80.

Results. Figure 1 shows that benchmark rankings diverge from human preferences. Both MT-AIME24 (left) and INCLUDE (middle) exhibit slight to moderate **negative** correlation with

[3]Unlike other models, Kimi K2 Thinking was released after the Non-Thinking variant.

[4]ZH, FR, DE, JA, KO, RU, and ES. We use mean ratings across available languages from the public LMArena leaderboards. In our snapshot, JA is unavailable for both DeepSeek variants and FR for Qwen3-235B (Thinking).



(a) Error distribution for MT-AIME24.

(b) Error distribution for INCLUDE.

Figure 3: Errors on multilingual benchmarks stem from reasoning and knowledge gaps—not language comprehension failures. We manually categorize all errors made by Qwen3-235B-A22B-Thinking across 11 languages. (a) On MT-AIME24, errors are overwhelmingly logical (arithmetic mistakes, flawed reasoning steps) rather than semantic (misunderstanding the translated question). (b) On INCLUDE, errors are predominantly factual (wrong facts, knowledge gaps) rather than semantic (misunderstanding the translated question). Both patterns confirm that these benchmarks measure reasoning and factual recall – not multilingual proficiency.

LMArena Elo scores. The disconnect is strongest between Thinking and Non-Thinking variants of the same model: Thinking models dramatically outperform their counterparts on MT-AIME24 and INCLUDE, yet on LMArena – where users rate actual multilingual outputs – they often perform no better. In contrast, round-trip translation scores on LiT correlate almost perfectly ($\rho = 0.94$) with LMArena, suggesting closer alignment with real-world multilingual performance. This disconnect also appears for Global-MMLU-Lite and MGSM, which show weak rank correlations with LMArena across 13 and 17 similarly rated models, respectively (Appendix G.3, Figure 11).

Analysis. Why do multilingual benchmarks fail to predict real-world performance? We observe that they conflate two distinct capabilities: reasoning/knowledge and multilingual understanding. We correlate each multilingual benchmark with an English benchmark measuring the same underlying skill: MT-AIME24 with AIME25, and INCLUDE with MMLU-Pro (Wang et al., 2024). Figures 2a and 2b show near perfect correlations across both benchmark pairs. This suggests that gains on multilingual benchmarks may largely reflect gains on English reasoning and knowledge tasks. Billions who speak languages other than English might see limited benefit from progress on these benchmarks.

Benchmark Scores Diverge from Real-World Use

Multilingual benchmarks conflate two distinct capabilities: reasoning/facts and multilingual understanding, i.e. progress on (MT-)AIME24 primarily stems from better mathematical capability, not improved English (language) proficiency. We need challenging *language-centric* multilingual benchmarks.

3.2 Errors in Current Benchmarks are not Linguistic

We next analyze error types in two common benchmark categories, multilingual reasoning and general-knowledge multiple choice, to test whether these benchmarks reflect reasoning or factual knowledge rather than language proficiency.

MT-AIME24 errors are logical, not semantic. We analyze errors for Qwen3-235B-A22B-Thinking-2507 on MT-AIME24 across 11 languages, with similar trends reproducible across a variety of models (detailed in Appendix F). Figure 3 shows the distribution. We observe that most errors stem from logical mistakes, calculation errors or flawed reasoning steps, and not from poor (multilingual) question comprehension. For French, Russian, and Thai, 100% of errors reflect failed reasoning, not failed understanding.

INCLUDE errors are factual, not semantic. One might argue that a general knowledge benchmark such as INCLUDE avoids this problem because it uses native questions rooted in regional knowledge rather than simply translating text from MMLU. However, the same pattern persists, including for other models in Appendix F. Figure 3 shows that most errors stem from incorrect factual knowledge. For Bengali and Chinese, over 96% of errors reflect knowledge gaps, not misunderstanding of the input question or MCQ options. Performance gains on INCLUDE, therefore, similarly reflect better elicitation and coverage of facts, not better multilingual capability.

Reasoning failures occur in English, not the target language. One could argue that reasoning in the target language is itself worth assessing. We test whether models actually reason in the target language by analyzing traces from multiple language models across both benchmarks. Figure 2c shows they do not: models almost always default to English for their reasoning traces. The knowledge gaps and logical errors we identify therefore occur in English, not even in the target language. This is further evidence suggesting that reasoning failures do not reflect multilingual failures – they are reasoning/factual mistakes made in the English language (for extensive results, please refer to the Appendix E and F).

Summary. Overall, both benchmark types fail to disentangle multilingual capability from orthogonal skills. A model with strong language proficiency, but weak mathematical reasoning (like DeepSeek-V3.2-Exp Non-Thinking) (as seen in Figure 1) scores poorly on MT-AIME24. Equivalently, a model with limited factual coverage underperforms on INCLUDE. Because models across newer generations often broadly improve both multilingual and reasoning capabilities at the same time, benchmark scores can appear to track multilingual progress even when they mainly reflect improvement in other areas.

Errors Don't Reflect Multilingual Failures

Errors in current multilingual benchmarks primarily reflect reasoning or knowledge gaps, and these gaps persist in English. Multilingual benchmarks should analyze the errors to understand the gaps in multilingual capabilities.

4 The Lost In Translation (LiT) Benchmark

Following our previous analysis of multilingual benchmark shortcomings, we propose Lost In Translation (LiT) – a natural language generation benchmark centered on backtranslation. LiT addresses the limitations of existing benchmarks summarized in Table 1: limited linguistic diversity, reliance on surface-level metrics, and failure to probe multilingual generation capabilities.

4.1 Dataset: Description and Setup

Data Curation. LiT contains 1,600 evaluation samples: 200 source texts evaluated across eight language sequences. The texts cover recent public paper abstracts (20%), pragmatics (60%), and informal language (20%). We authored all tasks and non-abstract texts, which were kept private to reduce contamination risk. Appendix B.1 provides the full category breakdown.

Language Sequences. Each sequence translates an English source through four intermediate languages before returning to English. The sequences span eight geographic regions and different language-resource levels; Appendix B.2 provides the language selection and ordering. Section 4.2 examines sensitivity to chain length.

The sequences are also grouped into sequences containing exclusively well-resourced languages or including at least one medium- or low-resource language.

Evaluation: LLM-as-a-Judge. We use Grok 4.1 Fast to score translations with MQM, which penalizes minor, major, and critical errors by 1, 5, and 25 points, respectively (Lommel et al., 2014; Kim, 2025). We report $\text{MQM}_{\geq 80}$, the percentage of translations meeting the 80-point pass threshold (Farinha et al., 2022; Lommel et al., 2024). Appendix B.3 details the rubric; Appendix D compares alternative metrics, and Section 4.2 tests judge choice. Sampling settings and the translation and judge prompts appear in Appendices B.5–B.7.

Table 2: LiT benchmark by linguistic category under $MQM_{\geq 80}$. We report the percentage of translations with MQM $\geq$ 80, aggregated within each category over eight translation sequences. Each cell shows the mean and the bootstrap-estimated standard error, computed via sentence-level resampling after averaging across sequences. Higher is better. Informal text is the hardest category overall (30.8 avg), while Core Semantics is the easiest (61.5).

Model	Average	Abstracts		Pragmatics					Informal
		Humanities	STEM	Core Semantics	Discourse	Implicit	Inference	Social	
Gemini-3-Flash (No-Thinking)	87.2 ± 0.7	90.0 ± 1.9	85.6 ± 1.8	87.5 ± 1.9	89.3 ± 1.5	86.9 ± 2.3	88.5 ± 1.6	89.7 ± 1.6	80.3 ± 2.2
Qwen3.5-397B (Thinking)	73.8 ± 1.0	70.0 ± 3.7	88.1 ± 2.7	81.2 ± 2.5	75.6 ± 3.2	76.2 ± 3.1	78.4 ± 2.0	75.9 ± 2.5	45.0 ± 3.5
Gemma-4-31B (Instruct)	73.0 ± 1.0	71.2 ± 3.4	76.2 ± 4.4	78.1 ± 2.3	75.0 ± 2.8	69.0 ± 3.1	79.8 ± 1.7	77.7 ± 1.6	56.9 ± 2.7
Gemma-4-31B (Thinking)	71.9 ± 1.0	70.6 ± 3.2	76.2 ± 3.4	75.5 ± 2.6	74.4 ± 2.5	70.8 ± 2.9	74.0 ± 2.5	77.7 ± 1.8	55.9 ± 2.9
GLM-5 (Thinking)	70.3 ± 1.1	66.2 ± 4.0	74.4 ± 2.9	77.1 ± 3.1	77.4 ± 1.4	71.4 ± 3.3	77.4 ± 1.9	72.8 ± 3.0	45.9 ± 3.2
Qwen3.5-397B (Instruct)	67.6 ± 1.2	71.2 ± 4.0	69.4 ± 4.4	71.9 ± 3.8	69.6 ± 2.2	66.7 ± 4.0	74.0 ± 1.9	71.4 ± 2.4	46.9 ± 3.2
Kimi-K2 (Thinking)	64.7 ± 0.9	63.1 ± 3.2	58.1 ± 4.0	70.6 ± 1.3	70.8 ± 1.4	67.6 ± 2.6	71.6 ± 1.5	68.3 ± 1.9	47.2 ± 2.4
GLM-4.7 (Thinking)	61.6 ± 1.0	63.8 ± 3.2	60.6 ± 3.5	72.4 ± 2.5	70.2 ± 2.0	61.3 ± 3.6	67.8 ± 1.9	65.2 ± 2.3	31.6 ± 2.9
Qwen3-235B (Thinking)	55.6 ± 1.0	49.4 ± 3.7	51.9 ± 3.8	67.2 ± 1.9	66.7 ± 2.1	60.1 ± 3.7	67.3 ± 1.8	59.8 ± 2.8	22.8 ± 2.5
DeepSeek-V3.2-Exp	53.4 ± 1.1	52.5 ± 3.6	42.5 ± 3.9	58.9 ± 2.9	57.1 ± 3.4	58.3 ± 2.8	63.9 ± 2.7	57.6 ± 2.6	36.6 ± 3.1
GLM-5 (Instruct)	53.3 ± 1.2	51.2 ± 4.2	45.0 ± 4.2	60.9 ± 2.2	60.7 ± 3.6	54.8 ± 3.7	59.1 ± 2.7	61.2 ± 2.6	33.1 ± 3.2
DeepSeek-V3.2-Exp (Thinking)	53.2 ± 1.3	55.0 ± 5.5	36.9 ± 5.1	65.6 ± 3.1	61.9 ± 2.9	60.1 ± 3.4	58.7 ± 3.5	59.8 ± 2.3	27.8 ± 2.8
Qwen3.5-35B (Thinking)	52.0 ± 1.2	48.1 ± 4.4	52.5 ± 3.7	63.5 ± 2.5	60.1 ± 3.4	54.2 ± 4.6	63.5 ± 2.5	58.5 ± 2.5	15.9 ± 2.8
Kimi-K2	49.7 ± 1.3	47.2 ± 3.8	43.3 ± 5.3	60.9 ± 2.9	55.6 ± 4.0	49.7 ± 4.0	57.4 ± 2.3	51.3 ± 3.1	32.1 ± 3.3
Gemma-3-27B (Instruct)	47.8 ± 1.3	43.8 ± 4.6	28.1 ± 5.3	60.4 ± 2.9	55.4 ± 3.4	49.4 ± 3.7	60.1 ± 1.9	52.2 ± 2.8	32.8 ± 3.4
Qwen3-235B (Instruct)	44.1 ± 1.2	39.4 ± 3.8	35.0 ± 4.4	53.6 ± 2.6	52.4 ± 3.2	48.2 ± 3.3	49.5 ± 2.3	49.6 ± 3.4	25.3 ± 3.2
GPT-OSS-120B (High)	41.9 ± 1.3	35.8 ± 4.4	41.2 ± 5.5	55.7 ± 2.5	54.8 ± 2.5	39.3 ± 4.6	56.2 ± 2.7	44.2 ± 3.3	7.8 ± 2.1
MiniMax-M2.5	39.9 ± 1.2	30.6 ± 3.2	36.2 ± 4.3	57.8 ± 3.4	51.8 ± 3.6	42.9 ± 4.9	49.0 ± 3.0	43.3 ± 2.9	7.2 ± 1.8
Qwen3.5-35B (Instruct)	38.2 ± 1.3	32.5 ± 4.3	36.9 ± 4.1	49.5 ± 3.3	41.7 ± 3.4	40.5 ± 5.0	51.9 ± 2.7	40.2 ± 3.1	12.8 ± 2.2
MiMo-V2-Flash	30.2 ± 1.1	25.0 ± 3.8	14.4 ± 3.6	44.3 ± 3.2	41.7 ± 3.3	34.5 ± 3.5	40.9 ± 3.2	30.8 ± 2.8	10.3 ± 2.3
Qwen3-30B (Instruct)	21.4 ± 1.0	14.4 ± 2.5	19.4 ± 4.1	30.7 ± 3.0	28.0 ± 3.0	23.2 ± 3.4	25.5 ± 2.3	26.8 ± 2.4	3.4 ± 1.1
Nemotron-3-Nano	5.5 ± 0.5	5.6 ± 1.9	1.2 ± 0.8	10.4 ± 1.8	7.7 ± 1.6	3.0 ± 1.4	10.6 ± 1.9	5.8 ± 1.2	0.0 ± 0.0
Average	52.6	49.8	48.8	61.5	59.0	54.0	60.2	56.4	30.8

4.2 Main Results

Model performance. Gemini-3-Flash leads overall and in every linguistic subcategory (Table 2). The table reports mean performance and bootstrap standard errors from 10,000 sentence-level resamples after averaging across sequences. Informal text and STEM abstracts are the hardest categories (average scores 30.8 and 48.8). Qualitative examples in Appendix G.4 illustrate how Gemini-3-Flash preserves jokes, nuance, and informal syntax that other models lose during translation.

Table 3: $MQM_{\geq 80}$ **collapses on low-resource language sequences.** We report $MQM_{\geq 80}$ rates, the percentage of subset examples with MQM $\geq$ 80, across the same eight language sequences grouped by resource availability. (a) High-resource sequences remain solvable for the strongest frontier models, with Gemini-3-Flash averaging 97.0 and reaching 98.0 on both Central European and Near Eastern chains. (b) Medium-resource sequences remain relatively stable for the best models, though separation widens below the frontier tier. (c) Low-resource sequences collapse sharply: only Gemini-3-Flash sustains a substantial pass rate, averaging 59.2 across African and South American chains; the next-best model drops to 32.0. (d) Global averages across these eight sequences show Gemini-3-Flash far ahead of all competitors.

(a) High & Med-High Resource Sequences

Model	E. Asia	C. Europe	N. East	Avg
Gemini-3-Flash (No-Think)	95.0	**98.0**	**98.0**	**97.0**
GLM-5 (Thinking)	**93.5**	93.5	82.5	89.8
Qwen3.5-397B (Thinking)	91.0	92.0	85.5	89.5
Gemma-4-31B (Instruct)	90.5	90.5	86.0	89.0
Gemma-4-31B (Thinking)	**93.5**	87.5	86.0	89.0
GLM-4.7 (Thinking)	92.0	90.0	80.5	87.5
Qwen3.5-397B (Instruct)	88.5	91.5	82.0	87.3
Kimi-K2 (Thinking)	84.0	89.0	79.0	84.0
DeepSeek-V3.2-Exp	87.5	85.0	71.5	81.3
GLM-5 (Instruct)	87.5	86.0	69.5	81.0
Qwen3-235B (Thinking)	82.5	86.0	71.5	80.0
DeepSeek-V3.2-Exp (Think)	84.5	75.0	69.5	76.3
Qwen3.5-35B (Thinking)	76.0	83.0	65.5	74.8
Gemma-3-27B (Instruct)	70.0	79.5	68.5	72.7
Kimi-K2	71.5	78.0	56.5	68.7
Qwen3-235B (Instruct)	83.0	62.5	57.5	67.7
MiniMax-M2.5	60.0	75.5	58.5	64.7
Qwen3.5-35B (Instruct)	71.5	73.5	46.5	63.8
GPT-OSS-120B (High)	68.5	65.5	49.0	61.0
MiMo-V2-Flash	57.0	61.5	30.5	49.7
Qwen3-30B (Instruct)	56.5	49.0	21.5	42.3
Nemotron-3-Nano	29.5	4.0	0.0	11.2

(b) Medium Resource Sequences

Model	SE. Asia	S. Asia	N. Europe	Avg
Gemini-3-Flash (No-Think)	**93.0**	**95.5**	**95.5**	**94.7**
Qwen3.5-397B (Thinking)	82.0	77.0	87.0	82.0
Gemma-4-31B (Instruct)	92.5	90.0	61.5	81.3
Gemma-4-31B (Thinking)	88.5	93.0	62.5	81.3
GLM-5 (Thinking)	86.5	74.0	79.5	80.0
Qwen3.5-397B (Instruct)	83.5	71.0	75.0	76.5
GLM-4.7 (Thinking)	70.5	66.0	63.5	66.7
Kimi-K2 (Thinking)	71.5	58.5	65.0	65.0
Qwen3-235B (Thinking)	74.5	55.0	53.5	61.0
DeepSeek-V3.2-Exp	75.0	46.5	51.5	57.7
GLM-5 (Instruct)	63.5	57.0	50.0	56.8
DeepSeek-V3.2-Exp (Think)	69.5	45.5	53.0	56.0
Qwen3.5-35B (Thinking)	60.5	48.0	56.0	54.8
Gemma-3-27B (Instruct)	60.5	50.0	46.5	52.3
Kimi-K2	57.0	42.0	54.0	51.0
Qwen3-235B (Instruct)	69.0	48.0	22.5	46.5
GPT-OSS-120B (High)	49.0	37.5	42.5	43.0
MiniMax-M2.5	45.0	16.0	43.0	34.7
Qwen3.5-35B (Instruct)	46.0	21.0	30.5	32.5
MiMo-V2-Flash	41.5	18.0	18.5	26.0
Qwen3-30B (Instruct)	28.0	5.0	1.5	11.5
Nemotron-3-Nano	2.5	0.0	0.0	0.8

(c) Low Resource / Imbalanced Sequences

Model	Africa	S. America	Avg
Gemini-3-Flash (No-Think)	**78.5**	**40.0**	**59.2**
Gemma-4-31B (Instruct)	60.0	4.0	32.0
Qwen3.5-397B (Thinking)	41.0	15.0	28.0
Gemma-4-31B (Thinking)	51.5	3.0	27.2
GLM-5 (Thinking)	30.5	7.0	18.8
Qwen3.5-397B (Instruct)	27.5	9.0	18.2
DeepSeek-V3.2-Exp (Think)	8.5	6.0	7.2
GLM-4.7 (Thinking)	8.0	3.0	5.5
Nemotron-3-Nano	1.5	4.5	3.0
Kimi-K2 (Thinking)	5.0	0.5	2.8
Qwen3.5-35B (Thinking)	4.0	1.5	2.8
MiMo-V2-Flash	0.0	4.5	2.2
GPT-OSS-120B (High)	1.5	1.0	1.3
Qwen3.5-35B (Instruct)	1.0	1.5	1.2
Qwen3-235B (Thinking)	0.5	2.0	1.2
Gemma-3-27B (Instruct)	2.0	0.0	1.0
DeepSeek-V3.2-Exp	1.5	0.5	1.0
GLM-5 (Instruct)	2.0	0.0	1.0
MiniMax-M2.5	1.0	0.5	0.8
Kimi-K2	1.0	0.0	0.5
Qwen3-235B (Instruct)	0.5	0.0	0.2
Qwen3-30B (Instruct)	0.0	0.0	0.0

(d) Overall Performance

Model	Global Average
Gemini-3-Flash (No-Think)	**86.7**
Gemma-4-31B (Instruct)	71.9
Qwen3.5-397B (Thinking)	71.3
Gemma-4-31B (Thinking)	70.7
GLM-5 (Thinking)	68.4
Qwen3.5-397B (Instruct)	66.0
GLM-4.7 (Thinking)	59.2
Kimi-K2 (Thinking)	56.6
Qwen3-235B (Thinking)	53.2
DeepSeek-V3.2-Exp	52.4
GLM-5 (Instruct)	51.9
DeepSeek-V3.2-Exp (Think)	51.4
Qwen3.5-35B (Thinking)	49.3
Gemma-3-27B (Instruct)	47.1
Kimi-K2	45.0
Qwen3-235B (Instruct)	42.9
GPT-OSS-120B (High)	39.3
MiniMax-M2.5	37.4
Qwen3.5-35B (Instruct)	36.4
MiMo-V2-Flash	28.9
Qwen3-30B (Instruct)	20.2
Nemotron-3-Nano	5.2

Performance drops sharply from high- to low-resource settings, with Gemini-3-Flash leading in both (Table 3). It scores 97.0 in high-resource sequences, followed by GLM-5 (Thinking) at 89.8. In low-resource sequences, Gemini scores 59.2, followed by Gemma-4-31B (Instruct) at 32.0, while Qwen3-30B and Qwen3-235B score 0.0 and 0.2. Qwen3-235B does not officially support some languages in this subset, including Quechua and Amharic (Qwen Team, 2025).

Reasoning does not reliably improve translation overall. Its effects are model-dependent: it helps Qwen3-235B more than DeepSeek-V3.2-Exp, while consistently hurting informal translation. In low-resource settings, DeepSeek-V3.2-Exp improves from 1.0 without thinking to 7.2 with thinking, whereas Gemma-4-31B (Instruct) matches or outperforms its thinking variant, especially on low-resource sequences. Appendices A.4 and B.4 provide the extended category and sequence analysis.

Robustness checks. We assess the stability of these comparisons across challenging texts, judges, and translation-chain lengths, with supporting analyses in Appendix G. Correct back-translations can conceal errors in intermediate translations (Somers, 2005; van Zaanen & Zwarts, 2006), so we evaluate a separate robustness set of 480 samples covering five classical translation challenges. Model rankings remain strongly correlated with performance across LiT categories (Spearman's $\rho = 0.78$–0.96). Appendix C provides the construction details, full results, and qualitative analysis.

To assess sensitivity to judge choice, we compare Grok 4.1 Fast, DeepSeek-V4-Pro, and Gemini-3.1-Flash-Lite on both the main benchmark and this robustness extension. The comparison covers ten model configurations and all eight language sequences, with 200 and 60 source texts per sequence, respectively. Model-level $\text{MQM}_{\geq 80}$ scores and rankings agree strongly (Table 4; Pearson r, Spearman ρ), with Kendall's $W = 0.987$ across judges. Appendix G.1 lists the configurations.

Table 4: **Judge agreement.** Model-level $\text{MQM}_{\geq 80}$ correlations.

Judge pair	ρ	r
Grok 4.1 Fast / DeepSeek-V4-Pro	0.96	0.99
Grok 4.1 Fast / Gemini-3.1-Flash-Lite	0.99	0.99
DeepSeek-V4-Pro / Gemini-3.1-Flash-Lite	0.99	0.99

Finally, we vary chain length for six models across four sequences. Using DeepSeek-V4-Pro as judge, we compare full four-language chains with chains returning to English after the first one, two, or three intermediate languages. Two- and three-language chains retain strong rank agreement with full chains (Table 5). One-language round trips are near ceiling and have less consistent rank agreement; longer chains provide a more demanding test of meaning preservation. Appendix G.2 gives the models and pass-rate curves.

Table 5: **Chain-length agreement.** Pearson r / Spearman ρ across six models, comparing shorter chains with full four-language chains.

Sequence	1 vs. 4	2 vs. 4	3 vs. 4
C. Europe	0.99 / 0.94	0.96 / 0.83	0.98 / 0.83
East Asia	0.79 / 0.32	0.95 / 0.94	0.96 / 0.83
Near East	0.82 / 0.61	0.93 / 0.75	0.99 / 1.00
SE. Asia	0.57 / 0.03	0.94 / 0.77	0.93 / 0.89

5 Conclusion

Benchmarks like MT-AIME24 and INCLUDE, used by frontier models to claim multilingual progress, actually might be confounded by improving reasoning and factual recall performance. First, thinking variants dramatically outperform instruct variants on these benchmarks, yet perform no better (and often worse) on LMArena, where real users rate multilingual outputs. Second, our error analysis reveals that failures are logical and factual, not linguistic: models parse translated questions correctly but fail to solve them. LiT evaluates meaning preservation through round-trip translation and correlates positively with multilingual user preferences on LMArena. It spans high-resource to low-resource languages and shows underperformance of reasoning models in informal settings as well as substantial degradation on low-resource languages. We hope this work contributes to multilingual benchmarks that better measure preservation of meaning, fluency, and cultural context across languages. See Appendix A for limitations and future directions.

AI USE STATEMENT

We utilized Large Language Models (LLMs) to assist with the writing, polishing and formatting of the manuscript. We also partially utilized LLMs during the construction of benchmark samples, which were manually verified, and as the LLM-as-a-judge, as described in the main text. All other work, including the experiments, analyses, and formulation of claims, was performed by the authors.

ETHICS STATEMENT

LiT studies failures of meaning preservation in multilingual translation. These failures can distort factual content, cultural meaning, or speaker intent, with consequences for users who rely on translated text. The benchmark combines recent public paper abstracts with author-created texts, including LLM-assisted examples that were manually verified. Automated translation and judging may reflect linguistic and cultural biases, particularly for underrepresented languages. We report these failures and the limits of the evaluation to support assessment of multilingual systems; performance on LiT does not establish comprehensive language competence or safe deployment.

REPRODUCIBILITY STATEMENT

To enable reproducibility of our results, we provide the full details of our benchmark composition, samples and model evaluation in Section 4.1 and Appendix B. Appendices C, D, and G report robustness analyses and ablations. Section 3.1 documents the public LMArena sources, language coverage, and aggregation procedure. We are planning to open-source the benchmark, evaluation code, and other materials to facilitate independent verification and future research.

REFERENCES

Miltiadis Allamanis, Sheena Panthaplackel, and Pengcheng Yin. Unsupervised evaluation of code llms with round-trip correctness. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. Which of these best describes multiple choice evaluation with LLMs? a) forced B) flawed C) fixable D) all of the above. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3394–3418, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.169. URL https://aclanthology.org/2025.acl-long.169/.

Regina Barzilay and Mirella Lapata. Modeling local coherence: an entity-based approach. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pp. 141–148, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219858. URL https://doi.org/10.3115/1219840.1219858.

Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385/.

Richard W Brislin. Back-translation for cross-cultural research. *Journal of cross-cultural psychology*, 1(3):185–216, 1970.

Penelope Brown and Stephen C Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Zanzotto. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–222, 2013. ISSN 1947-4040. doi: 10.2200/S00509ED1V01Y201305HLT023. Publisher Copyright: © Morgan and Claypool Publishers. All rights reserved.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Examining the tip of the iceberg: A data set for idiom translation. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1148/.

Ana C Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, and André F. T. Martins. Findings of the WMT 2022 shared task on chat translation. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 724–743, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wmt-1.70. URL https://aclanthology.org/2022.wmt-1.70/.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz (eds.), *Proceedings of the Sixth Conference on Machine Translation*, pp. 733–774, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.73/.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. Scaling laws for neural machine translation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hR_SMu8cxCV.

H Paul Grice. Logic and conversation. *Syntax and Semantics*, 3:41–58, 1975.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Pierre Isabelle, Colin Cherry, and George Foster. A challenge set approach to evaluating machine translation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2486–2496, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1263. URL https://aclanthology.org/D17-1263/.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Marzena Karpinska and Mohit Iyyer. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 419–451, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.41. URL https://aclanthology.org/2023.wmt-1.41/.

Ahrii Kim. RUBRIC-MQM : Span-level LLM-as-judge in machine translation for high-end models. In Georg Rehm and Yunyao Li (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 147–165, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-288-6. doi: 10.18653/v1/2025.acl-industry.12. URL https://aclanthology.org/2025.acl-industry.12/.

Hannah Calzi Kleidermacher and James Zou. Science across languages: Assessing llm multilingual translation of scientific papers, 2025. URL https://arxiv.org/abs/2502.17882.

Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 768–775, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.64. URL https://aclanthology.org/2023.wmt-1.64/.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, et al. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pp. 355–413, 2025.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In Thang Luong, Alexandra Birch, Graham Neubig, and Andrew Finch (eds.), *Proceedings of the First Workshop on Neural Machine Translation*, pp. 28–39, Vancouver, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3204. URL https://aclanthology.org/W17-3204/.

Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4796, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1512. URL https://aclanthology.org/D18-1512/.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 436–483, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.24. URL https://aclanthology.org/2025.wmt-1.24/.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=KfTf9vFvSn.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, pp. 455–463, 2014. Issue 12.

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In Marianna Martindale, Janice Campbell, Konstantin Savenkov, and Shivali Goel (eds.), *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pp. 75–94, Chicago, USA, September 2024. Association for Machine Translation in the Americas. URL https://aclanthology.org/2024.amta-presentations.6/.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. Error analysis prompting enables human-like translation evaluation in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8801–8816, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.520. URL https://aclanthology.org/2024.findings-acl.520/.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. Revisiting round-trip translation for quality estimation. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada (eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 91–104, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.11/.

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. *Advances in Neural Information Processing Systems*, 37:98180–98212, 2024.

Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. MultiPragEval: Multilingual pragmatic evaluation of large language models. In Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Amirhossein Kazemnejad, Christos Christodoulopoulos, Mario Giulianelli, and Ryan Cotterell (eds.), *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pp. 96–119, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.genbench-1.7. URL https://aclanthology.org/2024.genbench-1.7/.

Qwen Team. Qwen3: Think deeper, act faster, April 2025. URL https://qwen.ai/blog?id=qwen3.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, and Azril Hafizi et al. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k3gCieTXeY.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2), January 2022. ISSN 0360-0300. doi: 10.1145/3485766. URL https://doi.org/10.1145/3485766.

Oscar Sainz, Iker García-Ferrero, Alon Jacovi, Jon Ander Campos, Yanai Elazar, Eneko Agirre, Yoav Goldberg, Wei-Lin Chen, Jenny Chim, Leshem Choshen, Luca D'Amico-Wong, Melissa Dell, Run-Ze Fan, Shahriar Golchin, Yucheng Li, Pengfei Liu, Bhavish Pahwa, Ameya Prabhu, Suryansh Sharma, Emily Silcock, Kateryna Solonko, David Stap, Mihai Surdeanu, Yu-Min Tseng, Vishaal Udandarao, Zengzhi Wang, Ruijie Xu, and Jinglin Yang. Data contamination report from the 2024 CONDA shared task. *CoRR*, abs/2407.21530, 2024. doi: 10.48550/ARXIV.2407.21530. URL https://doi.org/10.48550/arXiv.2407.21530.

John R Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 86–96, 2016.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=fR3wGCk-IXp.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, and Wei Qi et al. Leong. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.919. URL https://aclanthology.org/2025.acl-long.919/.

Harold Somers. Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop 2005*, pp. 127–133, 2005.

Guijin Son, Jiwoo Hong, Hyunwoo Ko, and James Thorne. Linguistic generalizability of test-time scaling in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, July 2025. doi: 10.18653/v1/2025.acl-long.699. URL https://aclanthology.org/2025.acl-long.699/.

Lucas Spangher, Tianle Li, William F. Arnold, Nick Masiewicki, Xerxes Dotiwalla, Rama Kumar Pasumarthi, Peter Grabowski, Eugene Ie, and Daniel Gruhl. Chatbot arena estimate: towards a generalized performance benchmark for LLM capabilities. In Weizhu Chen, Yi Yang, Mohammad Kachuee, and Xue-Yong Fu (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 1016–1025, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-194-0. doi: 10.18653/v1/2025.naacl-industry.77. URL https://aclanthology.org/2025.naacl-industry.77/.

Chihiro Taguchi, Seng Mai, Keita Kurabe, Yusuke Sakai, Georgina Agyei, Soudabeh Eslami, and David Chiang. Languages still left behind: Toward a better multilingual machine translation benchmark. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20131–20143, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1018. URL https://aclanthology.org/2025.emnlp-main.1018/.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

NLLB Team, Marta R. Costa-jussà, and James Cross et al. No language left behind: Scaling human-centered machine translation, 2022. URL https://arxiv.org/abs/2207.04672.

Menno van Zaanen and Simon Zwarts. Unsupervised measurement of translation quality using multi-engine, bi-directional translation. In *Proceedings of the 19th Australian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence*, AI'06, pp. 1208–1214, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540497870. doi: 10.1007/11941439_149. URL https://doi.org/10.1007/11941439_149.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. Prompting PaLM for translation: Assessing strategies and performance. In Anna Rogers, Jordan

Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15406–15427, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.859. URL https://aclanthology.org/2023.acl-long.859/.

Yiming Wang, Pei Zhang, Jialong Tang, Hao-Ran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. Polymath: Evaluating mathematical reasoning in multilingual contexts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=B1vCImy6yI.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=y10DM6R2r3.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. The bitter lesson learned from 2,000+ multilingual benchmarks. *arXiv preprint arXiv:2504.15521*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, and Dayiheng Liu et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. Benchmarking machine translation with cultural awareness. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13078–13096, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.765. URL https://aclanthology.org/2024.findings-emnlp.765/.

Z.ai, 5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, 2025. URL https://arxiv.org/abs/2508.06471.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

Terry Yue Zhuo, Qiongkai Xu, Xuanli He, and Trevor Cohn. Rethinking round-trip translation for machine translation evaluation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 319–337, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.22. URL https://aclanthology.org/2023.findings-acl.22/.

Part I
Appendix

CONTENTS

A Extended Discussion and Future Directions **18**
A.1 Benchmark Design . . . 18
A.2 Statistical Methodology and Evidence . . . 18
A.3 Automated Evaluation and Validity . . . 18
A.4 Detailed LiT Results . . . 18
A.5 Limitations and Future Directions . . . 19

B Extended Experimental Details **20**
B.1 Dataset Composition . . . 20
B.2 Language Sequences . . . 20
B.3 Evaluation and Scoring . . . 20
B.4 Detailed LiT Results . . . 21
B.5 Sampling Hyperparameters . . . 22
B.6 Model Prompt . . . 23
B.7 Judge Prompt . . . 23

C Additional Robustness Benchmark Details **26**
C.1 Results . . . 26

D Robustness of Metric and Judge Choice **29**
D.1 Results . . . 29

E Robustness of Answer Language Analysis Across Models **34**

F Robustness of Error Distribution Analysis Across Models **39**

G Additional Experiments and Ablations **41**
G.1 Model Configurations for Inter-Judge Agreement . . . 41
G.2 Translation-chain length . . . 41
G.3 Extended benchmark comparisons . . . 42
G.4 Qualitative examples for Gemini-3-Flash . . . 43

A Extended Discussion and Future Directions

A.1 Benchmark Design

While serial language sequences inherently compound translation errors, we deliberately designed this mechanism to strictly stress-test a model's multilingual capabilities. A capable multilingual model must maintain semantic integrity across multiple translation hops; the catastrophic degradation observed outside high-resource sequences successfully isolates the boundary of this capability. By evaluating 200 highly dense, paragraph-length source texts (Läubli et al., 2018), LiT prioritizes linguistic depth over superficial breadth to prevent benchmark saturation (Bowman & Dahl, 2021). This provides an evaluation setting (Wu et al., 2025) that probes linguistic phenomena often missed by sentence-level benchmarks.

A.2 Statistical Methodology and Evidence

We report rank correlations across six same-tier frontier configurations (n=6), constrained as described in Section 3.1 to avoid scale-driven confounding. We therefore interpret the correlations together with three additional pieces of evidence: (1) across all three model families, Thinking variants significantly outperform on existing benchmarks yet perform comparably or worse on LMArena, with effect sizes exceeding 20 percentage points on MT-AIME24; (2) our error taxonomy (Section 3.2) independently shows benchmark failures are logical and factual, not linguistic; and (3) the reasoning-language analysis (Figure 2c) confirms models default to English reasoning regardless of input language. We present ρ values for directional contrast, not as standalone hypothesis tests.

Finally, LMArena remains the main large-scale source of multilingual human-preference data (Chiang et al., 2024), which constrains the set of models with reliable cross-lingual ratings. Future work can strengthen the quantitative evidence as multilingual evaluation platforms expand.

A.3 Automated Evaluation and Validity

To keep LiT scalable and reproducible, we use an LLM-as-a-judge framework. While this removes direct human-in-the-loop verification, recent literature demonstrates the effectiveness of this approach (Kocmi & Federmann, 2023; Lu et al., 2024), strongly correlating with human preference. Crucially, we validate the quality of our automated judge by demonstrating its high rank-order correlation with real-world human preference ratings from LMArena (Chiang et al., 2024). This confirms that the semantic degradation penalized by our round-trip translation approach strictly aligns with the cross-lingual generation failures penalized by actual multilingual users.

A.4 Detailed LiT Results

A Linguistic Category View. We present our results in Table 2. The trends from the previous section also appear across LiT categories: reasoning effects are model-dependent, helping Qwen3-235B-A22B-2507 much more than DeepSeek-V3.2-Exp, while consistently hurting informal translation. Gemini-3-Flash achieves the highest overall score and the highest score in each subcategory. The high gap between Gemini and other models indicates that our benchmark can measure performance across the "long tail" of scenarios.

Category-wise breakdown. To understand failure modes, Table 2 decomposes performance by linguistic category. STEM Abstracts and Informal registers emerge as the most challenging categories, with average performance of 48.8 and 30.8, respectively. Each entry reports the mean and a bootstrap-estimated standard error, computed from $N = 10,000$ sentence-level resamples after averaging across sequences. Technical content tests whether models can communicate domain knowledge fluently. Informal text tests whether models handle slang and colloquialisms in translation. We observe a striking pattern: reasoning models underperform on informal text. One possible explanation is that reasoning models overcomplicate simple colloquial inputs, producing a mismatch in register.

A Language-Centric View. Table 3 presents our results grouped by sequence and by high/medium/low resource classification. Most notably, performance drops sharply from high-resource to low-resource settings. Medium-resource language sequences largely follow high-resource trends.

High-Resource Stability. In East Asian, Near Eastern, and Central European sequences, most frontier models achieve strong performance. As observed in Table 3, Gemini-3-Flash leads with an $MQM_{\geq 80}$ score of 97.0, followed by GLM-5 (Thinking) at 89.8. Most models perform well on these sequences.

Low-Resource Collapse. In contrast, performance collapses in African and South American sequences. Except for Gemini-3-Flash, which maintains a score of 59.2, Table 3 shows that the $MQM_{\geq 80}$ score of the next best model – Gemma-4-31B (Instruct) – drops to merely 32.0. Several models, including Qwen3-30B (0.0) and Qwen3-235B (0.2), receive near-zero scores, signalling critical errors and unusable translations. This confirms the official result that Qwen3-235B does not officially support some low-resource languages in this subset, such as Quechua or Amharic (Qwen Team, 2025).

Impact of Reasoning on Translation. Comparing thinking variants with instruct counterparts in Table 3 suggests that inference-time reasoning helps most in low-resource languages. The thinking variant of DeepSeek-V3.2-Exp scores 7.2 in low-resource settings, outperforming the instruct model's $MQM_{\geq 80}$ score of 1.0. These results suggest that reasoning cannot compensate for missing lexical knowledge, but may mitigate some severe errors in unfamiliar linguistic settings. More broadly, however, reasoning does not reliably improve translation quality: in high-resource settings it is often only comparable to instruct models, and even recently released models such as Gemma-4-31B show the instruct variant matching or outperforming the thinking variant overall, with the clearest advantage in low-resource settings.

Key Findings

- Reasoning often hurts translation quality, especially on informal text.
- A catastrophic accuracy cliff separates high-resource from low-resource languages.
- Reasoning models outperform instruct models primarily in low-resource languages

A.5 Limitations and Future Directions

While LiT resolves critical limitations in existing multilingual benchmarks, it is merely the first step toward more comprehensive evaluation paradigms. We highlight three high-impact directions for future work which could address current limitations:

Scaling to Document-Level Discourse. Frontier models are increasingly deployed for full-document tasks involving reports, literature, and legal text. While LiT effectively evaluates paragraph-level pragmatics, document-scale translation introduces complex global dependencies. Future benchmarks must evaluate the preservation of long-range referential chains, stylistic consistency, and persistent terminology memory across massive context windows.

Disentangling Single Language Performance. While our serial language sequences bound overall cross-lingual robustness, they also obscure single-language performance. To isolate exact failure modes, future work should separate these sequences into independent, controlled evaluations. This would enable more granular capability profiles for individual low-resource languages without the relatively fuzzy intermediate error cascading.

Efficient and Culturally Grounded Generation. An important direction is to develop efficient proxies that better reflect real-world human utility. While round-trip translation serves as an exceptional proxy for semantic preservation, future evaluation suites must expand beyond translation entirely. Developing highly efficient, native generation tasks that evaluate cultural grounding and target-language fluency, without relying on an English source or pivot, will be critical. Ultimately, combining round-trip evaluation with such native generative tasks will yield a comprehensive multilingual suite, moving the needle for the billions of users who rely on frontier models for global knowledge access and digitization.

B Extended Experimental Details

B.1 Dataset Composition

Data Curation. We create a new benchmark of 1600 samples (200 unique texts x 8 language sequences). We authored all tasks and non-abstract texts and the abstracts (20%) come from recent public papers to prevent data contamination. We manually select samples with diverse linguistic complexities. The dataset spans three major categories, each targeting distinct aspects of translation difficulty, as detailed below:

We provide more details on the composition of our proposed LiT benchmark:

(a) Abstracts (20% samples): We divide this category equally between Humanities (10%) and STEM (10%) abstracts. Abstracts are useful benchmark units because they are self-contained, semantically dense, and complete, so errors are easier to detect (Isabelle et al., 2017; Kleidermacher & Zou, 2025). Humanities abstracts test whether models can preserve argumentative nuance and rhetorical force, while STEM abstracts test terminological precision and mathematical notation, where minor mistakes carry major consequences.

(b) Pragmatics (60% samples): This category tests whether models preserve meaning beyond literal content (Park et al., 2024). We subdivide it into five phenomena: *(i) Core Semantics* (20% samples): tests preservation of truth conditions and logical entailment (Dagan et al., 2013). *(ii) Discourse Coherence* (17.5% samples): evaluates maintenance of referential chains, topic continuity, and logical connectives across sentence boundaries Barzilay & Lapata (2005). *(iii) Implicit Content* (17.5% samples): probes handling of presuppositions, implicatures, and information that speakers convey without explicitly stating it (Grice, 1975). *(iv) Pragmatic Inference* (21.7% samples): tests understanding of speech acts, speaker intent, and context-dependent meaning (Searle, 1969). *(v) Social Interaction* (23.3% samples): evaluates preservation of politeness markers, formality levels, and sociolinguistic appropriateness (Brown & Levinson, 1987).

(c) Informal (20% samples): This category tests colloquial language, slang, idioms, and register shifts (Koehn & Knowles, 2017; Fadaee et al., 2018). Informal text requires preserving tone and social function, not just denotative meaning.

B.2 Language Sequences

To rigorously stress-test multilingual capabilities, we evaluate performance across 8 linguistic sequences. We select these sequences to span disjoint language families, distinct scripts (Latin, Cyrillic, Devanagari, Arabic, CJK), and varying levels of pretraining resource availability. We prioritize translation pairs between regions with frequent interactions to reflect real-world translation tasks. Each sequence starts with an English source text, which is serially translated through four intermediate languages before returning to English:

- *East Asia (E.Asia, High resource)*: Japanese → Korean → Chinese → Russian.
- *Central Europe (C.Europe, High resource)*: Romanian → Hungarian → German → Polish.
- *Near East (N.East, High resource)*: Bulgarian → Greek → Turkish → Persian.
- *North Europe (N.Europe, Medium resource)*: Icelandic → Swedish → Finnish → Lithuanian.
- *Southeast Asia (SE.Asia, Medium resource)*: Chinese → Thai → Vietnamese → Tagalog.
- *South Asia (S.Asia, Medium resource)*: Hindi → Tamil → Bengali → Punjabi.
- *Africa (Africa, Low resource)*: Swahili → Arabic → Hausa → Amharic.
- *South America (S.America, Low resource)*: Portuguese → Quechua → Spanish → Guarani.

B.3 Evaluation and Scoring

Evaluation: LLM-as-a-Judge. We employ Grok 4.1 Fast as our primary judge model, being an inexpensive, fast but simultaneously very capable model – consistently one of the highest ranked in multilingual LMArena (Chiang et al., 2024). Following standard practice Lommel et al. (2014); Kim (2025), we adopt the MQM framework, a weighted penalty system that categorizes errors by severity: (i) *Minor* (-1): Slight awkwardness or non-critical fluency issues that do not affect comprehension; (ii) *Major* (-5): Significant semantic shifts, structural failures, or mistranslations that alter meaning; (iii) *Critical* (-25): Complete loss of meaning, hallucinations, or safety-relevant

errors. The framework grounds evaluation in interpretable error categories, where low MQM scores indicate severe failure modes dominated by critical errors, with 80 widely considered to be a pass threshold (Farinha et al., 2022; Lommel et al., 2024). We additionally show in Appendix D that both the scores and rankings nearly perfectly correlate when using two alternatives: (i) raw MQM scores and (ii) an LLM judge providing a score from 0-100, demonstrating robustness to the metrics. The Appendix results additionally show robustness to the judge model, allowing us to fix the metric to $\text{MQM}_{\geq 80}$ and the Grok 4.1 Fast judge model for the rest of the section.

B.4 Detailed LiT Results

B.4.1 Linguistic Categories

We present our results in Table 2. Reasoning effects are model-dependent, helping Qwen3-235B-A22B-2507 much more than DeepSeek-V3.2-Exp, while consistently hurting informal translation. Gemini-3-Flash achieves the highest overall score and the highest score in each subcategory. The high gap between Gemini and other models indicates that our benchmark can measure performance across the "long tail" of scenarios.

Category-wise breakdown. To understand failure modes, Table 2 decomposes performance by linguistic category. STEM Abstracts and Informal registers emerge as the most challenging categories, with average performance of 48.8 and 30.8, respectively. Each entry reports the mean and a bootstrap-estimated standard error, computed from $N = 10,000$ sentence-level resamples after averaging across sequences. Technical content tests whether models can communicate domain knowledge fluently. Informal text tests whether models handle slang and colloquialisms in translation. We observe a striking pattern: reasoning models underperform on informal text. One possible explanation is that reasoning models overcomplicate simple colloquial inputs, producing a mismatch in register.

B.4.2 Language Sequences

Table 3 presents our results grouped by sequence and by high/medium/low resource classification. Most notably, performance drops sharply from high-resource to low-resource settings. Medium-resource language sequences largely follow high-resource trends.

High-Resource Stability. In East Asian, Near Eastern, and Central European sequences, most frontier models achieve strong performance. As observed in Table 3, Gemini-3-Flash leads with an $MQM_{\geq 80}$ score of 97.0, followed by GLM-5 (Thinking) at 89.8. Most models perform well on these sequences.

Low-Resource Collapse. In contrast, performance collapses in African and South American sequences. Except for Gemini-3-Flash, which maintains a score of 59.2, Table 3 shows that the $MQM_{\geq 80}$ score of the next best model – Gemma-4-31B (Instruct) – drops to merely 32.0. Several models, including Qwen3-30B (0.0) and Qwen3-235B (0.2), receive near-zero scores, signalling critical errors and unusable translations. This confirms the official result that Qwen3-235B does not officially support some low-resource languages in this subset, such as Quechua or Amharic (Qwen Team, 2025).

B.4.3 Reasoning Effects

Comparing thinking variants with instruct counterparts in Table 3 suggests that inference-time reasoning helps most in low-resource languages. The thinking variant of DeepSeek-V3.2-Exp scores 7.2 in low-resource settings, outperforming the instruct model's $MQM_{\geq 80}$ score of 1.0. These results suggest that reasoning cannot compensate for missing lexical knowledge, but may mitigate some severe errors in unfamiliar linguistic settings. More broadly, however, reasoning does not reliably improve translation quality: in high-resource settings it is often only comparable to instruct models, and even recently released models such as Gemma-4-31B show the instruct variant matching or outperforming the thinking variant overall, with the clearest advantage in low-resource settings.

Summary. Reasoning often hurts translation quality, especially on informal text. A catastrophic accuracy cliff separates high-resource from low-resource languages. Reasoning models outperform instruct models primarily in low-resource languages.

B.5 SAMPLING HYPERPARAMETERS

Table 6: Hyperparameter details for round-trip translation.

Model	Temp.	Top-p	Effort
Gemini-3-Flash (No-Thinking)	1.0	0.95	–
GLM-4.7	1.0	0.95	–
GLM-5 (Thinking)	1.0	0.95	–
GLM-5 (Instruct)	1.0	0.95	–
Qwen3-235B (Thinking)	0.6	0.95	–
Qwen3-235B (Instruct)	0.7	0.8	–
Qwen3-30B (Instruct)	0.7	0.8	–
Qwen3.5-397B (Thinking)	0.6	0.95	–
Qwen3.5-397B (Instruct)	0.7	0.8	–
Qwen3.5-35B (Thinking)	0.6	0.95	–
Qwen3.5-35B (Instruct)	0.7	0.8	–
Kimi-K2 (Thinking)	1.0	1.0	–
Kimi-K2 (Instruct)	1.0	1.0	–
DeepSeek-V3.2 (Thinking)	1.0	0.95	–
DeepSeek-V3.2 (Instruct)	1.0	0.95	–
GPT-OSS-120B	1.0	1.0	High
MiniMax-M2.5	1.0	0.95	–
MiMo-V2-Flash (Instruct)	1.0	0.95	–
Nemotron-3-Nano	0.6	0.95	–
Gemma-3-27B (Instruct)	1.0	0.95	–
Gemma-4-31B (Instruct)	1.0	0.95	–
Gemma-4-31B (Thinking)	1.0	0.95	–

Table 7: Hyperparameter details for MT-AIME24 and INCLUDE

Model	Temp.	Top-p	Max Tokens		Effort
			AIME24	INCLUDE	
Gemini-3-Flash	1.0	0.95	38,912	32,768	–
GLM-4.7	1.0	0.95	38,912	32,768	–
Qwen3-235B (Thinking)	0.6	0.95	38,912	32,768	–
Kimi-K2 (Thinking)	1.0	1.0	131,072	32,768	–
DeepSeek-V3.2-Exp (Think)	1.0	0.95	38,912	32,768	–
DeepSeek-V3.2-Exp	1.0	0.95	38,912	32,768	–
GPT-OSS-120B	1.0	1.0	38,912	32,768	High
MiMo-V2-Flash	1.0	0.95	38,912	32,768	–
Nemotron-3-Nano	0.6	0.95	38,912	32,768	–
Gemma-3-27B (Instruct)	1.0	0.95	38,912	32,768	–
Qwen3-235B (Instruct)	0.7	0.8	38,912	32,768	–
Qwen3-30B (Instruct)	0.7	0.8	38,912	32,768	–

The hyperparameters used for the model sampling utilizing OpenRouter are presented in Table 6 for the LiT benchmark and Table 7 for MT-AIME24 and INCLUDE. We follow the official technical reports for the corresponding models and use the official sampling parameters unless specified otherwise. If no suggested default is officially given, we set the temperature to a default of 1.0 and Top-p sampling to 0.95.

B.6 MODEL PROMPT

We provide the translation prompt used in our evaluation, for reproducibility of our pipeline:

System Prompt

```
You are a professional translator.
Task:  Translate the SOURCE TEXT from {src_lang} into {target_lang}.
Instructions:
     • Use natural, idiomatic {target_lang}, avoid unnatural
       word-for-word translation.
     • Preserve meaning, tone, and register.  Do not add, omit, or
       summarize.
     • Output the translated text in {target_lang}.  Do not include
       any additional text.
SOURCE TEXT: "{text}"
```

B.7 JUDGE PROMPT

We provide the judge prompt used in our evaluation, for reproducibility of our pipeline:

System Prompt

You are an annotator and expert linguist. Your task is to evaluate the back-translation given the original text and output a score, classification of machine translation quality, and a list of issues found.
Score measures back-translation quality on a continuous scale from 0 to 100, where a score of zero means "little-to-no meaning preserved" and score of one hundred means "perfect meaning preserved".
The classification of the quality of machine translation must be into one of following 10 categories:

- '1-nonsense': The text is gibberish, in the wrong language, or completely unrelated to the source.
- '2-severe distortion': Contains unrecognizable fragments. The core meaning is often lost or dangerously misleading.
- '3-failed gist': The topic is broadly correct, but the translation is mostly misleading or incomprehensible due to severe errors.
- '4-unreliable': Meaning is often preserved but significant meaning errors (critical mistranslations) are present.
- '5-machine-like': The meaning is roughly preserved (no critical errors), but the phrasing is overly literal ("translationese"), awkward, or grammatically poor.
- '6-understandable but Flawed': Meaning is preserved. Grammar is mostly functional but contains distracting errors or very unnatural stiffness.
- '7-good': Accurate meaning. Grammatically correct with only minor, non-impeding errors (e.g., wrong punctuation, slight awkwardness).
- '8-very good': Fluent and accurate. No grammatical errors. However, it may miss minor nuances of tone or style found in the Reference.
- '9-excellent': Native-level fluency. Captures the exact meaning and tone. Indistinguishable from professional human translation.
- '10-perfect': Flawless. Captures distinct cultural nuances, idioms, and subtext perfectly. Equivalent to the reference.

The list of issues is a comprehensive list of errors based on the following MQM Core dimensions: Accuracy, Fluency, and Terminology and Style/Locale.

1. Accuracy: (Mistranslation, Omission, Addition, Untranslated). Does the target text accurately reflect the source meaning?
2. Fluency: (Grammar, Spelling, Punctuation, Unintelligible). Is the target text linguistically correct and natural?
3. Terminology: (Inconsistent, Wrong Term). Does it adhere to domain standards?
4. Style/Locale: Does it follow local formats (dates, currencies) and cultural norms? Does the translation match the required formality/register (e.g., formal vs. casual)?

The three severity categories are:

- 'minor': Has a limited impact on accuracy, stylistic quality, consistency, fluency, clarity, or general appeal of the content.
- 'major': Seriously affects the understandability, reliability, or usability of the content for its intended purpose. For example, it causes significant loss or change in meaning or because the error appears in a highly visible or important part of text.

```
    • 'critical': Hallucination, completely changes meaning or
      catastrophic failure rendering the sentence unusable or poses
      serious reputational harm.
Issues is a list of tuples of each issue being a tuple of (severity
category, issue description).
Output Format: You must output a single valid JSON object. Do not
include markdown formatting (like "`json) or conversational text. The
JSON must follow this schema: {"score": <0-100>, "classification":
"<one of the 10 quality categories>", "issues": [{"severity":"< one
of three severity categories>", "description":"<issue description>"} ,
...]}
Original text: "{original}"
Back-translation: "{translation}"
```

C Additional Robustness Benchmark Details

While LiT uses round-trip translation, the method has several known limitations (Somers, 2005; van Zaanen & Zwarts, 2006). A correct round-trip translation does not necessarily imply correct intermediate translations, which may contain awkward phrasing, inappropriate formality, register mismatches, or missing cultural nuance that disappear in the final back-translation. To assess whether these classical concerns still affect current models, we evaluate a separate 480-sample robustness benchmark. Our results show that modern LLMs are largely robust to these failure modes, and that round-trip translation remains informative on such challenging cases.

Dataset Construction. We construct a dedicated robustness extension of our LiT benchmark consisting of 480 samples (60 source sentences $\times$ 8 language sequences), targeting five categories of classically challenging phenomena:

(i) Conceptual and Abstract Nuance: Passages where physical or spatial vocabulary is used metaphorically to describe abstract concepts (e.g., "a heavy decision"), testing whether models preserve figurative meaning across languages.

(ii) Idioms and Cultural Metaphors: Figurative expressions whose meaning cannot be recovered from literal word-by-word translation, testing whether models preserve communicative intent rather than surface form.

(iii) Polysemy and Lexical Ambiguity: Words carrying multiple distinct meanings where only surrounding context disambiguates the intended sense (e.g., "bank" as financial institution vs. riverbank).

(iv) Register and Tone Shifts: Passages that shift between formal, colloquial, and technical registers within a single example, testing sensitivity to sociolinguistic appropriateness.

(v) Syntactic Complexity and Garden Paths: Sentences whose grammatical structure remains ambiguous until late in the sentence, forcing re-parsing and testing whether models maintain structural fidelity through the translation chain.

C.1 Results

Results. Table 8 reports category-wise $MQM_{\geq 80}$ scores on the robustness subset. Idioms & Cultural Metaphors emerge as the most challenging category (35.3% on average), consistent with the difficulty of preserving figurative meaning across typologically diverse languages, whereas Conceptual & Abstract Nuance is the easiest (54.8%), suggesting that frontier models handle metaphorical extension more reliably than fixed idiomatic expressions. Although these items were specifically designed to stress classical failure modes of round-trip translation, model rankings remain highly consistent with the main LiT benchmark: Table 9 shows that the Spearman correlation between each LiT category and the robustness benchmark average is significant at $p < 0.001$ for all categories, ranging from $\rho = 0.78$ for Informal to $\rho = 0.96$ for Core Semantics. As in the main benchmark, we report mean performance together with bootstrap-estimated standard errors.

Manual inspection helps explain this robustness: when an intermediate translation is awkward or overly literal, stronger models often preserve that awkwardness rather than correct it, because they follow the instruction to translate faithfully. In this sense, instruction-following models can faithfully propagate errors instead of repairing them, suggesting that round-trip translation is less vulnerable to some classical critiques and can still probe whether models genuinely understand idiomatic or structurally complex constructions.

Implications. These results suggest that round-trip translation, when applied to instruction-following LLMs, largely overcomes the classical criticisms against it. The benchmark successfully surfaces genuine cross-lingual weaknesses (e.g., idiom handling, register preservation) rather than being artificially derailed by the phenomena it was designed to probe. The high rank-order consistency with the main benchmark (Table 9) further validates that robustness cases do not introduce a separate, orthogonal axis of difficulty, but rather stress-test the same underlying multilingual generation capabilities measured by the main LiT benchmark.

Table 8: $MQM_{\geq 80}$ on the Backtranslation robustness subset, which targets classical round-trip translation challenges: idioms, polysemy, register shifts, and garden-path syntax. We report the percentage of sentence-sequence translations with MQM $\geq$ 80, aggregated within each category over the same eight translation sequences. Each cell shows the mean and bootstrap-estimated standard error, computed via sentence-level resampling after averaging across sequences. Higher is better. Each category contains 12 source sentences. Idioms & Cultural Metaphors is the hardest category overall (35.3 avg), while Conceptual & Abstract Nuance is the easiest (54.8).

Model	Average	Conceptual & Abstract Nuance	Idioms & Cultural Metaphors	Polysemy & Lexical Ambiguity	Register & Tone Shifts	Syntactic Complexity & Garden Paths
Gemini-3-Flash (No-Thinking)	71.0 ± 2.1	83.3 ± 3.4	65.6 ± 5.9	68.8 ± 5.4	60.4 ± 5.3	77.1 ± 2.5
Qwen3.5-397B (Thinking)	63.5 ± 2.0	75.0 ± 3.6	62.5 ± 5.9	59.4 ± 4.2	58.3 ± 3.4	62.5 ± 4.2
GLM-5 (Thinking)	59.8 ± 2.0	74.0 ± 2.7	54.2 ± 4.0	50.0 ± 6.9	57.3 ± 3.5	63.5 ± 4.0
Gemma-4-31B (Thinking)	55.8 ± 2.4	75.0 ± 3.6	47.9 ± 6.0	54.2 ± 6.2	55.2 ± 4.7	46.9 ± 5.9
Kimi-K2 (Thinking)	54.6 ± 1.8	67.7 ± 2.3	49.0 ± 4.8	55.2 ± 3.6	50.0 ± 5.1	51.0 ± 3.3
Gemma-4-31B (Instruct)	52.5 ± 2.2	70.8 ± 3.7	44.8 ± 4.0	49.0 ± 5.1	53.1 ± 6.1	44.8 ± 5.2
Qwen3.5-397B (Instruct)	51.9 ± 2.1	70.8 ± 3.7	38.5 ± 5.0	56.2 ± 3.5	52.1 ± 5.5	41.7 ± 5.3
GLM-4.7	48.8 ± 2.0	61.5 ± 2.7	43.8 ± 5.6	47.9 ± 4.6	43.8 ± 5.3	46.9 ± 4.2
Qwen3-235B (Thinking)	47.5 ± 2.1	60.4 ± 3.2	42.7 ± 5.6	45.8 ± 5.9	37.5 ± 5.1	51.0 ± 3.4
Qwen3.5-35B (Thinking)	43.5 ± 2.4	57.3 ± 4.5	40.6 ± 5.5	38.5 ± 5.8	43.8 ± 5.9	37.5 ± 5.1
DeepSeek-V3.2-Exp (Thinking)	41.0 ± 2.4	57.3 ± 4.6	29.2 ± 5.0	43.8 ± 6.6	34.4 ± 4.2	40.6 ± 6.1
DeepSeek-V3.2-Exp	39.2 ± 2.2	53.1 ± 4.2	34.4 ± 5.6	31.2 ± 5.2	32.3 ± 5.2	44.8 ± 4.6
MiniMax-M2.5	38.8 ± 2.5	43.8 ± 4.8	33.3 ± 5.4	41.7 ± 5.9	32.3 ± 6.8	42.7 ± 4.3
Gemma-3-27B (Instruct)	37.5 ± 1.9	57.3 ± 3.1	31.2 ± 4.3	34.4 ± 5.2	37.5 ± 4.4	27.1 ± 4.2
GLM-5 (Instruct)	36.9 ± 2.1	50.0 ± 4.9	26.0 ± 4.5	34.4 ± 5.6	30.2 ± 4.0	43.8 ± 4.0
GPT-OSS-120B (High)	34.6 ± 2.4	37.5 ± 5.1	37.5 ± 3.6	30.2 ± 5.9	38.5 ± 5.8	29.2 ± 5.7
Kimi-K2	32.1 ± 2.0	51.4 ± 4.8	22.5 ± 4.1	30.5 ± 5.6	24.6 ± 3.8	31.5 ± 4.3
Qwen3-235B (Instruct)	31.9 ± 2.3	41.7 ± 5.2	26.0 ± 5.6	33.3 ± 5.6	32.3 ± 5.2	26.0 ± 4.1
Qwen3.5-35B (Instruct)	24.6 ± 2.0	44.8 ± 6.2	17.7 ± 4.0	21.9 ± 3.9	19.8 ± 3.7	18.8 ± 4.4
MiMo-V2-Flash	24.4 ± 2.2	39.6 ± 5.9	15.6 ± 3.0	22.9 ± 5.6	13.5 ± 4.0	30.2 ± 5.6
Qwen3-30B (Instruct)	14.0 ± 1.3	25.0 ± 2.9	7.3 ± 2.8	11.5 ± 2.7	12.5 ± 2.5	13.5 ± 4.0
Nemotron-3-Nano	5.2 ± 0.8	8.3 ± 1.7	5.2 ± 1.8	1.0 ± 1.0	5.2 ± 2.7	6.2 ± 1.8
Average	41.3	54.8	35.3	39.2	37.5	39.9

Table 9: Rank correlation between LiT categories and the robustness benchmark under $MQM_{\geq 80}$. We report Spearman rank correlations between model performance on each LiT category and the robustness benchmark average, using two-sided tests over all models.

LiT Category	**Spearman ρ**	**p-value**	**Significance**
Average	0.946	3.11×10^{-10}	***
Humanities	0.917	1.27×10^{-8}	***
STEM	0.901	5.98×10^{-8}	***
Core Semantics	0.959	2.46×10^{-11}	***
Discourse Coherence	0.953	8.98×10^{-11}	***
Implicit Content	0.954	7.78×10^{-11}	***
Pragmatic Inference	0.949	1.88×10^{-10}	***
Social Interaction	0.940	7.46×10^{-10}	***
Informal	0.777	5.49×10^{-5}	***

D ROBUSTNESS OF METRIC AND JUDGE CHOICE

A potential concern with any LLM-as-a-judge evaluation is sensitivity to the choice of metric formulation and judge model. If model rankings shifted substantially under alternative scoring rubrics, the benchmark's conclusions would be fragile. We address this concern by evaluating all models under two complementary scoring metrics and verifying that rankings remain stable across them.

Metrics. In the main paper, we report $\text{MQM}_{\geq 80}$: the percentage of translations whose MQM score meets or exceeds the 80-point threshold widely considered to indicate fit-for-purpose translation quality. We additionally evaluate two alternative metrics:

(i) Raw MQM scores (Tables 10 and 12): Rather than binarizing at the 80-point threshold, we report the average continuous MQM score for each model–sequence combination. This preserves the full distribution of translation quality and avoids potential artifacts introduced by a fixed cutoff.

(ii) Direct judge scores (Tables 11 and 13): We prompt the judge model to assign a holistic quality score on a continuous 0–100 scale, without reference to the MQM error taxonomy. This tests whether the structured MQM framework and a simple judge-based scoring rubric converge on the same model ordering.

D.1 RESULTS

Comparing global averages across the three scoring paradigms reveals highly stable rankings. The top tier is unchanged across all three metrics: Gemini-3-Flash leads decisively, followed by Gemma-4-31B (Instruct) and Qwen3.5-397B (Thinking). The bottom tier is equally stable, with Nemotron-3-Nano and Qwen3-30B (Instruct) consistently occupying the lowest positions. Mid-table models exhibit only modest reshuffling (typically within 1–2 rank positions), which is expected given that these models perform similarly and minor scoring differences can reorder near-tied entries.

The language-sequence breakdown further confirms robustness. All three metrics agree on the central finding: performance collapses catastrophically from high-resource to low-resource language sequences. Under raw MQM (Table 12), only Gemini-3-Flash maintains a clearly usable average score (75.4) on low-resource sequences, while the next-best model drops to 43.7. Under judge scores (Table 13), the same pattern holds, with Gemini-3-Flash at 79.4 and the runner-up at 64.4. The qualitative conclusion – that a steep accuracy cliff separates high-resource from low-resource performance – is invariant to the metric.

Summary. Across three metrics ($\text{MQM}_{\geq 80}$, Raw MQM, and direct judge scores) and validated against an external human-preference benchmark (LMArena), model rankings on LiT remain highly consistent. This stability indicates that our findings – including the disconnect between reasoning benchmarks and multilingual proficiency, the low-resource performance collapse, and the underperformance of reasoning models on informal text – are robust properties of the models themselves, not artifacts of a particular evaluation configuration.

Table 10: LiT benchmark by linguistic category under MQM. We report mean MQM scores (higher is better), aggregated within each category over the same eight translation sequences. Each cell shows the mean and the bootstrap-estimated standard error, computed via sentence-level resampling after averaging across sequences. Informal text is the hardest category overall (36.7 avg), while Core Semantics is the easiest (64.4).

Model	Average	Abstracts		Pragmatics					Informal
		Humanities	STEM	Core Semantics	Discourse	Implicit	Inference	Social	
Gemini-3-Flash (No-Thinking)	86.9 ± 0.3	88.6 ± 0.8	86.4 ± 0.9	88.8 ± 0.8	87.7 ± 0.8	85.2 ± 1.0	87.5 ± 0.8	87.5 ± 0.7	83.8 ± 0.8
Qwen3.5-397B (Thinking)	78.1 ± 0.6	76.9 ± 2.1	87.5 ± 1.1	85.2 ± 1.1	80.8 ± 1.4	80.0 ± 1.8	79.3 ± 1.5	79.0 ± 1.2	56.5 ± 2.2
Gemma-4-31B (Instruct)	75.2 ± 0.5	77.4 ± 1.5	82.0 ± 1.8	75.6 ± 1.6	77.8 ± 1.6	72.3 ± 1.8	76.8 ± 1.2	75.3 ± 1.1	64.7 ± 1.6
Gemma-4-31B (Thinking)	73.7 ± 0.6	77.2 ± 1.5	81.6 ± 1.7	74.9 ± 1.8	75.0 ± 1.5	71.3 ± 2.3	74.0 ± 1.5	73.6 ± 1.3	62.4 ± 1.9
GLM-5 (Thinking)	73.5 ± 0.7	76.1 ± 2.1	79.9 ± 2.1	79.9 ± 1.5	77.8 ± 1.3	74.0 ± 2.3	74.1 ± 1.7	74.6 ± 1.8	51.4 ± 2.3
Qwen3.5-397B (Instruct)	72.5 ± 0.8	76.1 ± 2.4	78.6 ± 2.9	76.6 ± 2.1	73.2 ± 1.7	73.5 ± 2.4	74.6 ± 1.4	72.8 ± 1.4	54.4 ± 2.9
GLM-4.7 (Thinking)	63.2 ± 0.8	67.1 ± 2.6	69.4 ± 2.4	68.5 ± 2.3	67.5 ± 1.8	64.4 ± 2.3	65.8 ± 1.6	62.6 ± 1.8	40.1 ± 2.1
DeepSeek-V3.2-Exp (Thinking)	62.9 ± 0.9	64.8 ± 3.0	51.3 ± 4.7	72.4 ± 1.7	69.2 ± 1.6	67.6 ± 2.4	66.0 ± 2.1	69.0 ± 1.4	43.1 ± 2.7
Kimi-K2 (Thinking)	61.5 ± 0.8	58.4 ± 3.2	59.9 ± 3.9	70.0 ± 1.3	67.8 ± 1.8	63.2 ± 2.0	66.8 ± 1.5	63.5 ± 1.4	42.5 ± 2.5
DeepSeek-V3.2-Exp	59.2 ± 1.0	63.1 ± 3.0	49.7 ± 4.9	64.2 ± 1.8	61.7 ± 2.3	62.9 ± 2.6	65.1 ± 1.5	62.1 ± 1.6	44.6 ± 2.4
Gemma-3-27B (Instruct)	58.8 ± 1.1	59.2 ± 2.9	33.7 ± 6.9	69.1 ± 1.3	67.5 ± 1.2	61.5 ± 2.0	69.3 ± 1.3	63.4 ± 1.9	47.1 ± 2.0
Qwen3.5-35B (Thinking)	57.8 ± 1.0	55.0 ± 3.3	67.0 ± 3.0	66.9 ± 2.0	64.5 ± 1.7	60.2 ± 3.9	65.2 ± 1.2	59.7 ± 1.9	24.3 ± 3.5
GLM-5 (Instruct)	57.6 ± 1.0	56.9 ± 3.7	56.3 ± 4.9	65.9 ± 1.8	60.1 ± 2.3	60.7 ± 2.4	60.6 ± 1.9	58.1 ± 1.8	41.9 ± 2.2
Qwen3-235B (Thinking)	54.9 ± 0.8	53.2 ± 2.4	56.4 ± 3.2	62.8 ± 1.4	60.8 ± 1.8	56.8 ± 2.8	62.0 ± 1.0	55.8 ± 1.5	31.5 ± 2.5
Kimi-K2	52.1 ± 1.1	47.6 ± 4.0	38.5 ± 5.8	61.0 ± 1.6	56.8 ± 2.2	55.5 ± 2.5	59.3 ± 1.6	56.4 ± 1.8	41.9 ± 2.8
GPT-OSS-120B (High)	52.1 ± 0.9	48.7 ± 3.1	61.2 ± 2.5	62.4 ± 1.5	60.0 ± 2.6	53.7 ± 4.1	58.7 ± 1.4	52.8 ± 2.0	19.3 ± 3.3
Qwen3-235B (Instruct)	49.4 ± 0.9	45.5 ± 3.0	42.4 ± 4.4	58.9 ± 1.5	52.9 ± 1.9	53.6 ± 1.7	53.1 ± 1.7	54.4 ± 1.8	34.3 ± 2.5
Qwen3.5-35B (Instruct)	49.4 ± 1.0	45.1 ± 3.2	51.9 ± 4.3	56.3 ± 1.7	55.6 ± 1.9	51.1 ± 3.4	60.5 ± 1.5	52.9 ± 1.8	21.4 ± 3.6
MiniMax-M2.5	47.2 ± 1.0	43.4 ± 3.6	50.0 ± 3.6	61.3 ± 1.9	58.7 ± 1.9	52.1 ± 3.8	54.2 ± 1.9	51.7 ± 1.9	6.5 ± 3.6
MiMo-V2-Flash	38.0 ± 1.0	33.8 ± 2.3	12.2 ± 5.4	51.7 ± 1.9	52.1 ± 2.7	44.8 ± 2.5	48.9 ± 1.8	43.8 ± 1.3	16.6 ± 3.0
Qwen3-30B (Instruct)	31.1 ± 1.1	24.8 ± 3.7	19.3 ± 5.4	42.3 ± 2.2	40.2 ± 3.0	39.4 ± 2.2	39.7 ± 1.5	37.8 ± 2.1	5.1 ± 2.5
Nemotron-3-Nano	-6.3 ± 1.2	-11.9 ± 3.2	-7.2 ± 5.0	1.2 ± 3.0	0.2 ± 2.8	-5.0 ± 3.9	0.4 ± 2.9	-1.4 ± 2.5	-26.6 ± 3.4
Average	56.8	55.8	54.9	64.4	62.2	59.0	61.9	59.3	36.7

Table 11: LiT benchmark by linguistic category under raw Judge Score. We report mean judge scores (higher is better), aggregated within each category over the same eight translation sequences. Each cell shows the mean and the bootstrap-estimated standard error, computed via sentence-level resampling after averaging across sequences. Informal text is the hardest category overall (63.3 avg), while STEM is the easiest (71.3).

Model	Average	Abstracts		Pragmatics					Informal
		Humanities	STEM	Core Semantics	Discourse	Implicit	Inference	Social	
Gemini-3-Flash (No-Thinking)	89.1 ± 0.2	91.2 ± 0.6	90.5 ± 0.7	89.6 ± 0.8	88.9 ± 0.9	86.2 ± 0.9	88.5 ± 0.6	88.2 ± 0.5	89.5 ± 0.5
Qwen3.5-397B (Thinking)	83.9 ± 0.4	85.2 ± 1.2	91.1 ± 0.7	86.8 ± 0.8	84.9 ± 0.9	81.9 ± 1.8	82.5 ± 0.9	83.1 ± 0.7	75.7 ± 1.5
Gemma-4-31B (Instruct)	80.9 ± 0.3	83.1 ± 1.0	87.6 ± 1.0	79.9 ± 0.9	80.7 ± 1.1	76.3 ± 1.2	80.1 ± 0.7	79.4 ± 0.6	80.1 ± 0.6
GLM-5 (Thinking)	80.7 ± 0.4	82.7 ± 1.0	86.4 ± 1.0	82.9 ± 1.1	82.6 ± 0.8	77.8 ± 1.7	80.0 ± 0.8	79.5 ± 0.9	73.4 ± 1.5
Qwen3.5-397B (Instruct)	80.1 ± 0.4	83.6 ± 1.6	85.0 ± 1.7	80.4 ± 1.5	78.9 ± 1.1	77.2 ± 1.6	79.7 ± 0.9	78.6 ± 0.7	77.4 ± 0.8
Gemma-4-31B (Thinking)	79.7 ± 0.4	82.2 ± 0.8	84.0 ± 1.4	79.6 ± 1.2	79.3 ± 1.2	76.0 ± 1.4	79.3 ± 0.9	78.6 ± 0.8	79.0 ± 0.7
GLM-4.7 (Thinking)	75.2 ± 0.4	77.7 ± 1.1	81.3 ± 1.2	77.7 ± 1.1	77.7 ± 0.7	72.5 ± 1.7	74.4 ± 0.7	73.5 ± 0.9	67.0 ± 1.3
DeepSeek-V3.2-Exp (Thinking)	73.7 ± 0.5	77.2 ± 1.7	72.1 ± 2.2	76.3 ± 1.1	75.2 ± 1.2	71.7 ± 1.5	73.6 ± 0.9	73.4 ± 0.9	70.0 ± 1.6
Kimi-K2 (Thinking)	72.3 ± 0.5	72.7 ± 1.4	75.3 ± 2.0	74.1 ± 0.8	74.1 ± 0.9	69.8 ± 1.9	72.7 ± 0.9	72.1 ± 1.0	67.9 ± 1.6
GLM-5 (Instruct)	71.9 ± 0.5	72.8 ± 1.7	74.4 ± 2.2	73.6 ± 0.9	72.5 ± 1.1	70.1 ± 1.5	71.4 ± 0.8	70.7 ± 0.9	69.6 ± 0.8
DeepSeek-V3.2-Exp	71.5 ± 0.5	75.2 ± 1.6	68.7 ± 2.4	72.1 ± 1.3	72.1 ± 1.2	70.2 ± 1.3	72.3 ± 0.9	70.0 ± 0.9	71.3 ± 0.7
Qwen3.5-35B (Thinking)	69.8 ± 0.5	69.4 ± 1.4	78.7 ± 1.4	74.1 ± 1.3	71.9 ± 1.4	67.8 ± 2.1	71.3 ± 0.8	67.4 ± 1.0	57.9 ± 1.2
Qwen3-235B (Thinking)	67.3 ± 0.4	68.6 ± 1.1	71.9 ± 1.4	69.8 ± 0.9	70.2 ± 0.9	65.2 ± 2.0	67.6 ± 0.6	65.6 ± 0.9	59.6 ± 1.1
Gemma-3-27B (Instruct)	67.1 ± 0.5	69.5 ± 1.2	60.2 ± 2.5	70.0 ± 1.0	70.1 ± 1.0	64.5 ± 1.6	69.6 ± 0.8	66.1 ± 0.9	67.2 ± 0.8
GPT-OSS-120B (High)	65.4 ± 0.5	68.0 ± 1.4	71.9 ± 1.7	69.7 ± 0.8	70.0 ± 1.4	64.4 ± 2.0	67.5 ± 0.9	64.2 ± 1.1	47.7 ± 1.8
MiniMax-M2.5	63.6 ± 0.5	62.1 ± 1.6	70.4 ± 1.6	68.9 ± 1.1	68.8 ± 1.2	63.2 ± 2.1	64.3 ± 0.9	61.2 ± 1.1	50.1 ± 1.2
Qwen3.5-35B (Instruct)	63.0 ± 0.5	62.6 ± 1.4	69.8 ± 1.9	65.1 ± 1.0	63.2 ± 1.2	60.7 ± 1.9	65.1 ± 0.9	60.4 ± 1.1	57.4 ± 1.1
Kimi-K2	62.5 ± 0.6	64.3 ± 1.6	60.5 ± 3.0	65.3 ± 1.2	63.4 ± 1.3	59.8 ± 1.9	63.1 ± 0.8	62.2 ± 1.0	61.6 ± 1.0
Qwen3-235B (Instruct)	56.6 ± 0.6	54.6 ± 2.4	57.1 ± 2.4	61.0 ± 1.4	57.1 ± 1.8	57.3 ± 1.8	56.7 ± 1.4	56.0 ± 1.6	53.3 ± 1.6
MiMo-V2-Flash	56.2 ± 0.5	55.5 ± 1.4	49.1 ± 2.6	60.6 ± 0.8	60.3 ± 1.2	56.2 ± 1.7	58.2 ± 1.1	56.1 ± 1.1	53.8 ± 0.9
Qwen3-30B (Instruct)	47.9 ± 0.5	47.0 ± 1.2	47.8 ± 2.1	50.4 ± 1.2	51.4 ± 1.2	47.1 ± 1.5	49.2 ± 0.6	48.6 ± 1.1	41.3 ± 0.9
Nemotron-3-Nano	29.2 ± 0.5	30.1 ± 1.4	35.6 ± 2.0	31.4 ± 1.2	31.0 ± 1.0	27.0 ± 1.4	28.9 ± 1.3	28.9 ± 1.3	21.0 ± 0.9
Average	68.5	69.8	71.3	70.9	70.2	66.5	68.9	67.4	63.3

Table 12: **Raw MQM scores similarly collapse on low-resource languages.** We report the raw MQM scores across eight language sequences grouped by resource availability. (a) High-resource sequences (East Asian, Central European, Near Eastern): top frontier models remain strong, with Gemini-3-Flash leading at 91.0 average and GLM-5 (Thinking) taking the East Asian column at 90.4. (b) Medium-resource sequences (Southeast Asian, South Asian, North European): performance remains stable for the strongest models, with Gemini-3-Flash averaging 89.9. (c) Low-resource sequences (African, South American): catastrophic collapse remains, with only Gemini-3-Flash (75.4) maintaining clearly usable performance; the next-best model (Qwen3.5-397B (Thinking)) drops to 43.7. (d) Global averages across these eight sequences show Gemini-3-Flash holding roughly a 10.5-point lead over the next competitor.

(a) High & Med-High Resource Sequences

Model	E. Asia	C. Europe	N. East	Avg
Gemini-3-Flash (No-Think)	90.0	**91.6**	**91.6**	**91.0**
Qwen3.5-397B (Thinking)	89.3	90.3	87.1	88.9
GLM-5 (Thinking)	**90.4**	89.7	85.4	88.5
Gemma-4-31B (Thinking)	89.6	88.0	86.8	88.1
Gemma-4-31B (Instruct)	88.6	88.2	86.0	87.6
Qwen3.5-397B (Instruct)	88.1	89.3	85.2	87.5
GLM-4.7 (Thinking)	89.3	89.0	82.9	87.1
Kimi-K2 (Thinking)	86.1	87.2	82.8	85.4
DeepSeek-V3.2-Exp	87.6	86.2	79.9	84.6
Qwen3-235B (Thinking)	85.0	87.3	80.0	84.1
GLM-5 (Instruct)	86.3	85.9	77.8	83.3
DeepSeek-V3.2-Exp (Think)	85.8	82.9	79.7	82.8
Qwen3.5-35B (Thinking)	81.2	85.3	75.7	80.8
Gemma-3-27B (Instruct)	78.6	83.1	77.8	79.8
Kimi-K2	79.4	82.8	70.5	77.6
Qwen3.5-35B (Instruct)	79.3	80.7	64.1	74.7
Qwen3-235B (Instruct)	83.7	68.3	72.0	74.7
MiniMax-M2.5	71.6	80.2	68.7	73.5
GPT-OSS-120B (High)	78.3	76.8	63.9	73.0
MiMo-V2-Flash	70.2	74.5	54.3	66.4
Qwen3-30B (Instruct)	71.3	67.5	44.4	61.0
Nemotron-3-Nano	50.9	11.5	-17.1	15.1

(b) Medium Resource Sequences

Model	SE. Asia	S. Asia	N. Europe	Avg
Gemini-3-Flash (No-Think)	**89.3**	**90.3**	**90.2**	**89.9**
Qwen3.5-397B (Thinking)	86.0	83.1	86.3	85.1
GLM-5 (Thinking)	86.3	83.3	83.7	84.5
Gemma-4-31B (Instruct)	88.6	88.2	76.0	84.2
Gemma-4-31B (Thinking)	86.5	89.0	74.7	83.4
Qwen3.5-397B (Instruct)	85.8	79.7	81.9	82.5
GLM-4.7 (Thinking)	80.9	78.1	74.6	77.9
Kimi-K2 (Thinking)	79.6	73.2	77.6	76.8
GLM-5 (Instruct)	75.3	75.5	68.0	72.9
DeepSeek-V3.2-Exp	81.4	66.4	70.8	72.9
Qwen3-235B (Thinking)	81.1	71.5	63.5	72.0
DeepSeek-V3.2-Exp (Think)	79.9	65.2	68.8	71.3
Gemma-3-27B (Instruct)	76.5	66.9	66.8	70.1
Qwen3.5-35B (Thinking)	74.4	62.7	66.6	67.9
Kimi-K2	72.7	58.3	67.0	66.0
Qwen3-235B (Instruct)	77.2	65.6	43.7	62.2
GPT-OSS-120B (High)	67.7	58.0	59.4	61.7
Qwen3.5-35B (Instruct)	68.0	46.6	53.6	56.1
MiniMax-M2.5	61.6	34.8	57.9	51.4
MiMo-V2-Flash	62.3	38.7	40.1	47.0
Qwen3-30B (Instruct)	53.3	18.1	1.7	24.4
Nemotron-3-Nano	-8.1	-41.8	-46.0	-31.9

(c) Low Resource / Imbalanced Sequences

Model	Africa	S. America	Avg
Gemini-3-Flash (No-Think)	**84.4**	**66.4**	**75.4**
Qwen3.5-397B (Thinking)	58.5	28.9	43.7
Gemma-4-31B (Instruct)	74.8	3.5	39.2
Gemma-4-31B (Thinking)	72.6	-6.2	33.2
Qwen3.5-397B (Instruct)	45.9	10.7	28.3
GLM-5 (Thinking)	49.1	3.9	26.5
DeepSeek-V3.2-Exp (Think)	29.6	0.4	15.0
Gemma-3-27B (Instruct)	16.8	1.0	8.9
Kimi-K2 (Thinking)	11.7	-17.5	-2.9
Qwen3.5-35B (Thinking)	13.8	-19.7	-3.0
GLM-4.7 (Thinking)	18.4	-24.7	-3.2
DeepSeek-V3.2-Exp	13.6	-21.0	-3.7
GPT-OSS-120B (High)	3.4	-13.2	-4.9
Nemotron-3-Nano	-12.8	1.5	-5.6
Qwen3.5-35B (Instruct)	-11.4	-3.2	-7.3
Kimi-K2	-16.2	-1.0	-8.6
GLM-5 (Instruct)	9.2	-27.6	-9.2
Qwen3-30B (Instruct)	-12.7	-8.9	-10.8
Qwen3-235B (Instruct)	-17.3	-5.9	-11.6
MiniMax-M2.5	-15.1	-8.4	-11.8
Qwen3-235B (Thinking)	-35.8	-8.4	-22.1
MiMo-V2-Flash	-34.6	-11.1	-22.9

(d) Overall Performance

Model	Global Average
Gemini-3-Flash (No-Think)	**86.7**
Qwen3.5-397B (Thinking)	76.2
Gemma-4-31B (Instruct)	74.2
Gemma-4-31B (Thinking)	72.6
GLM-5 (Thinking)	71.5
Qwen3.5-397B (Instruct)	70.8
DeepSeek-V3.2-Exp (Think)	61.5
GLM-4.7 (Thinking)	61.1
Kimi-K2 (Thinking)	60.1
Gemma-3-27B (Instruct)	58.4
DeepSeek-V3.2-Exp	58.1
GLM-5 (Instruct)	56.3
Qwen3.5-35B (Thinking)	55.0
Qwen3-235B (Thinking)	53.0
Kimi-K2	51.7
GPT-OSS-120B (High)	49.3
Qwen3-235B (Instruct)	48.4
Qwen3.5-35B (Instruct)	47.2
MiniMax-M2.5	43.9
MiMo-V2-Flash	36.8
Qwen3-30B (Instruct)	29.3
Nemotron-3-Nano	-7.7

Table 13: **Judge scores also drop on low-resource languages.** We report judge scores across the same eight language sequences on the 200-example subset, grouped by resource availability. (a) High-resource sequences remain strong for frontier models, with Gemini-3-Flash averaging 93.0 and taking both Central Europe (94.2) and Near East (93.5), while GLM-5 (Thinking) narrowly leads East Asia (91.5). (b) Medium-resource sequences remain comparatively stable for the strongest models, with Gemini-3-Flash averaging 91.6. (c) Low-resource sequences show a substantial drop, with Gemini-3-Flash still clearly first at 79.4 average, while the next-best model, Qwen3.5-397B (Thinking), falls to 64.4. (d) Global averages across these eight sequences preserve the same ranking pattern, with Gemini-3-Flash leading at 89.1 and holding a 6.0-point lead over the next competitor.

(a) High & Med-High Resource Sequences

Model	E. Asia	C. Europe	N. East	Avg
Gemini-3-Flash (No-Think)	91.2	**94.2**	**93.5**	**93.0**
Gemma-4-31B (Thinking)	91.1	91.3	90.3	90.9
Qwen3.5-397B (Thinking)	91.0	91.5	90.0	90.8
Qwen3.5-397B (Instruct)	89.7	91.9	90.1	90.6
GLM-5 (Thinking)	**91.5**	91.3	88.7	90.5
Gemma-4-31B (Instruct)	90.5	90.9	89.9	90.4
GLM-4.7 (Thinking)	90.2	90.3	87.3	89.3
Kimi-K2 (Thinking)	88.0	88.6	87.8	88.1
DeepSeek-V3.2-Exp	88.6	89.0	85.3	87.6
GLM-5 (Instruct)	88.6	89.6	83.9	87.3
Qwen3-235B (Thinking)	87.4	89.3	85.3	87.3
Qwen3.5-35B (Thinking)	83.8	88.3	82.8	85.0
DeepSeek-V3.2-Exp (Think)	87.5	83.0	84.1	84.8
Gemma-3-27B (Instruct)	82.7	87.2	84.4	84.8
Kimi-K2	82.4	86.9	78.9	82.8
Qwen3.5-35B (Instruct)	83.6	85.6	77.0	82.1
MiniMax-M2.5	77.4	84.5	79.0	80.3
GPT-OSS-120B (High)	81.3	79.5	76.1	79.0
Qwen3-235B (Instruct)	86.3	70.2	78.6	78.4
MiMo-V2-Flash	77.6	81.7	70.1	76.5
Qwen3-30B (Instruct)	79.1	76.4	65.2	73.6
Nemotron-3-Nano	66.4	48.5	29.9	48.3

(b) Medium Resource Sequences

Model	SE. Asia	S. Asia	N. Europe	Avg
Gemini-3-Flash (No-Think)	**90.3**	**91.7**	**92.8**	**91.6**
Qwen3.5-397B (Thinking)	86.9	86.7	89.8	87.8
Gemma-4-31B (Instruct)	89.9	89.4	82.7	87.3
GLM-5 (Thinking)	86.7	84.7	88.2	86.5
Qwen3.5-397B (Instruct)	88.1	84.2	86.9	86.4
Gemma-4-31B (Thinking)	86.0	90.2	81.7	86.0
GLM-4.7 (Thinking)	84.4	83.0	83.7	83.7
Kimi-K2 (Thinking)	82.7	78.6	84.3	81.9
Qwen3-235B (Thinking)	83.3	80.3	78.5	80.7
DeepSeek-V3.2-Exp	84.3	77.3	80.1	80.6
GLM-5 (Instruct)	80.4	80.3	78.4	79.7
DeepSeek-V3.2-Exp (Think)	81.4	74.1	79.4	78.3
Gemma-3-27B (Instruct)	81.1	76.6	77.2	78.3
Qwen3.5-35B (Thinking)	78.5	74.3	79.6	77.5
Kimi-K2	77.3	71.6	78.0	75.6
GPT-OSS-120B (High)	71.8	72.8	72.8	72.5
Qwen3.5-35B (Instruct)	76.2	65.8	71.6	71.2
Qwen3-235B (Instruct)	79.1	73.6	53.8	68.8
MiniMax-M2.5	72.9	58.0	74.2	68.4
MiMo-V2-Flash	72.6	63.7	65.8	67.4
Qwen3-30B (Instruct)	66.5	50.0	41.3	52.6
Nemotron-3-Nano	33.2	15.9	17.0	22.0

(c) Low Resource / Imbalanced Sequences

Model	Africa	S. America	Avg
Gemini-3-Flash (No-Think)	**86.2**	**72.6**	**79.4**
Qwen3.5-397B (Thinking)	72.4	56.3	64.4
Gemma-4-31B (Instruct)	79.4	32.7	56.1
GLM-5 (Thinking)	66.4	41.9	54.1
Qwen3.5-397B (Instruct)	66.2	41.0	53.6
Gemma-4-31B (Thinking)	74.2	31.9	53.1
DeepSeek-V3.2-Exp (Think)	57.7	39.7	48.7
GLM-4.7 (Thinking)	52.1	23.9	38.0
GLM-5 (Instruct)	48.4	23.6	36.0
DeepSeek-V3.2-Exp	46.6	20.4	33.5
Kimi-K2 (Thinking)	46.8	18.5	32.6
Qwen3.5-35B (Thinking)	47.1	14.8	31.0
GPT-OSS-120B (High)	36.7	18.8	27.8
MiniMax-M2.5	33.4	19.4	26.4
Gemma-3-27B (Instruct)	46.3	2.6	24.5
Qwen3.5-35B (Instruct)	33.2	6.8	20.0
Qwen3-235B (Thinking)	16.3	12.0	14.2
Kimi-K2	19.8	5.0	12.4
MiMo-V2-Flash	7.9	9.4	8.7
Nemotron-3-Nano	5.7	10.8	8.3
Qwen3-235B (Instruct)	7.2	2.2	4.7
Qwen3-30B (Instruct)	0.0	0.1	0.1

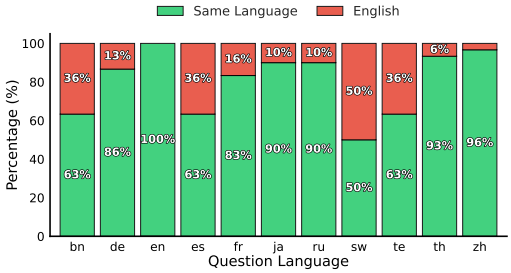
(d) Overall Performance

Model	Global Average
Gemini-3-Flash (No-Think)	**89.1**
Qwen3.5-397B (Thinking)	83.1
Gemma-4-31B (Instruct)	80.7
GLM-5 (Thinking)	79.9
Qwen3.5-397B (Instruct)	79.8
Gemma-4-31B (Thinking)	79.6
GLM-4.7 (Thinking)	74.3
DeepSeek-V3.2-Exp (Think)	73.4
Kimi-K2 (Thinking)	71.9
GLM-5 (Instruct)	71.6
DeepSeek-V3.2-Exp	71.4
Qwen3.5-35B (Thinking)	68.7
Gemma-3-27B (Instruct)	67.3
Qwen3-235B (Thinking)	66.5
GPT-OSS-120B (High)	63.7
Kimi-K2	62.5
Qwen3.5-35B (Instruct)	62.5
MiniMax-M2.5	62.3
Qwen3-235B (Instruct)	56.4
MiMo-V2-Flash	56.1
Qwen3-30B (Instruct)	47.3
Nemotron-3-Nano	28.4

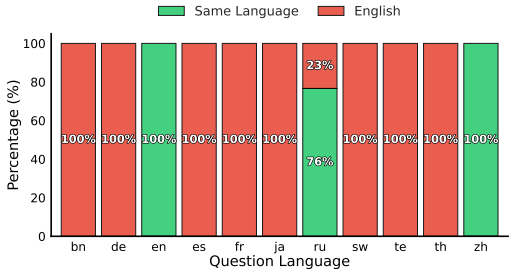
E ROBUSTNESS OF ANSWER LANGUAGE ANALYSIS ACROSS MODELS

In this section, we provide results on the languages in which models answer and reason, extending the analysis in Section 3.2 across models. We show results for MT-AIME24 in Figures 4 and 5 and Figures 6 and 7 for INCLUDE to separate two failure modes that can confound multilingual benchmark evaluation. First, some models do not consistently answer in the language of the prompt. For example, Qwen3-32B answers in Swahili only about half of the time when prompted in Swahili.

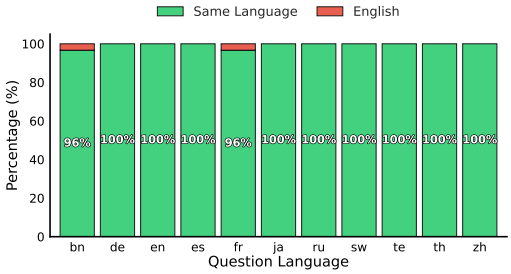
Second, an even stronger effect appears in the reasoning traces: most models reason predominantly in English, with only a few partial exceptions, such as Chinese and Russian for some Qwen3 models. Taken together, these results further support our claim that multilingual reasoning and general-knowledge benchmarks often measure English-centered reasoning ability more than genuine multilingual capability.



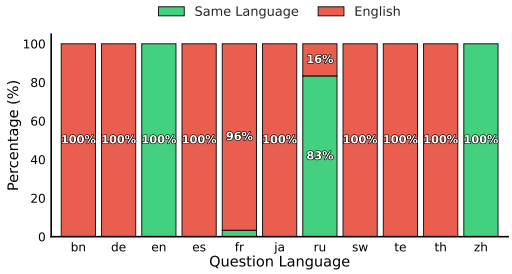
(a) **Answer language**: the language the model Qwen3-32B answers in.



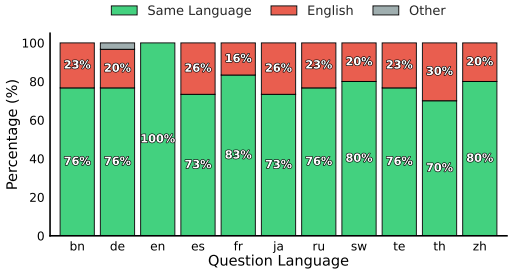
(b) **Reasoning language**: the language the model Qwen3-32B reasons in.



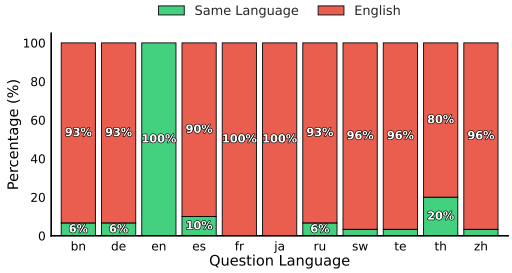
(c) **Answer language**: the language the model Qwen3-235B-A22B-Thinking-2507 answers in.



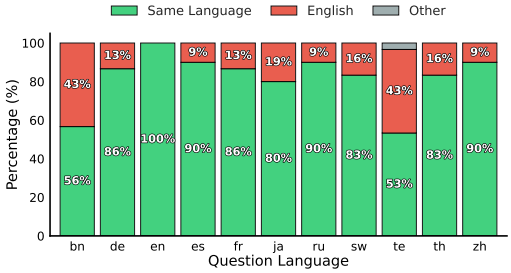
(d) **Reasoning language**: the language the model Qwen3-235B-A22B-Thinking-2507 reasons in.



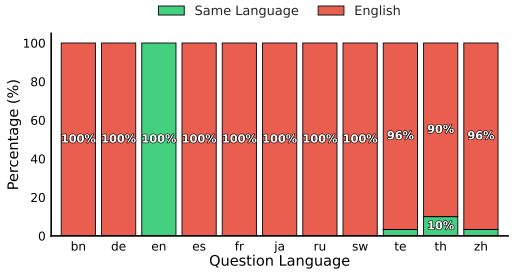
(e) **Answer language**: the language the model GPT-OSS-20B answers in.



(f) **Reasoning language**: the language the model GPT-OSS-20B reasons in.

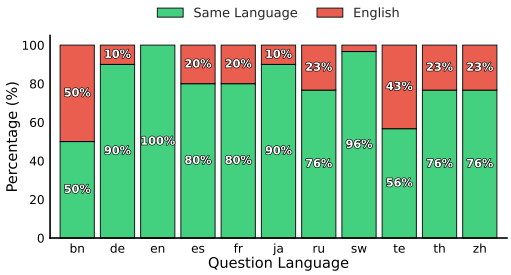


(g) **Answer language**: the language the model GPT-OSS-120B answers in.

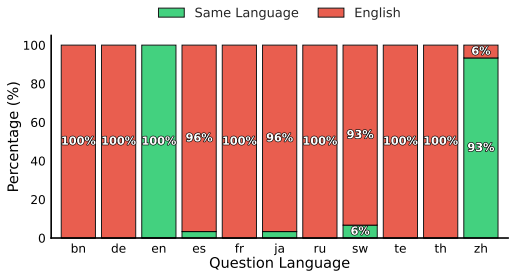


(h) **Reasoning language**: the language the model GPT-OSS-120B reasons in.

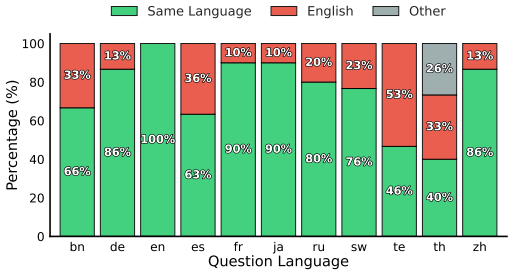
Figure 4: Qwen3 and GPT-OSS models default to English reasoning on MT-AIME24. We analyze the language used for answers (left column) and reasoning traces (right column) across 11 languages. (a-b) Qwen3-32B answers in the target language inconsistently but reasons almost entirely in English. (c-d) Qwen3-235B-Thinking answers consistently in the target language (100%) but still reasons predominantly in English, especially for Swahili. (e-f) GPT-OSS-20B shows mixed answering behavior but reasons 93–100% in English. (g-h) GPT-OSS-120B answers mostly in the target language but reasons almost entirely in English. These patterns confirm that mathematical reasoning occurs in English regardless of input language.



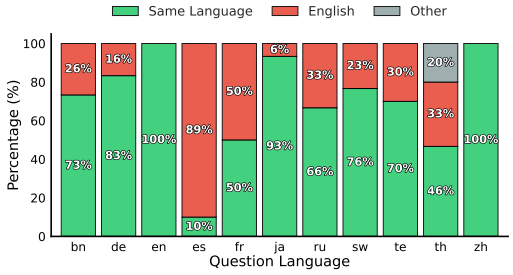
(a) **Answer language**: the language the model GLM-4.7 Z.ai et al. (2025) answers in.



(b) **Reasoning language**: the language the model GLM-4.7 Z.ai et al. (2025) reasons in.

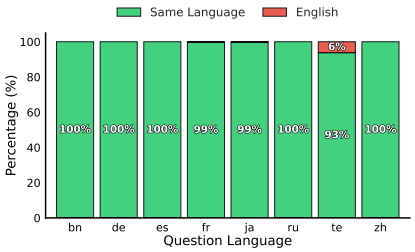


(c) **Answer language**: the language the model MiMo-V2-Flash answers in.

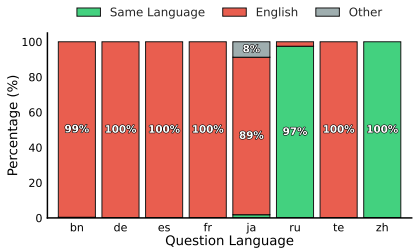


(d) **Reasoning language**: the language the model MiMo-V2-Flash reasons in.

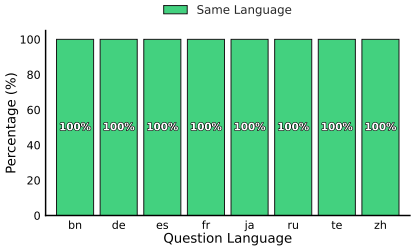
Figure 5: GLM-4.7 and MiMo-V2-Flash show contrasting reasoning language patterns on MT-AIME24. (a-b) GLM-4.7 answers in mixed languages across different inputs but reasons overwhelmingly in English (93–100% for most languages), with slight exceptions for Swahili and Chinese. (c-d) MiMo-V2-Flash displays the most diverse reasoning behavior: it reasons natively 33–100% of the time depending on the language, making it an outlier among tested models. However, this native reasoning does not translate to better benchmark performance, further suggesting that MT-AIME24 measures reasoning ability rather than multilingual proficiency.



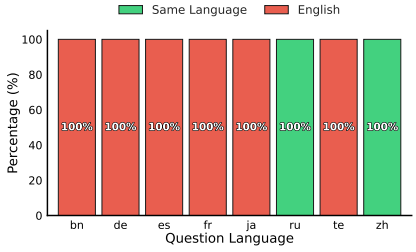
(a) **Answer language**: the language the model Qwen3-32B answers in.



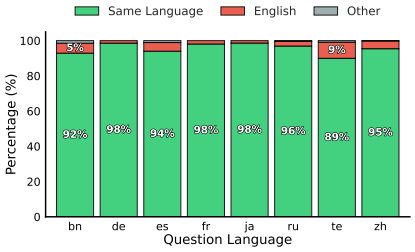
(b) **Reasoning language**: the language the model Qwen3-32B reasons in.



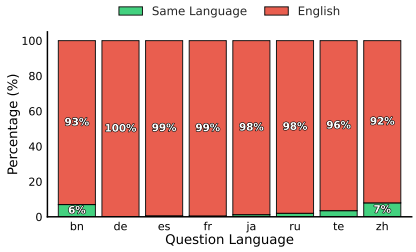
(c) **Answer language**: the language the model Qwen3-235B-A22B-Thinking-2507 answers in.



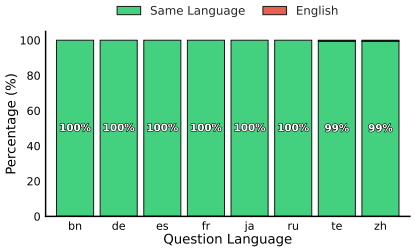
(d) **Reasoning language**: the language the model Qwen3-235B-A22B-Thinking-2507 reasons in.



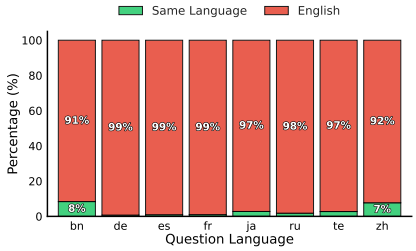
(e) **Answer language**: the language the model GPT-OSS-20B answers in.



(f) **Reasoning language**: the language the model GPT-OSS-20B reasons in.

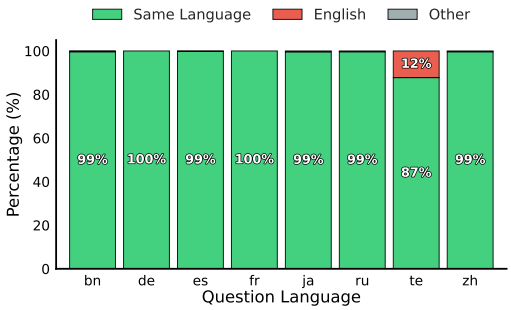


(g) **Answer language**: the language the model GPT-OSS-120B answers in.

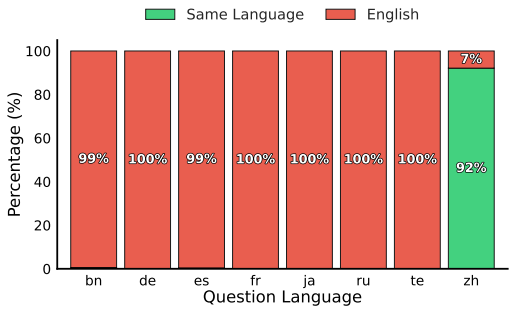


(h) **Reasoning language**: the language the model GPT-OSS-120B reasons in.

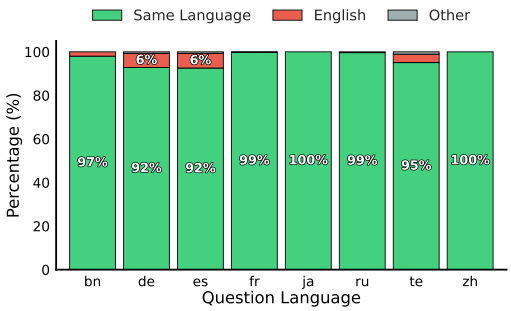
Figure 6: Qwen3 and GPT-OSS models predominantly reason in English on INCLUDE despite answering in target languages. (a-b) Qwen3-32B generally answers in the target language but reasons predominantly in English, except for Russian and Chinese. (c-d) Qwen3-235B-Thinking answers in the target language but reasons in English except for Russian and Chinese. (e-f) GPT-OSS-20B answers mostly in the target language (89–98%) but reasons in English 92–100% of the time. (g-h) GPT-OSS-120B similarly answers in the target language but reasons predominantly in English. The disconnect between answer language and reasoning language explains why INCLUDE performance tracks English knowledge benchmarks.



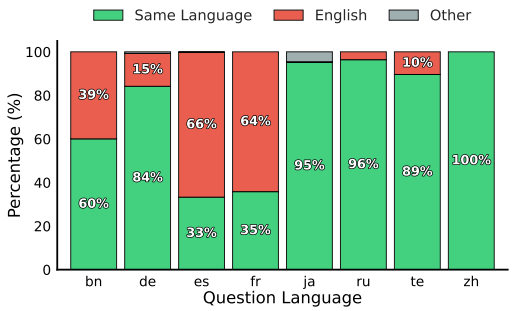
(a) **Answer language**: the language the model GLM-4.7 Z.ai et al. (2025) answers in.



(b) **Reasoning language**: the language the model GLM-4.7 Z.ai et al. (2025) reasons in.



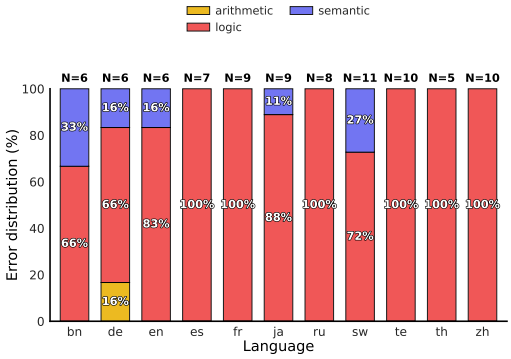
(c) **Answer language**: the language the model MiMo-V2-Flash answers in.



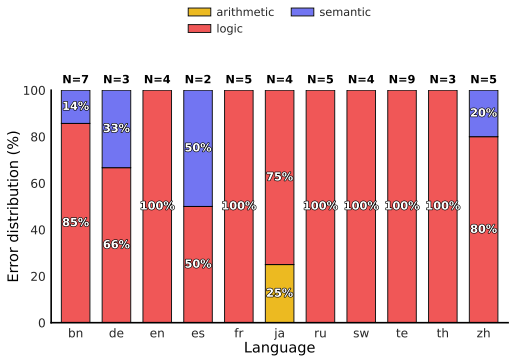
(d) **Reasoning language**: the language the model MiMo-V2-Flash reasons in.

Figure 7: GLM-4.7 reasons in English while MiMo-V2-Flash shows mixed reasoning patterns on INCLUDE. (a-b) GLM-4.7 answers in the target language 87–100% of the time and reasons in English 92–100% of the time across languages. Russian (87%) and Telugu (12%) show the most answer-language variation. (c-d) MiMo-V2-Flash displays highly variable reasoning behavior: it reasons natively 33–100% of the time depending on the language, with particularly high native reasoning for Telugu (100%), Bengali (60%), and Spanish (84%). This variability makes MiMo-V2-Flash an interesting case study for understanding how reasoning language affects downstream task performance.

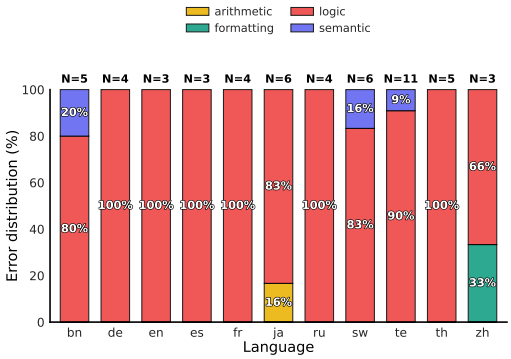
F Robustness of Error Distribution Analysis Across Models



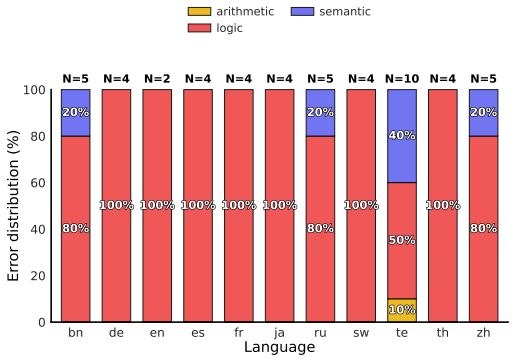
(a) Error distribution for Qwen3-32B



(b) Error distribution for Qwen3-235B-A22B-Thinking-2507



(c) Error distribution for GPT-OSS-120B

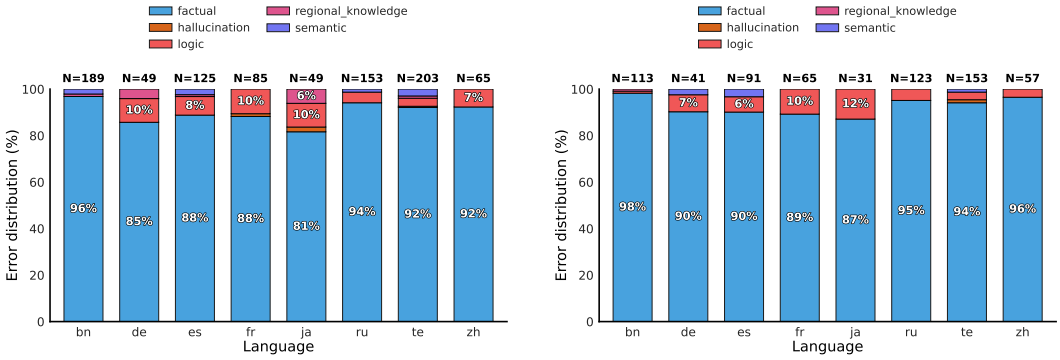


(d) Error distribution for GLM-4.7

Figure 8: Error analysis across four additional models confirms MT-AIME24 errors are logical, not linguistic. We categorize errors for four models on MT-AIME24 across 11 languages. (a-d) Across all models, errors are predominantly logical (blue) or arithmetic (green) rather than semantic (red). The semantic error rate rarely exceeds 25% for any language-model combination. For several languages (e.g., French in panels a and c), 100% of errors stem from reasoning failures. This consistent pattern across diverse model architectures confirms that MT-AIME24 does not effectively measure multilingual comprehension.

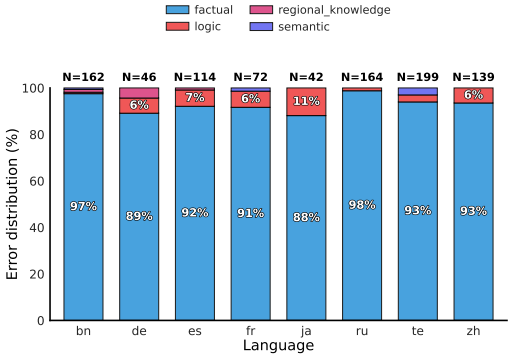
We provide additional error distribution analyses for MT-AIME24 in Figure 8 and INCLUDE in Figure 9 across additional model families. We observe that MT-AIME24 errors are consistently dominated by logical and arithmetic failures rather than semantic misunderstanding, indicating that the benchmark primarily tests mathematical reasoning rather than multilingual comprehension.

Figure 9 shows a parallel pattern for INCLUDE: errors are overwhelmingly factual, with smaller contributions from regional knowledge gaps and hallucinations, while semantic errors due to multilingual misunderstanding remain rare. Together, these results strengthen our conclusion that current multilingual benchmarks mainly reflect reasoning and factual recall, not genuine cross-lingual understanding.

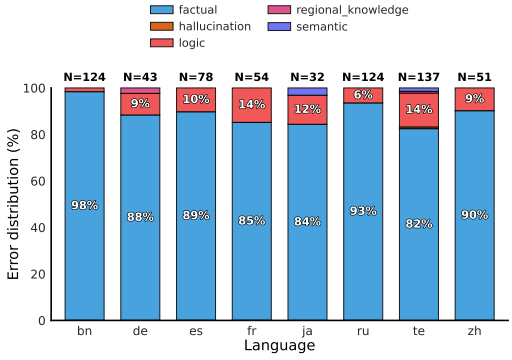


(a) Qwen3-32B

(b) Qwen3-235B-A22B-Thinking



(c) GPT-OSS-120B



(d) GLM-4.7

Figure 9: Error analysis on INCLUDE confirms errors are factual and knowledge-based, not linguistic. We categorize errors for four models across eight languages on INCLUDE. (a-d) Across all models, the dominant error type is factual (blue), accounting for 81–98% of errors depending on language and model. Semantic errors (red) rarely exceed 10% for any configuration. Regional knowledge gaps (orange) contribute modestly (6–14% in some cases). Hallucinations appear occasionally but are not the primary failure mode. This consistent pattern confirms that INCLUDE measures factual knowledge coverage rather than multilingual comprehension ability.

G Additional Experiments and Ablations

We provide evaluation details and supporting figures for the main-text validation experiments, followed by extended benchmark comparisons and qualitative examples.

G.1 Model Configurations for Inter-Judge Agreement

The inter-judge agreement analysis in Section 4.2 (Table 4) uses ten model configurations: DeepSeek-V3.2-Exp, Gemma-4-31B, Qwen3.5-35B, and Qwen3.5-397B (each with and without thinking), plus Gemini-3-Flash (No-Thinking) and GPT-OSS-120B (High).

G.2 Translation-chain length

To evaluate the performance across our benchmark sequences, we vary the number of translation hops for six diversely chosen models shown in Figure 10. We use the same benchmark sequences and ordering from Appendix B.2 for Central Europe, East Asia, Near East, and Southeast Asia, returning to English after the first one, two, three, or all four intermediate languages. We use DeepSeek-V4-Pro as the judge.

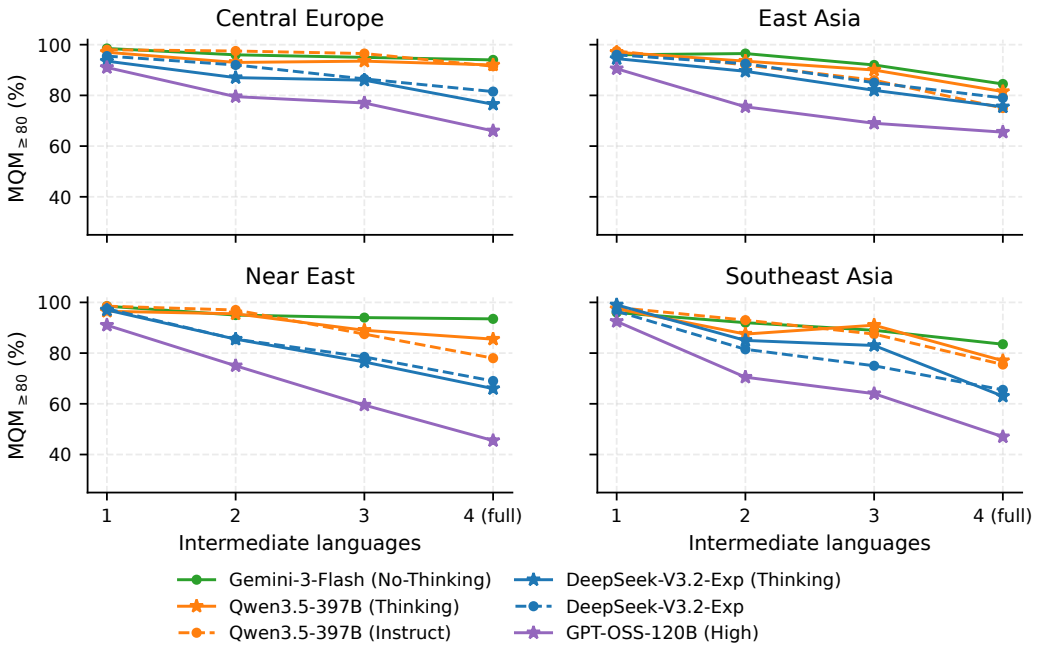


Figure 10: **Effect of translation-chain length.** $MQM_{\geq 80}$ pass rates for six models from Table 2. Full denotes all four intermediate languages of the sequences in Appendix B.2, followed by the translation back to English.

Longer chains generally lower pass rates while retaining strong agreement with two- and three-language chains (Table 5). One-language round trips are near ceiling and have less consistent rank agreement. This supports longer chains as a more demanding test of meaning preservation.

G.3 Extended benchmark comparisons

Figure 11 extends Figure 1 to two additional multilingual benchmarks with public scores, Global-MMLU-Lite (Singh et al., 2025) from Artificial Analysis[5] and MGSM (Shi et al., 2023) from Vals.ai,[6] using all models in the LMArena range of Figure 1, widened by 10 Elo. MGSM scores are averaged over French, German, Japanese, Russian, and Spanish.

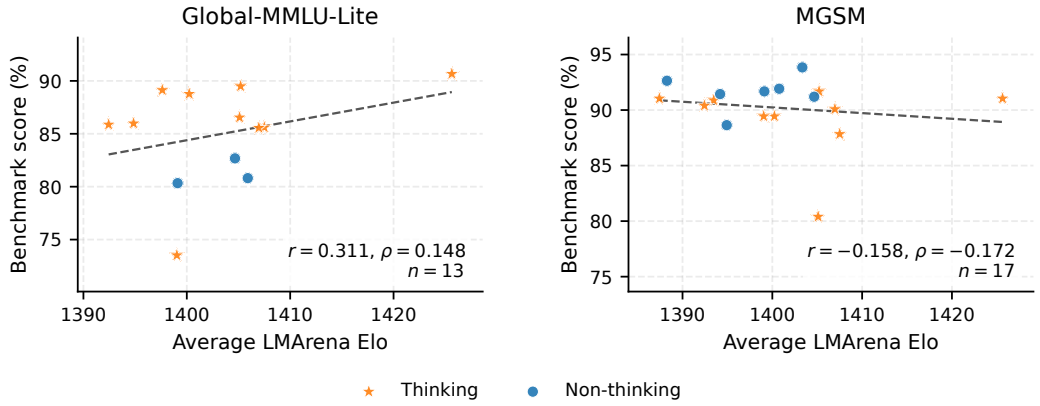


Figure 11: **Additional multilingual knowledge and reasoning benchmarks show no association with preferences among similarly rated models.** Global-MMLU-Lite (general knowledge, like INCLUDE) and MGSM (mathematical reasoning, like MT-AIME24) for models with a six-language LMArena average between 1384 and 1426 Elo; dashed lines are linear fits.

[5] https://artificialanalysis.ai/evaluations/global-mmlu-lite
[6] https://www.vals.ai/benchmarks/mgsm

G.4 QUALITATIVE EXAMPLES FOR GEMINI-3-FLASH

We examine Gemini-3-Flash (No-Thinking), which achieves the highest overall and informal-text scores on LiT (Table 2). The examples in Table 14 illustrate how LiT distinguishes models' ability to preserve jokes, nuance, and informal syntax during translation.

Table 14: **Qualitative examples for Gemini-3-Flash.** Gemini-3-Flash preserves jokes, nuance, and informal syntax in these examples, whereas other models introduce errors such as translating "vibe" as "air". Gemini denotes Gemini-3-Flash (No-Thinking).

What the original tests	Gemini behavior	Failure mode in other models
Product group chat *informal-4* Multi-speaker chat, slang, "shipping," "to the public/team/void," accountability joke	Preserves speaker turns, release meaning, jokes, urgency, and casual tone	Qwen3-235B (Thinking) scores 15/100: omits most of the chat and cuts off; DeepSeek-V3.2-Exp (Thinking) loses key jokes.
Relationship post *informal-3* Informal advice-post style, emotional nuance, concrete cancellation example	Keeps "am I overreacting," false agreement pattern, dinner cancellation, and later update	Qwen3.5-397B (Thinking) distorts it into "do you love me / crooked path"; GPT-OSS-120B (High) collapses into unrelated nonsense
Brunch planning chat *informal-5* Internet humor and idioms: "two leaves and a vibe," "international diplomacy," QR-code menu joke	Keeps the brunch negotiation, social dynamics, and most jokes intact	Other models translate "vibe" as "air," "wind," or "feeling," losing the food/aesthetic joke
Anxiety meme post *informal-1* Meme phrasing: "built different (derogatory)," "microwave beep in my soul," "stop being perceived"	Preserves the self-deprecating voice and understands "stop being perceived" as wanting social invisibility	Other outputs literalize it as "stop having an identity" or shift "forest creature" into "feral creature"