

Multilingual Reasoning Cascades Need More Context

Arnav Mazumder^{1*} Dengjia Zhang^{2*}

Shuyue Stella Li¹ Yulia Tsvetkov¹ Niyati Bafna²

¹University of Washington ²Johns Hopkins University
am2107@cs.washington.edu dzhang98@jhu.edu

Abstract

Translation cascades for reasoning translate the query from another language to English, reason in English, and translate the answer back to the original language. This is a competitive approach to multilingual reasoning, but structurally lossy, since each stage discards information later stages may need, including cues for cultural grounding, register, and disambiguation. We examine the benefits of a simple and training-free intervention: a *context-aware* translation cascade, which additionally provides the original question, the English-translated question, and the reasoning trace to the final cascade stage for target-language generation. We evaluate gains across ten multilingual benchmarks including various task types, three backbone models, and 285 high-, mid-, and low-resource languages, and demonstrate strong gains for open-ended generation across models and resource regimes. We show that the original language question carries most of the beneficial context. Our study emphasizes the need to better design information flow in multilingual reasoning cascades for mitigating error propagation, and provides a simple and actionable default strategy: preserve the original user question until the end of the pipeline.¹

1 Introduction

Large language models reason more effectively in English than in most other languages (Shi et al., 2023). Translation cascades exploit this for multilingual reasoning by translating the target-language input into English (MT1), reasoning in English, then translating the answer back (MT2). This approach matches or outperforms end-to-end multilingual generation on a range of tasks (Liu et al., 2025; Intrator et al., 2024). However, cascaded systems notoriously suffer from error propagation (Zhu et al.,

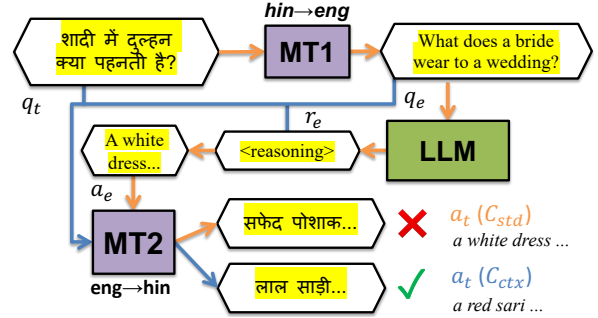


Figure 1: Standard cascade (C_{std} , orange) vs. context-aware cascade (C_{ctx} , orange+blue). In C_{std} , MT2 translates only the English answer a_e . In C_{ctx} , the final stage additionally receives the original target-language question q_t , the English question q_e , and the English reasoning trace r_e , and acts as a second reasoning stage for error recovery and translation

2019). We target a structural weakness: while MT2 only has access to the final English answer but not the original language and English question, accurate and grounded translation of the final English answer may require informational cues present only in the original question, depending on the syntax of the target language, cultural context, or simply to recover from errors in intermediate stages.

Contextual information could be critical in tasks requiring cultural context and grounding, given certain answer formats (Veselovsky et al., 2026), or for certain target languages that have poor translation support (Liu et al., 2019). However, in some settings, additional context can be damaging, e.g., distracting or confusing (Yang et al., 2025). In this work, we thus explore: *when does extra context benefit the multilingual reasoning cascade?*

We introduce a minimal and targeted intervention: a *context-aware MT cascade* (C_{ctx}). This differs from the *standard cascade* (C_{std}) in that C_{ctx} additionally feeds into MT2 the original input and intermediate outputs, as depicted in Figure 1. Although we retain the conventional label MT2, the context-aware final stage performs more than iso-

*Equal contribution.

¹We release the code for our experiments at <https://github.com/adoptedirelia/Multiling-reasoning>.

Source	Cascade state	Outputs (a_t)	Why context helps
Reasoning Error	q_t : Qual o ingrediente principal da chanfana? q_e : What is the main ingredient of chanfana? r_e : "Chanfana is a traditional dish from Extremadura in Spain... " a_e : The main ingredient is lamb .	C_{std} : O ingrediente principal da chanfana é o cordeiro (lamb) . C_{ctx} : O ingrediente principal da chanfana é o cabrito (goat) .	English reasoning misidentifies the dish as Spanish. Seeing the Portuguese question lets C_{ctx} recover the intended Portuguese dish.
MT2 Error	q_t : Apabila orang Melayu berbuka puasa di bulan Ramadan, makanan pertama yang dimakan biasanya ialah q_e : When Malays break their fast during Ramadan, the first food they usually eat is r_e : "... start their iftar with dates and water ..." a_e : Dates	C_{std} : Tarikh (calendar date) C_{ctx} : Kurma (date fruit)	The word "dates" is ambiguous. Without context, MT2 chooses the calendar sense; with the Malay question and reasoning, it chooses the food sense.
MT1 Error	q_t : 街边有个小摊在做糖人。糖人师傅用勺子舀起熬好的糖稀，在石板上快速勾勒。糖人与泥塑相比，更容易被碰碎的是 q_e : Which is more likely to be broken, candy figurines or clay figurines? r_e : "... clay figurines are more likely to be broken ..." a_e : Candy figurines are less likely to be broken.	C_{std} : 街头小贩制作的 糖果塑像比陶土塑像更容易保持完整 . (less likely) C_{ctx} : 糖人比泥塑更容易被碰碎 . (more likely)	The English translation abstracts the culture-specific item (food made of sugar) into generic "candy figurines." Seeing the original Chinese lets C_{ctx} recover this information and reason correctly.

Table 1: Standard cascade (C_{std}) vs. context-aware cascade (C_{ctx}) as described in Figure 1. Errors are shown in **red**; correct outputs are shown in **green**. C_{ctx} resolves target-language ambiguity and recovers from upstream translation or reasoning errors using the retained context.

lated translation: it generates the target-language answer while using upstream context for disambiguation and, when necessary, error recovery.

We analyze the usefulness of a context-aware cascade given the task type (open-ended QA, MCQ-style QA, and Math), and resourcedness of the target language. We find that the context-aware cascade shows strong gains for open-ended generation with cultural reasoning as opposed to classification style or exact-answer tasks, allowing small open-source models to considerably bridge the gap with strong proprietary models in these settings. The benefit extends to languages of varying resourcedness. Finally, our ablation study shows that the original user question is the most beneficial context to propagate. Our study provides a simple recommendation for translation cascades with reasoning: in settings requiring grounding and nuance, preserve the user’s question until the end of the pipeline.

2 Method

Let q_t be the original question in language t . A standard cascade system (1) translates q_t to English to get q_e (MT1); (2) reasons in English, producing reasoning trace r_e ; (3) answers in English with a_e ; and (4) translates the answer to the original language to produce a_t (MT2). Our methodology provides a controlled setup to compare the **standard cascade** (C_{std}) and the **context-aware cascade** (C_{ctx}), which differ only in the information available in the prompt context of the MT2 module, as shown in Figure 1. Specifically, in C_{std} , MT2 receives only a_e , whereas in C_{ctx} , MT2 additionally receives q_t ,

q_e , and r_e . For multiple-choice and math tasks, where a_e is already language-neutral, C_{std} requires no additional MT2 realization step; C_{ctx} nevertheless applies the context-aware final stage to permit error recovery.

Note that as a consequence, the two systems differ in the role played by the final stage. In C_{std} , MT2 performs answer translation from a_e alone. In C_{ctx} , it functions as a second reasoning step: translation remains the primary goal when the upstream answer is correct, while the model may revise the answer when the provided context reveals an upstream translation or reasoning error.

3 Experiments

Models. We use three backbone models: Llama-3.1-8B-Instruct (Llama Team, AI @ Meta, 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and GPT-4o-mini (OpenAI, 2024). All components use greedy decoding ($T=0$). See Appendix A for prompts. We use the same model checkpoint for every role (MT1, reasoner, MT2), avoiding confounds from mixed-model pipelines.

Datasets. We evaluate on ten multilingual benchmarks grouped by output regime:

- **Open-ended QA:** We use Aya Evaluation Suite (human-annotated subset) (Singh et al., 2024), BLENd (Myung et al., 2024), MKQA (Longpre et al., 2021), and Global-PIQA-OE² (Chang et al., 2025) to evalu-

²While Global-PIQA is typically an MCQ dataset, we elicit open-ended responses for questions without providing

Dataset	$C_{ctx} - C_{std}$		
	Llama	Mistral	GPT
<i>Open-ended QA (chrF)</i>			
Aya	+4.70	+2.28*	+0.87*
BLEnD	+2.67	+0.77	0.00*
Global-PIQA-OE	+2.31	+1.68	+0.58
MKQA	+1.25	+0.8	+1.66
<i>Multiple-choice QA (accuracy)</i>			
Global-MMLU	+2.36	+5.78	−0.56*
Belebele	+1.29	+1.32	+4.74
Global-PIQA	+0.63	+7.39	−3.74
MCSQA	+0.50*	+2.00*	−0.41*
<i>Math (accuracy)</i>			
MGSM	+9.37	−1.27	−0.90*
MMath	+2.70	+0.30*	+0.00*

Table 2: Gains of the context-aware cascade (C_{ctx}) over the standard cascade (C_{std}) across ten benchmarks and three backbone models. Columns give signed deltas, with green for improvement and red for degradation, shade intensity scaled to magnitude. Results that are *not* statistically significant with a BH-corrected Wilcoxon signed-rank test (Wilcoxon, 1945; Benjamini and Hochberg, 1995) with $p > 0.05$ are denoted with *.

ate open-ended generation requiring cultural, commonsense, and factual grounding.

- **Multiple-choice QA:** We evaluate on Global-MMLU (Singh et al., 2025), Global-PIQA, MCSQA (Sakai et al., 2024), and Belebele (Bandarkar et al., 2024) as MCQ-style tasks for cultural, factual, and commonsense knowledge and reading comprehension.
- **Math and exact-answer reasoning:** We use MGSM (Shi et al., 2023) and MMath (Luo et al., 2025) to evaluate exact-answer reasoning.

Languages and resource levels. We evaluate on all languages supported by each dataset, totaling 285 languages across all benchmarks, consisting of 94 high-resource languages, 77 mid-resource languages, and 114 low-resource languages.³ We provide dataset language coverage in Appendix B.

Evaluation. For multiple-choice tasks we report **accuracy** after extracting the predicted option with a deterministic parser. For open-ended tasks we report **chrF** (Popović, 2015) against the gold reference after Unicode NFKC normalization and lower-casing, and refer to this as “Global-PIQA-OE”.

³These are categorized into resource tiers following Singh et al. (2025) based on the amount of available web data.

		% Languages where $C_{ctx} > C_{std}$		
Dataset	# Langs	Llama	Mistral	GPT
<i>Open-ended QA (chrF)</i>				
Aya	6	+100.0	+83.3*	+83.3*
BLEnD	14	+100.0	+85.7	+42.9*
Global-PIQA-OE	116	+88.8	+86.2	+70.7
MKQA	25	+92.0	+100.0	+88.0
<i>Multiple-choice QA (accuracy)</i>				
Global-MMLU	42	+84.6	+100.0	+20.0*
Belebele	121	+78.3	+66.3	+89.9
Global-PIQA	116	+65.2*	+92.5	+7.1
MCSQA	8	+66.7*	+62.5*	+33.3*
<i>Math (accuracy)</i>				
MGSM	11	+100.0	+0.0	+0.0*
MMath	10	+100.0	+75.0*	+0.0*

Table 3: Language-level win rates (%) for the context-aware cascade (C_{ctx}) over the standard cascade (C_{std}). Results that are *not* statistically significant with a BH-corrected (Benjamini and Hochberg, 1995) binomial sign test, $p > 0.05$ are marked with *.

casing. We also evaluate open-ended answer quality on a sampled subset of outputs using GPT-5.4-mini as an LLM judge to verify that differences in chrF are consistent with differences in answer quality. For math tasks we report **exact-match accuracy** on the final numerical answer after stripping units and normalizing decimal separators.

4 Results and Discussion

4.1 When Does Context Help?

Table 2 summarizes results for all datasets and models, averaged over languages.⁴ In general, we see broad improvement: C_{ctx} significantly outperforms C_{std} on 18 of 30 model $\times$ dataset settings and is neutral (no significant difference) on 10. We also look at **language wins** for each setting in Table 3, i.e. the fraction of languages per dataset for which C_{ctx} wins over C_{std} (ties excluded from the denominator), and find that C_{ctx} benefits more than half of the languages for Llama and Mistral. We provide absolute scores and per-language breakdowns in Appendix D.

Tasks with context-sensitive responses benefit from context-aware cascade. C_{ctx} outperforms C_{std} consistently on tasks requiring cultural grounding, including on open-ended benchmarks like Aya

⁴We also evaluate the end-to-end approach E2E, where the model solves the task natively in the target language, shown in Appendix D. We find that a cascaded approach wins in most model-dataset settings, confirming previous studies about the relevance of cascaded systems (Liu et al., 2025).

Dataset	Llama-3.1-8B-Instruct		Mistral-7B-Instruct-v0.3	
	C_{ctx}	gap closed	C_{ctx}	gap closed
Aya	14.26	92%	11.99	46%
Global-PIQA-OE	14.31	38%	14.58	32%
BLEnD	12.14	34%	9.05	9%
MKQA	14.62	10%	12.09	5%

Table 4: Fraction of the open-model-to-GPT-4o-mini gap closed by the context-aware cascade, on the four open-ended benchmarks. *Gap closed* is $(C_{ctx} - C_{std}) / (GPT-4o-mini_{std} - C_{std})$, the share of the standard-cascade gap from the open model to GPT-4o-mini that the context-aware cascade recovers. On Aya, the most culturally grounded benchmark, Llama with C_{ctx} nearly closes the gap to the larger proprietary model. GPT-4o-mini standard-cascade reference scores: Aya 14.65, BLEnD 17.26, Global-PIQA-OE 18.15, MKQA 26.25 (chrF).

Evaluation Suite (e.g., with Aya Arabic going from 8.67 $\rightarrow$ 13.63 chrF), as well as MCQ tasks like Global-PIQA. This is intuitive: C_{ctx} is helpful when the target-language answer is likely to require information that may be lost in the English language reasoning, such as language- and culture-specific grounding information. On the other hand, for short, self-contained responses, as in Math, C_{ctx} may match C_{std} or even degrade performance. Including more context may mislead the final answer, and therefore is not trivially beneficial, motivating application only in relevant scenarios.

Answer quality evaluation supports chrF results.

We use GPT-5.4-mini as an LLM judge on sampled open-ended outputs to test whether chrF improvements reflect genuine improvements in answer quality. Evaluation takes into consideration factuality, relevance, and cultural grounding. We sample 100 open-ended examples—25 from each dataset—with each example evaluated for all three backbone models, yielding 300 pairwise judgments. The sample spans 70 dataset-specific language or locale groups (6 in Aya, 14 in BLEnD, and 25 each in Global-PIQA-OE and MKQA), providing broad multilingual coverage despite its limited size. As shown in Table 5, GPT-5.4-mini prefers C_{ctx} in 36.3% of comparisons, compared with 16.3% for C_{std} , while 47.3% are judged as ties. Among non-tied comparisons, C_{ctx} accounts for 69.0% of wins and has more wins than losses in 11 of 12 model–dataset settings. These results provide complementary evidence that the gains observed with chrF reflect improvements in answer quality rather than only greater surface-form overlap. We provide details on evaluation and

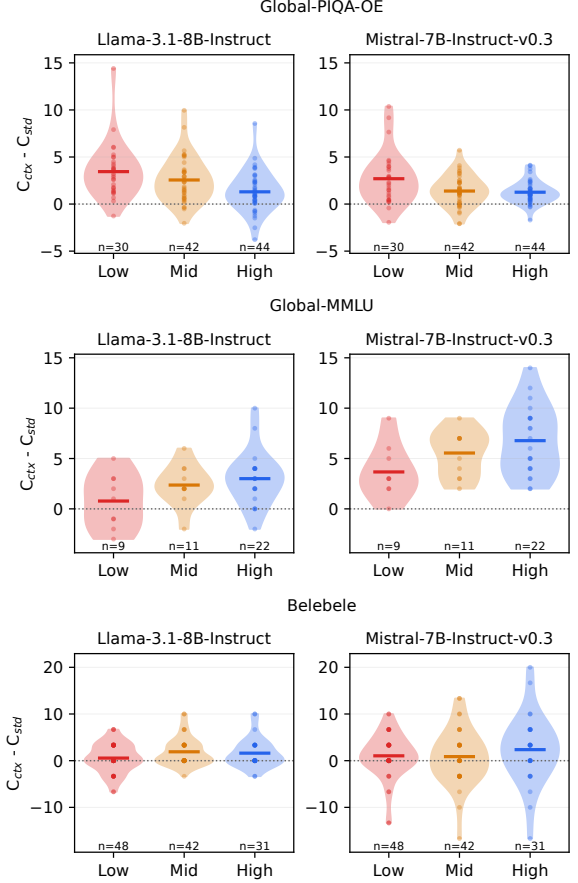


Figure 2: Gains of the context-aware cascade (C_{ctx}) over the standard cascade (C_{std}) across resource levels on Global-PIQA-OE, Global-MMLU, and Belebele for Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3.

human validation of the judge in Appendix E.

Benefits are consistent across language resourcedness. Figure 2 shows gains made by C_{ctx} over C_{std} for each tier of resourcedness for Llama and Mistral over all languages in Global-PIQA-OE, Global-MMLU, and Belebele. We see that C_{ctx} benefits languages in all tiers, indicating that a considerable part of the gains come from lossiness inherent to the structure of the standard cascade.

4.2 Weaker backbone models benefit more from context

We find that C_{ctx} benefits smaller models such as Llama and Mistral broadly, but gives fewer benefits for GPT-4o-mini, while still helping for several datasets and languages. In fact, the gap-closing results in Table 4 show that improvements from the context-aware final stage can recover a substantial portion of the performance gap between the open models and GPT-4o-mini. The recovered share is

Dataset	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3			GPT-4o-mini			All models		
Aya	48	36	16	36	36	28	32	52	16	38.7	41.3	20.0
BLEnD	36	52	12	32	60	8	28	68	4	32.0	60.0	8.0
Global-PIQA-OE	44	36	20	24	52	24	48	32	20	38.7	40.0	21.3
MKQA	52	32	16	28	56	16	28	56	16	36.0	48.0	16.0
Overall	45	39	16	30	51	19	34	52	14	36.3	47.3	16.3

Table 5: Cells report C_{ctx} win / tie / C_{std} win percentages in a pairwise semantic evaluation of sampled open-ended outputs using GPT-5.4-mini as an LLM judge. Shade intensity is scaled to the magnitude of the non-tied preference between C_{ctx} and C_{std} . Each model–dataset cell contains 25 judgments.

Model	OK	STRUCT.	ENTITY	EVENT	CULTURAL	HALLUC.
GPT-4o-mini	65.05	20.78	0.87	7.52	1.12	4.66
Llama-3.1-8B-Instruct	40.00	23.33	2.94	16.72	1.84	15.16
Mistral-7B-Instruct-v0.3	26.70	16.14	4.56	19.09	2.19	31.32

Table 6: Distribution over all datasets of GPT-5.4-mini translation-judged error labels (%). STRUCT. denotes structural/formatting error, ENTITY denotes referent/entity substitution, EVENT denotes event/constraint distortion, CULTURAL denotes cultural/local-term mistranslation, and HALLUC. denotes hallucination/over-answering. The most prominent errors across all models are structural/formatting errors, event/constraint distortions, and hallucination/over-answering.

largest on Aya, where C_{ctx} closes 92% of the gap for Llama. This may be because C_{ctx} serves as an error recovery mechanism for reasoning or translation flaws in the pipeline. Weaker models may make “unforced” translation or reasoning errors which additional context may mitigate, whereas stronger models are already capable of avoiding such errors. For example, as shown in the Math block in Table 2, Llama still benefits from C_{ctx} on tasks that require minimal language context: C_{ctx} improves over C_{std} by +9.37 points on MGSM and +2.70 points on MMath. This suggests that translation flaws in the pipeline can introduce errors in math tasks and C_{ctx} helps mitigate these errors. The extent of recovery varies considerably across datasets, suggesting that C_{ctx} can address pipeline errors like translation but cannot fully compensate for broader reasoning capability gaps between models.

Weaker backbone models produce less faithful translations. We verify this for MT1 translation via an analysis of translation errors made by the MT1 modules of our compared systems. Specifically, for each (q_t, q_e) pair, we use GPT-5.4-mini as a judge to determine whether the English text preserves the meaning of the original input. We then develop a translation error taxonomy from a manual inspection of model translations, focusing on recurring failure modes that could change the meaning of the downstream task. More details on how the error taxonomy was developed and human-judge agreement

can be found in Appendix F.

- **STRUCTURAL/FORMATting ERROR** covers errors such as leaked templates, untranslated source-script text, missing question bodies, repeated options, placeholders, or meta-text.
- **REFERENT/ENTITY SUBSTITUTION** covers cases where a central person, place, object, or other noun phrase is replaced by an incorrect one.
- **EVENT/CONSTRAINT DISTORTION** covers cases where the same core referents remain but the action, relation, negation, condition, comparison, or quantity changes.
- **CULTURAL/LOCAL-TERM MISTRANSlation** covers cases where a culture-specific term such as a food, festival, institution, clothing item, household item, or local artifact is translated literally or mapped to the wrong referent.
- **HALLUCINATION/OVER-ANSWERING** covers cases where the model invents content not present in the source, replaces the source with an unrelated question or statement, or answers/explains the task instead of translating it.
- **OK** indicates that the translation is faithful.

Table 6 reports the resulting label distribution for GPT-4o-mini, Llama-3.1-8B-Instruct, and Mistral-

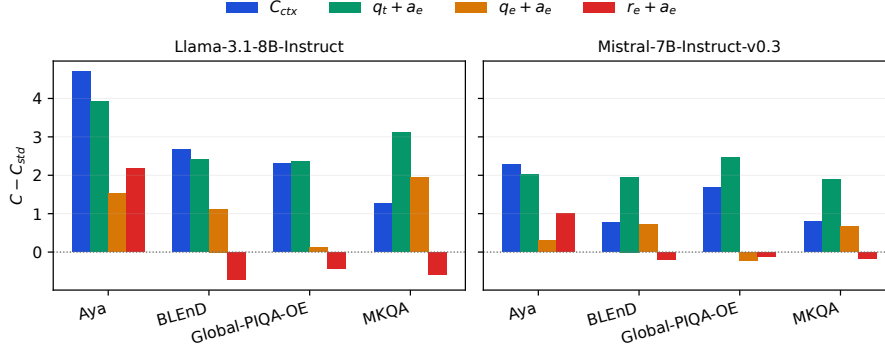


Figure 3: Improvements over C_{std} on open-ended datasets for Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. Bars show the change in chrF when MT2 is given full context (C_{ctx}) or ablated context variants: $q_t + a_e$, $q_e + a_e$, and $r_e + a_e$.

7B-Instruct-v0.3. We find that GPT-4o-mini produces higher-quality translations than the other two models (65% faithful translations versus 40% and 27%), with a considerable proportion of errors related to hallucination across all models. The high rate of translation errors from the weaker backbone models indicates that they have more scope for error recovery from context than stronger models. Thus, for weaker models, better cascade choices serve as a cheap strategy for boosting performance.

4.3 Which Context Should Be Retained?

We perform an ablation by individually providing (a) the original question q_t , (b) the English question q_e , and (c) the reasoning trace r_e , alongside the English answer a_e , in order to understand which component provides the most benefit. Figure 3 indicates that providing only the original question q_t and English answer a_e is competitive on open-ended tasks: it outperforms C_{ctx} on three of four open-ended datasets with Mistral-7B-Instruct-v0.3 and two of four with Llama-3.1-8B-Instruct. Providing the reasoning trace r_e alongside the English answer a_e , on the other hand, yields little improvement and can even degrade performance. Detailed per-dataset ablation results are provided in Appendix G.

5 Takeaways

We present two actionable takeaways from our study: (1) Treat the information flow of the cascade as a design factor rather than as fixed preprocessing. (2) Treat the output-side translation stage as a translation-oriented reasoning module rather than isolated translation, and preserve the user’s original input until that stage.

6 Related Work

Translation cascades match or exceed end-to-end multilingual generation on many tasks (Liu et al., 2025; Benvivogli et al., 2021; Intrator et al., 2024) such as multilingual question answering (Clark et al., 2020), cross-lingual transfer (Ebing and Glavaš, 2024), and mathematical reasoning (Shi et al., 2023), but suffer from error propagation across stages (Zhu et al., 2019; Asai et al., 2021). In end-to-end models, studies show that reasoning failures on non-English inputs arise partly because models cannot bridge between the target language and the English-dominant reasoning language (Kang et al., 2026; Bafna et al., 2025). Liu et al. (2025) show that cascades underperform on culturally-grounded inputs where translating into English discards culture-specific framing. Relatedly, Lee et al. (2026) improve culture-aware multilingual moral reasoning through culture-specific theoretical grounding and native-language reasoning; in contrast, our method preserves cultural information already present in the input throughout the cascade. Prior work also suggests that preserving cross-lingual semantic information can help: translation-following demonstrations improve cross-lingual alignment (Ranaldi et al., 2024), while selective input span translation can improve translation cascades (Paul et al., 2025). These findings motivate structural improvements to the cascade architecture.

7 Conclusion

This paper shows that default translate-reason-translate cascades for multilingual reasoning suffer from loss of cultural and language-specific cues and weak translation performance for low-resource

languages. We refashion the final translation step into a translation-oriented reasoning stage and show resulting benefits. Our studies show the value of information flow analysis for cascaded approaches, and we recommend our training-free approach for robustness in multilingual reasoning cascades.

Limitations

Our study explores a training-free modification to the information flow of the standard cascade. While automatic metrics and our sampled semantic evaluation confirm benefits of the context-aware cascade, our human evaluation is limited to validating the LLM judge on a small subset; we leave larger-scale human evaluation and more detailed qualitative analyses to future work.

We test three models, including two smaller open-source models and one strong proprietary model, and show that our approach largely benefits smaller models in our studied settings. This suggests that cascade design choices matter less for models with stronger native multilingual capabilities, similarly to how cascaded systems themselves grow less relevant in settings where native multilinguality is strong (Bentivogli et al., 2021). We leave it to future work to explore the needs of stronger models.

Finally, we evaluate the cascade as deployed (single pass reasoning with greedy decoding): exploring further decoding strategies is orthogonal to this work. We believe that better information flow design of cascades is a fundamental improvement that would benefit from further exploration discussed above.

Ethics Statement

Our work is about improving translation cascade design with large language models. We foresee no negative ethical implications of our work beyond general concerns with research in LLMs. We used AI for coding assistance and typo and grammatical error correction.

References

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021. [One question answering model for many languages with cross-lingual dense passage retrieval](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 7547–7560. Curran Associates, Inc.

Niyati Bafna, Tianjian Li, Kenton Murray, David R. Mortensen, David Yarowsky, Hale Sirin, and Daniel

Khashabi. 2025. [The translation barrier hypothesis: Multilingual generation with large language models suffers from implicit translation failure](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1541–1568, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. [The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. [Controlling the false discovery rate: A practical and powerful approach to multiple testing](#). *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. [Cascade versus direct speech translation: Do the differences still make a difference?](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, and Authors at the 5th Multilingual Representation Learning (MRL) Workshop. 2025. [Global PIQA: Evaluating physical commonsense reasoning across 100+ languages and cultures](#). Preprint, arXiv:2510.24081v1.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jenimaria Palomaki. 2020. [TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages](#). *Transactions of the Association for Computational Linguistics*, 8:454–470.

Benedikt Ebing and Goran Glavaš. 2024. [To translate or not to translate: A systematic investigation of translation-based cross-lingual transfer to low-resource languages](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5325–5344, Mexico City, Mexico. Association for Computational Linguistics.

Yotam Intrator, Matan Halfon, Roman Goldenberg, Reut Tsarfaty, Matan Eyal, Ehud Rivlin, Yossi Matias, and Natalia Aizenberg. 2024. [Breaking the language barrier: Can direct inference outperform pre-translation](#)

- in multilingual LLM applications? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 829–844, Mexico City, Mexico. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7B*. *Preprint*, arXiv:2310.06825.
- Deokhyung Kang, Seonjeong Hwang, Daehui Kim, Hyounghun Kim, and Gary Lee. 2026. *Why do multilingual reasoning gaps emerge in reasoning language models?* In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 31684–31716, San Diego, California, United States. Association for Computational Linguistics.
- Ayoung Lee, Ryan Kwon, Yunxiang Zhang, Yuxuan Liu, Peter Railton, and Lu Wang. 2026. *MET: Theory-grounded and culture-aware multilingual moral reasoning*. *Preprint*, arXiv:2607.11736.
- Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. *Is translation all you need? A study on solving multilingual tasks with large language models*. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.
- Ding Liu, Ning Ma, Fangtao Yang, and Xuebin Yang. 2019. *A survey of low resource neural machine translation*. In *2019 4th International Conference on Mechanical, Control and Computer Engineering (ICM-CCE)*, pages 39–42. IEEE.
- Llama Team, AI @ Meta. 2024. *The Llama 3 herd of models*. *Preprint*, arXiv:2407.21783.
- Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. *MKQA: A linguistically diverse benchmark for multilingual open domain question answering*. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.
- Wenyang Luo, Wayne Xin Zhao, Jing Sha, Shijin Wang, and Ji-Rong Wen. 2025. *MMATH: A multilingual benchmark for mathematical reasoning*. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11187–11202, Suzhou, China. Association for Computational Linguistics.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, V  ctor Guti  rrez-Basulto, Yazm  n Ib    ez-Garc  a, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. *BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages*. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc.
- OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>. Accessed: 2026-05-19.
- Rakesh Paul, Anusha Kamath, Kanishk Singla, Raviraj Joshi, Utkarsh Vaidya, Sanjay Singh Chauhan, and Niranjan Wartikar. 2025. *Aligning large language models to low-resource languages through LLM-based selective translation: A systematic study*. In *Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025)*, pages 69–82, Mumbai, India. Association for Computational Linguistics.
- Maja Popovi  . 2015. *chrF: character n-gram F-score for automatic MT evaluation*. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2024. *Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7961–7973, Bangkok, Thailand. Association for Computational Linguistics.
- Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. *mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14182–14214, Bangkok, Thailand. Association for Computational Linguistics.
- 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. 2023. *Language models are multilingual chain-of-thought reasoners*. In *The Eleventh International Conference on Learning Representations*.
- Shivalika Singh, Angelika Romanou, Cl  mentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. *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), pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D’souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O’Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. [Aya dataset: An open-access collection for multilingual instruction tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Veniamin Veselovsky, Berke Argın, Benedikt Stroebel, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths, and Arvind Narayanan. 2026. [Localized cultural knowledge is conserved and controllable in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 43152–43178, San Diego, California, United States. Association for Computational Linguistics.

Frank Wilcoxon. 1945. [Individual comparisons by ranking methods](#). *Biometrics Bulletin*, 1(6):80–83.

Minglai Yang, Ethan Huang, Liang Zhang, Mihai Surdeanu, William Yang Wang, and Liangming Pan. 2025. [How is LLM reasoning distracted by irrelevant context? An analysis using a controlled benchmark](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13329–13347, Suzhou, China. Association for Computational Linguistics.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. [NCLS: Neural cross-lingual summarization](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3054–3064, Hong Kong, China. Association for Computational Linguistics.

A Prompts

We provide prompts for each system (C_{ctx} , C_{std} , and E2E) and task-type (open-ended QA, multiple-choice QA, and math). For open-ended tasks, C_{std} MT2 receives only the English answer. For multiple-choice and math tasks, whose outputs are already language-neutral option labels or numerical values, C_{std} requires no additional target-language realization step. In C_{ctx} , the prompt additionally supplies the original target-language question, the English question, and the English reasoning trace, and explicitly instructs the model to answer in the cultural context of the target language.

A.1 Open-ended

Figures 4–8 show the prompts used for E2E, MT1, English reasoning, and the C_{std} and C_{ctx} MT2 stages on open-ended tasks.

```
You are an advanced reasoning assistant. You will be provided with a question in {language}.
Your task is to analyze the query step-by-step and provide a direct answer in {language}.

Instructions:
1. Reasoning Process: Before answering, break down the problem logically. Analyze the
constraints, perform necessary calculations, or outline your arguments. Enclose this entire
thought process within <think> tags.
2. Final Output: Provide only the final, concise result in {language} within <answer> tags.

Required Format:
<think> [Detailed step-by-step logic and analysis in {language} goes here] </think>
<answer> [Final answer in {language} goes here] </answer>

Input: {question}
Output:
```

Figure 4: E2E prompt for Open-ended QA.

```
Translate the following question to English: {question}

Required Format:
<translation> [English translation of the question goes here] </translation>

Output:
```

Figure 5: MT1 prompt for Open-ended QA.

```
You are an advanced reasoning assistant. You will be provided with a question. Your task is
to analyze the query step-by-step and provide a direct answer in English.

Instructions:
1. Reasoning Process: Before answering, break down the problem logically. Analyze the
constraints, perform necessary calculations, or outline your arguments. Enclose this entire
thought process within <think> tags.
2. Final Output: Provide only the final, concise result in English within <answer> tags.

Required Format:
<think> [Detailed step-by-step logic and analysis in English goes here] </think>
<answer> [Final answer in English goes here] </answer>

Input: {question}
Output:
```

Figure 6: LLM prompt for Open-ended QA.

You are an advanced translation assistant. You will be provided with a sentence in English. You need to translate the sentence to {language}.

Required Format:

<answer> [Final answer in {language} goes here] </answer>

Input: {english sentence}

Output:

Figure 7: MT2 C_{std} prompt for Open-ended QA.

You are an advanced reasoning assistant. You will be provided with a question, the English question, the English thinking process and answer. You need to answer the question in the {language}. You need to answer the question in cultural context of the {language}.

Required Format:

<answer> [Final answer in {language} goes here] </answer>

Input:{question}

English Question: {english question}

English Thinking Process: {english thinking process}

English Answer: {english answer}

Output:

Figure 8: MT2 C_{ctx} prompt for Open-ended QA.

A.2 Multiple-choice

Figures 9–12 show the prompts used for E2E, MT1, English reasoning, and the C_{ctx} MT2 stage on multiple-choice tasks. Because the standard-cascade output is already an option label, C_{std} requires no additional target-language realization (MT2) step. We provide the available options (e.g., A, B, C, D) when specifying the answer format; this varies over datasets since datasets may have different numbers of options.

You are an advanced reasoning assistant. You will be provided with a multiple-choice question in {language}. Your task is to analyze each option step-by-step and select the correct answer in {language}.

Instructions:

1. Reasoning Process: Before answering, break down the problem logically. Evaluate each option and explain why it is correct or incorrect. Enclose this entire thought process within <think> tags.
2. Final Output: Provide only the letter of the correct option ({possible answer choice letters}) within <answer> tags.

Required Format:

<think> [Detailed step-by-step logic and analysis of each option in {language} goes here]
</think>

<answer> [{possible answer choice letters}] </answer>

Input: {question}

{options text}

Output:

Figure 9: E2E prompt for Multiple-choice.

Translate the following multiple-choice question and all its options to English.

Required Format:

<question_translation> [English translation of the question goes here]
</question_translation>

{options translation format}

```
Input: {question}
{options text}
Output:
```

Figure 10: MT1 prompt for Multiple-choice.

You are an advanced reasoning assistant. You will be provided with a multiple-choice question. Your task is to analyze each option step-by-step and select the correct answer in English.

Instructions:

1. Reasoning Process: Before answering, break down the problem logically. Evaluate each option and explain why it is correct or incorrect. Enclose this entire thought process within `<think>` tags.
2. Final Output: Provide only the letter of the correct option ({possible answer choice letters}) within `<answer>` tags.

Required Format:

```
<think> [Detailed step-by-step logic and analysis of each option in English goes here]
</think>
<answer> [{possible answer choice letters}] </answer>
```

```
Input: {question}
{options text}
Output:
```

Figure 11: LLM prompt for Multiple-choice.

You are an advanced reasoning assistant. You will be provided with a multiple-choice question, the English question, the English thinking process and answer. You need to select the correct option for the question. You need to answer the question in the cultural context of the {language}.

Required Format:

```
<answer> [{possible answer choice letters}] </answer>
```

```
Input: {question}
{options text}
English Question: {english question}
English Thinking Process: {english thinking process}
English Answer: {english answer}
Output:
```

Figure 12: MT2 C_{ctx} prompt for Multiple-choice.

A.3 Math

Figures 13–16 show the prompts used for E2E, MT1, English reasoning, and the C_{ctx} MT2 stage on math tasks. Since the standard-cascade output is already a numerical value, C_{std} requires no additional target-language realization step. Unlike the open-ended setting, the C_{ctx} prompt does not request cultural grounding. It receives the original problem, its English translation, and the English solution, and produces the final answer in the required format. This setting therefore evaluates whether the retained context can support error recovery even when the output is an exact answer rather than culturally situated text.

You are an advanced mathematics reasoning assistant. You will be provided with a math problem in {language}. Your task is to solve it step-by-step and present the final answer.

Instructions:

1. Reasoning Process: Solve the problem step-by-step in {language}, showing all calculations and logical deductions. Enclose this entire process within `<think>` tags.
2. Final Answer: Place only the final answer inside `\\boxed{}` immediately after the closing `</think>` tag.

Required Format:
`<think> [Detailed step-by-step solution in {language} goes here] </think> \\boxed{[Final answer]}`
 Input: {question}
 Output:

Figure 13: E2E prompt for Math.

Translate the following question to English: {question}
 Required Format:
`<translation> [English translation of the question goes here] </translation>`
 Output:

Figure 14: MT1 prompt for Math.

You are an advanced mathematics reasoning assistant. You will be provided with a math problem. Your task is to solve it step-by-step and present the final answer.

Instructions:
 1. Reasoning Process: Solve the problem step-by-step in English, showing all calculations and logical deductions. Enclose this entire process within `<think>` tags.
 2. Final Answer: Place only the final answer inside `\\boxed{}` immediately after the closing `</think>` tag.

Required Format:
`<think> [Detailed step-by-step solution in English goes here] </think> \\boxed{[Final answer]}`
 Input: {question}
 Output:

Figure 15: LLM prompt for Math.

You are an advanced mathematics reasoning assistant. You will be provided with a math problem in {language}, along with its English translation and full English solution. Present the final answer in {language}.

Required Format:
`<answer> \\boxed{[Final answer]}` `</answer>`

Input: {question}
 English Question: {english question}
 English Thinking Process: {english thinking process}
 English Answer: {english answer}
 Output:

Figure 16: MT2 C_{ctx} prompt for Math.

A.4 Translation quality

Figure 17 shows the prompt used for our translation quality analysis in Appendix F. It contains descriptions and examples of our error taxonomy.

You are a translation quality auditor. You will be given a source-language text and a candidate translation into English. The input comes in two forms: some items are a question only; others bundle the answer options into the same sentence as the question. In both cases, compare the candidate translation against the source and decide whether it contains an error. Error categories:

1. Structural / formatting error

The output is not a clean translation --- it leaks annotation templates, repeats identical translations for distinct options, drops the question body, uses placeholders, outputs meta text, or leaves untranslated source-script text inside the English output.

Source = ``正确翻译 ( Cho tam giác ABCABC ABC nội tiếp trên đường tròn  $\omega\omega\omega$ . ... Tính giá trị  $m + nm + n m+n$ . ): ``

Correct: only the Vietnamese question, translated (no template wrapper).

Error: the stored translation keeps the Chinese annotation template 正确翻译 (...): wrapped around the text instead of containing only the translated question.

2. Referent / entity substitution

The overall topic changes because a key referent is swapped for a different one. This includes person, team, place, organization, object, instrument, food item, artifact, or other central noun phrase.

Source = ``ラムズはいつスーパーボウルでプレーしましたか``

Correct: ``When did the Rams play in the Super Bowl?``

Error: ``When was the Super Bowl that the Lions played in?`` --- the team Rams is replaced by Lions.

Source = ``He put the flute in a hot room.``

Correct: ``He put the flute in a hot room.``

Error: ``He put the clock in a hot room.`` --- the main object changes from flute to clock.

3. Event / constraint distortion

The same core referents remain, but who-did-what, the key action/relation, negation, condition, comparison, or quantity is changed or dropped.

Source = ``Akeredolu fofin de awon ọlọkada l'Ondo.``

Correct: ``Akeredolu has given new motorcycles to the riders in Ondo.``

Error: ``Akeredolu appoints new Ondo commissioners`` --- the action ``giving motorcycles`` becomes ``appointing commissioners.``

Source = ``Who was the bridge sold to?``

Correct: ``Who was the bridge sold to?``

Error: ``Who sold the bridge?`` --- the same bridge remains, but the relation changes.

4. Cultural / local-term mistranslation

A culture-specific food, idiom, institution, festival, clothing item, household item, or local artifact is translated literally or mapped to the wrong referent.

Source = ``Um misto quente é um sanduiche feito com pão.``

Correct: ``A ``misto quente`` is a sandwich made with bread.``

Error: ``A hot mix is a sandwich made with bread`` --- the Brazilian food term ``misto quente`` (ham-and-cheese toastie) is literally rendered as ``hot mix.``

5. Hallucination / over-answering

The model invents content not in the source, replaces the source with an unrelated question or statement, or solves/explains the task instead of translating it.

Source (a math problem) = ``A group of 7 friends split a bill of \$84 equally. How much does each pay?``

Correct: preserves the question.

Error: ``Each person pays \$12.`` --- the model answered the problem instead of translating it, producing a number absent from the source.

Source = ``What snack is popular at a local fair?``

Correct: ``What snack is popular at a local fair?``

Error: ``What is the current role of the national parliament?`` --- this invents a different, unrelated question.

Decision order (when more than one category could apply, assign the FIRST that matches):

1 -> The output is not a clean translation: it leaks code / templates / annotations / placeholders, drops the question body, repeats or merges options, outputs meta text like ``unknown question``, OR leaves part of the text untranslated in the source script.

5 -> Otherwise, the model invents content absent from the source, replaces the source with an unrelated question or statement, or answers / solves / explains the task instead of translating it.

2 -> Otherwise, a key referent is swapped for a DIFFERENT one (person / team / place / org / object / instrument / item / topic). A mere spelling / transliteration variant of the SAME referent (e.g. ``Ishikawa`` -> ``Ishekawa``) is NOT category 2 --- that is OK.

4 -> Otherwise, a culture-specific term (food / idiom / institution / festival / clothing / artifact) is mapped to the wrong referent.

3 -> Otherwise, the same core referents remain but an action / relation / quantity / condition / comparison / negation is changed or dropped. Use 3 ONLY as the residual meaning error --- never as a catch-all for messy output (that is 1), unrelated invented content (that is 5), or referent swaps (that are 2).

OK -> none of the above; the translation is faithful.

Mark OK when the meaning is preserved even if the English wording changes. Do NOT mark an error just because:

- the translation is a paraphrase rather than literal - a search-query fragment stays a fragment - a bare name is copied as a bare name - singular/plural, article choice, or phrasing differences do not change meaning - a close wording variant like ``type'' vs ``kind'' preserves the same meaning

The source may be a full sentence, a question fragment, a title, a search query, or just a name. Judge meaning fidelity, not grammatical style.

Output format:
Return a single JSON object and nothing else (no preamble, no markdown fences):
json{ ``error type': ``' }
Rules:

1. question: the source-language text, copied verbatim.
2. translation: the candidate English translation, copied verbatim.
3. error type: one of 1, 2, 3, 4, 5, or OK if the translation is faithful.
4. For items with bundled options, a single dropped or merged option still counts as category 1.
5. A correct answer value does not make the translation correct; judge by meaning only.
6. Before deciding, explicitly check these high-risk, easy-to-miss changes and treat any of them as category 3:
 - Negation flips: a ``not / no / never'' (or its source-language equivalent) present on one side but missing on the other (e.g. ``which is NOT a demand'' -> ``which is a demand').
 - Quantity / condition changes: numbers, percentages, quantifiers (``all / some / at least / only / most'), or time / conditional phrases (``before / after / if / except')
 - Relation changes: same referents but changed role structure (e.g. ``sold to'' vs ``sold by'', ``caused by'' vs ``caused').

A fluent, well-formed English sentence can still be wrong this way --- do NOT mark OK just because the translation reads naturally.

Now evaluate the following item:
question: {question} translation: {translation}

Figure 17: Prompt used for analyzing translation errors.

A.5 Answer quality

Figure 18 shows the prompt used for our pairwise semantic evaluation in [Appendix E](#). It instructs the judge to compare blinded candidate answers based on semantic answer quality and to treat purely stylistic or paraphrastic differences as ties.

You are a multilingual semantic answer-quality evaluator. Compare two candidate answers to the same question. Candidate order is randomized and carries no meaning.

Judge semantic answer quality, prioritizing:

- factual correctness;
- whether the answer addresses the intended question;
- preservation of culturally and linguistically relevant meaning;
- whether either answer introduces a factual error, mistranslation, or inappropriate interpretation.

Reference answers are evidence, but may be incomplete: a semantically correct alternative formulation can still be valid. Differences that are only stylistic, grammatical, or paraphrastic should normally be a tie. Do not prefer an answer merely because it is longer, more detailed, or has greater surface overlap with a reference. Use TIE when neither answer has a meaningful semantic advantage.

Return a short reason grounded in the question, reference, and candidates. Do not speculate about which system produced either candidate.

Question (original target language):
{question}

Reference answer(s), where available:
- {reference_answer_1}
- {reference_answer_2}

Candidate A:
{candidate_a}

Candidate B:
{candidate_b}

Choose A, B, or TIE according to the semantic evaluation criteria.

Figure 18: Prompt used for pairwise answer quality evaluation.

B Languages

Table 7 lists the languages evaluated per dataset with their resource tier assignment. Resource tiers follow Singh et al. (2025): high-resource languages have substantial web presence and strong multilingual model coverage; low-resource languages have limited digital resources.

Dataset	Total	High	Mid	Low
<i>Open-ended QA</i>				
Aya*	6	4	1	1
BLEnD*	14	9	1	4
Global-PIQA-OE*	116	44	42	30
MKQA	25	19	6	0
<i>Multiple-choice QA</i>				
Global-MMLU*	42	22	11	9
Belebele	121	31	42	48
Global-PIQA*	116	44	42	30
MCSQA	8	8	0	0
<i>Math</i>				
MGSM	11	7	4	0
MMath	10	9	1	0

Table 7: Resource-level breakdown by dataset. * denotes culturally grounded datasets.

Aya Evaluation Suite (human-annotated subset) (6 languages): Telugu, Yoruba, Arabic, Turkish, Portuguese, Chinese.

BLEnD (14 languages): Chinese (Simplified), Spanish (Spain), Spanish (Mexico), Indonesian, Korean (South Korea), Korean (North Korea), Greek, Persian, Arabic (Algeria), Azerbaijani, Javanese, Assamese, Hausa, Amharic.

MCSQA (8 languages): Chinese, Dutch, English, French, German, Japanese, Portuguese, Russian.

MGSM (11 languages): Chinese, English, French, German, Japanese, Russian, Spanish, Bengali, Swahili, Telugu, Thai.

MMath (10 languages): Chinese, English, French, Japanese, Spanish, Arabic, Korean, Portuguese, Thai, Vietnamese.

MKQA (25 languages): Arabic, Chinese (Hong Kong, Simplified, Taiwan), Danish, Finnish, French, German, Hebrew, Hungarian, Italian, Japanese, Khmer, Korean, Malay, Norwegian, Dutch, Polish, Portuguese, Russian, Spanish, Swedish, Thai, Turkish, Vietnamese.

Full language lists for Global-MMLU (42), Belebele (121), and Global-PIQA/Global-PIQA-OE (116) are available in the per-language results tables in §D.

C Experimental details

Hardware and API costs. All experiments were run on NVIDIA A100 (80GB) and NVIDIA L40S GPUs. For each model (Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3) and each method (E2E, C_{std} , C_{ctx}), inference across all datasets took approximately 74 hours. We used the OpenAI API for querying GPT-4o-mini and GPT-5.4-mini. Our experiments cost about 100 USD in total.

Generation parameters. For all model components (MT1, reasoner, and MT2), we use the following decoding configuration:

- `max_new_tokens`: 4096
- `batch_size`: 16

These settings are held constant across all backbone models (Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, and GPT-4o-mini), datasets, and cascade configurations (C_{std} , C_{ctx} , E2E) to ensure comparability.

D Detailed results

See our absolute score averages over languages per dataset and model in Table 8, and per-language breakdown per model below.

Dataset	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3			GPT-4o-mini		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
<i>Open-ended QA (chrF)</i>									
Aya*	9.56	11.27	14.26	9.71	12.61	11.99	14.65	15.96	15.52
BLEnD*	9.47	11.91	12.14	8.27	7.14	9.05	17.26	22.10	17.26
Global-PIQA-OE*	11.99	12.46	14.31	12.90	12.32	14.58	18.15	15.41	18.73
MKQA	13.36	11.51	14.62	11.29	10.17	12.09	26.25	23.03	27.92
<i>Multiple-choice QA (acc.)</i>									
Global-MMLU*	40.76	36.12	43.12	26.29	26.36	32.07	70.40	61.90	69.84
Belebele	56.89	53.50	58.18	24.60	33.69	25.92	53.50	71.43	58.24
Global-PIQA*	65.60	61.95	66.24	51.72	50.60	59.11	80.69	49.02	76.95
MCSQA	64.62	67.12	65.12	57.12	60.38	59.12	68.33	72.08	67.92
<i>Math (acc.)</i>									
MGSM	40.18	51.82	49.55	35.27	27.91	34.00	92.42	86.36	91.52
MMath	13.50	13.80	16.20	5.40	4.10	5.70	38.33	36.00	38.33

Table 8: Overall results per dataset and model. **Bold** marks the best system for each dataset and model; ties are all bolded. * denotes culturally grounded datasets. C_{std} = standard cascade; E2E = end-to-end; C_{ctx} = context-aware cascade.

Table 9: Per-language results on Global-MMLU (accuracy) for Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. C_{std} = standard cascade; E2E = end-to-end; C_{ctx} = context-aware cascade. **Bold** marks the best system per row within each model.

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
<i>High-resource (22)</i>						
Arabic	47.00	41.00	50.00	23.00	14.00	28.00
Chinese	54.00	35.00	54.00	27.00	33.00	31.00
Czech	48.00	34.00	49.00	33.00	27.00	39.00
Dutch	43.00	44.00	46.00	34.00	41.00	42.00
English	52.00	62.00	57.00	32.00	39.00	40.00
French	42.00	39.00	46.00	33.00	35.00	42.00
German	44.00	48.00	48.00	31.00	31.00	35.00
Greek	35.00	36.00	43.00	25.00	28.00	29.00
Indonesian	50.00	32.00	54.00	27.00	26.00	36.00
Italian	45.00	40.00	48.00	31.00	40.00	45.00
Japanese	42.00	36.00	46.00	30.00	26.00	42.00
Korean	39.00	34.00	41.00	28.00	27.00	31.00
Persian	42.00	43.00	42.00	36.00	24.00	38.00
Polish	47.00	38.00	47.00	38.00	35.00	43.00
Portuguese	40.00	37.00	50.00	26.00	30.00	36.00
Romanian	46.00	41.00	48.00	34.00	26.00	37.00
Russian	47.00	32.00	51.00	30.00	38.00	39.00
Spanish	43.00	39.00	45.00	34.00	31.00	41.00
Swedish	46.00	42.00	44.00	26.00	37.00	37.00
Turkish	43.00	40.00	45.00	32.00	33.00	37.00
Ukrainian	38.00	36.00	41.00	34.00	31.00	43.00
Vietnamese	36.00	39.00	40.00	27.00	31.00	29.00
Avg	44.05	39.45	47.05	30.50	31.05	37.27
<i>Mid-resource (11)</i>						
Bengali	42.00	33.00	46.00	20.00	24.00	29.00
Hebrew	40.00	35.00	38.00	28.00	21.00	32.00

Continued on next page

Table 9 – continued

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
Hindi	36.00	36.00	40.00	25.00	29.00	32.00
Kyrgyz	35.00	31.00	37.00	19.00	21.00	21.00
Lithuanian	42.00	33.00	44.00	23.00	23.00	26.00
Malay	41.00	42.00	44.00	31.00	25.00	38.00
Nepali	43.00	39.00	49.00	21.00	28.00	26.00
Serbian	41.00	34.00	43.00	28.00	30.00	35.00
Sinhala	37.00	34.00	39.00	20.00	4.00	23.00
Swahili	37.00	40.00	38.00	22.00	23.00	29.00
Telugu	41.00	25.00	43.00	14.00	12.00	21.00
Avg	39.55	34.73	41.91	22.82	21.82	28.36
<i>Low-resource (9)</i>						
Amharic	37.00	29.00	36.00	29.00	6.00	31.00
Chichewa	28.00	26.00	30.00	15.00	22.00	15.00
Filipino	49.00	32.00	52.00	28.00	25.00	34.00
Hausa	30.00	36.00	29.00	13.00	20.00	22.00
Igbo	31.00	29.00	32.00	15.00	28.00	20.00
Malagasy	36.00	26.00	34.00	25.00	23.00	28.00
Shona	30.00	29.00	33.00	22.00	22.00	25.00
Somali	34.00	30.00	31.00	19.00	19.00	21.00
Yoruba	33.00	30.00	38.00	16.00	19.00	19.00
Avg	34.22	29.67	35.00	20.22	20.44	23.89
Overall (42)	40.76	36.12	43.12	26.29	26.36	32.07

Table 10: Per-language results on MCSQA (accuracy) for Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. C_{std} = standard cascade; E2E = end-to-end; C_{ctx} = context-aware cascade. **Bold** marks the best system per row within each model.

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
<i>High-resource (8)</i>						
Chinese	60.00	64.00	60.00	55.00	63.00	59.00
Dutch	73.00	79.00	74.00	69.00	67.00	75.00
English	67.00	77.00	66.00	65.00	65.00	69.00
French	72.00	70.00	74.00	54.00	62.00	58.00
German	65.00	71.00	65.00	58.00	59.00	61.00
Japanese	72.00	74.00	73.00	66.00	63.00	64.00
Portuguese	73.00	64.00	72.00	57.00	63.00	56.00
Russian	35.00	38.00	37.00	33.00	41.00	31.00
Avg	64.62	67.12	65.12	57.12	60.38	59.12
Overall (8)	64.62	67.12	65.12	57.12	60.38	59.12

Table 11: Per-language results on Belebele (accuracy) for Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. C_{std} = standard cascade; E2E = end-to-end; C_{ctx} = context-aware cascade. **Bold** marks the best system per row within each model.

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
<i>High-resource (31)</i>						
Chinese (Simplified)	53.33	80.00	63.33	30.00	66.67	26.67
Chinese (Traditional)	63.33	86.67	70.00	33.33	56.67	40.00
Czech	76.67	86.67	76.67	20.00	63.33	16.67

Continued on next page

Table 11 – continued

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
Danish	76.67	83.33	76.67	23.33	56.67	26.67
Dutch	90.00	86.67	90.00	26.67	60.00	46.67
Egyptian Arabic	66.67	53.33	66.67	33.33	33.33	23.33
English	80.00	90.00	83.33	56.67	66.67	56.67
French	80.00	83.33	80.00	33.33	63.33	30.00
German	86.67	90.00	90.00	26.67	60.00	36.67
Greek	83.33	73.33	86.67	20.00	36.67	26.67
Hungarian	76.67	80.00	76.67	16.67	60.00	23.33
Indonesian	80.00	70.00	80.00	20.00	43.33	26.67
Italian	76.67	80.00	80.00	40.00	56.67	56.67
Japanese	63.33	60.00	63.33	36.67	50.00	36.67
Korean	76.67	70.00	76.67	16.67	46.67	20.00
Mesopotamian Arabic	66.67	53.33	63.33	20.00	46.67	23.33
Modern Standard Arabic	70.00	73.33	80.00	50.00	56.67	33.33
Modern Standard Arabic (Romanized)	20.00	23.33	20.00	23.33	13.33	16.67
Najdi Arabic	50.00	53.33	53.33	20.00	36.67	20.00
North Levantine Arabic	56.67	53.33	53.33	20.00	33.33	23.33
Polish	86.67	76.67	90.00	26.67	56.67	30.00
Portuguese	80.00	93.33	80.00	23.33	63.33	33.33
Romanian	70.00	73.33	73.33	16.67	50.00	23.33
Russian	66.67	76.67	66.67	33.33	60.00	30.00
Slovak	86.67	63.33	86.67	16.67	40.00	20.00
Spanish	93.33	96.67	93.33	40.00	56.67	40.00
Swedish	76.67	76.67	80.00	26.67	70.00	26.67
Turkish	66.67	73.33	66.67	36.67	43.33	36.67
Ukrainian	73.33	60.00	73.33	43.33	50.00	50.00
Vietnamese	83.33	83.33	86.67	26.67	56.67	30.00
Western Persian	76.67	66.67	76.67	43.33	50.00	43.33
Avg	72.69	73.23	74.30	29.03	51.72	31.40
<i>Mid-resource (42)</i>						
Afrikaans	86.67	83.33	86.67	30.00	43.33	30.00
Armenian	73.33	63.33	76.67	16.67	36.67	16.67
Basque	73.33	56.67	73.33	30.00	16.67	26.67
Bengali	53.33	53.33	53.33	30.00	23.33	26.67
Bengali (Romanized)	40.00	30.00	40.00	23.33	16.67	20.00
Bulgarian	80.00	76.67	80.00	26.67	53.33	26.67
Burmese	33.33	43.33	36.67	23.33	10.00	23.33
Catalan	76.67	83.33	76.67	20.00	76.67	23.33
Croatian	83.33	80.00	86.67	30.00	63.33	33.33
Eastern Panjabi	40.00	36.67	40.00	20.00	26.67	16.67
Finnish	66.67	83.33	70.00	26.67	40.00	23.33
Georgian	66.67	56.67	66.67	16.67	20.00	13.33
Gujarati	56.67	53.33	56.67	26.67	20.00	26.67
Hebrew	80.00	66.67	80.00	33.33	40.00	43.33
Hindi	66.67	60.00	63.33	30.00	26.67	40.00
Hindi (Romanized)	46.67	53.33	46.67	20.00	26.67	16.67
Icelandic	80.00	60.00	80.00	23.33	43.33	36.67
Kannada	60.00	56.67	63.33	10.00	20.00	16.67
Kazakh	60.00	60.00	60.00	10.00	30.00	13.33
Khmer	53.33	20.00	53.33	13.33	16.67	26.67
Kyrgyz	66.67	63.33	73.33	16.67	26.67	20.00
Lithuanian	66.67	56.67	70.00	33.33	43.33	40.00
Macedonian	60.00	86.67	70.00	40.00	60.00	36.67
Malayalam	46.67	43.33	50.00	10.00	23.33	13.33
Marathi	73.33	86.67	73.33	26.67	43.33	30.00
Nepali	40.00	56.67	40.00	13.33	20.00	20.00
North Azerbaijani	60.00	63.33	60.00	43.33	33.33	43.33
Northern Uzbek	70.00	70.00	73.33	36.67	26.67	40.00
Norwegian Bokmål	83.33	83.33	83.33	30.00	40.00	13.33
Serbian	76.67	76.67	80.00	36.67	56.67	36.67
Sinhala	46.67	23.33	50.00	20.00	3.33	16.67
Sinhala (Romanized)	30.00	26.67	33.33	20.00	20.00	23.33
Slovenian	70.00	73.33	76.67	26.67	50.00	23.33

Continued on next page

Table 11 – continued

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
Standard Malay	90.00	83.33	90.00	30.00	43.33	33.33
Swahili	56.67	50.00	66.67	13.33	33.33	16.67
Tagalog	53.33	56.67	56.67	26.67	33.33	23.33
Tamil	50.00	60.00	50.00	20.00	40.00	20.00
Telugu	40.00	46.67	40.00	26.67	33.33	20.00
Thai	60.00	73.33	66.67	23.33	30.00	23.33
Tosk Albanian	86.67	80.00	90.00	26.67	33.33	26.67
Urdu	90.00	70.00	90.00	46.67	30.00	53.33
Urdu (Romanized)	33.33	16.67	33.33	36.67	20.00	26.67
Avg	62.54	60.08	64.44	25.32	33.17	26.19
<i>Low-resource (48)</i>						
Amharic	33.33	20.00	33.33	30.00	13.33	16.67
Assamese	50.00	56.67	53.33	33.33	16.67	43.33
Bambara	26.67	20.00	23.33	13.33	16.67	13.33
Cebuano	60.00	50.00	56.67	36.67	30.00	36.67
Central Kurdish	73.33	60.00	70.00	30.00	26.67	16.67
Estonian	70.00	66.67	76.67	23.33	26.67	26.67
Ganda	33.33	26.67	26.67	3.33	20.00	6.67
Guarani	30.00	36.67	30.00	20.00	36.67	20.00
Haitian Creole	63.33	33.33	63.33	23.33	30.00	26.67
Halh Mongolian	46.67	33.33	50.00	30.00	13.33	30.00
Hausa	43.33	30.00	50.00	23.33	16.67	23.33
Igbo	40.00	16.67	43.33	13.33	13.33	16.67
Ilocano	46.67	43.33	46.67	33.33	26.67	36.67
Javanese	66.67	56.67	66.67	16.67	26.67	16.67
Jingpho	20.00	16.67	23.33	10.00	13.33	10.00
Kabuverdianu	63.33	46.67	63.33	20.00	33.33	23.33
Kinyarwanda	23.33	30.00	23.33	20.00	20.00	16.67
Lao	30.00	26.67	36.67	26.67	23.33	33.33
Lingala	23.33	23.33	23.33	13.33	23.33	13.33
Luo	23.33	23.33	23.33	16.67	16.67	20.00
Maltese	66.67	56.67	70.00	30.00	23.33	36.67
Maori	60.00	33.33	60.00	26.67	16.67	26.67
Nepali (Romanized)	46.67	46.67	50.00	20.00	26.67	20.00
Nigerian Fulfulde	36.67	23.33	40.00	23.33	23.33	23.33
Northern Sotho	30.00	30.00	30.00	23.33	30.00	23.33
Nyanja	33.33	46.67	33.33	16.67	20.00	23.33
Odia	53.33	33.33	56.67	20.00	13.33	16.67
Plateau Malagasy	26.67	36.67	26.67	13.33	30.00	13.33
Shan	26.67	30.00	23.33	23.33	13.33	16.67
Shona	36.67	26.67	36.67	20.00	33.33	16.67
Sindhi	53.33	46.67	56.67	23.33	20.00	26.67
Somali	16.67	23.33	16.67	10.00	23.33	10.00
Southern Pashto	56.67	43.33	60.00	16.67	16.67	16.67
Southern Sotho	26.67	33.33	30.00	20.00	13.33	23.33
Standard Latvian	60.00	73.33	56.67	30.00	36.67	23.33
Standard Tibetan	36.67	13.33	33.33	13.33	6.67	13.33
Sundanese	53.33	43.33	53.33	23.33	30.00	30.00
Swati	46.67	33.33	46.67	26.67	26.67	26.67
Tajik	66.67	50.00	70.00	20.00	23.33	23.33
Tigrinya	23.33	23.33	26.67	6.67	23.33	13.33
Tsonga	30.00	40.00	33.33	26.67	16.67	30.00
Tswana	30.00	26.67	26.67	23.33	20.00	23.33
Waray	56.67	26.67	56.67	30.00	43.33	30.00
West Central Oromo	43.33	23.33	43.33	6.67	13.33	10.00
Wolof	36.67	13.33	33.33	33.33	30.00	33.33
Xhosa	23.33	16.67	23.33	13.33	16.67	23.33
Yoruba	26.67	36.67	20.00	16.67	10.00	23.33
Zulu	33.33	33.33	33.33	20.00	36.67	20.00
Avg	41.74	35.00	42.29	21.11	22.50	22.15
Overall (121)	56.89	53.50	58.18	24.60	33.69	25.92

Table 12: Per-language results on Global-PIQA (accuracy) for Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. C_{std} = standard cascade; E2E = end-to-end; C_{ctx} = context-aware cascade. **Bold** marks the best system per row within each model.

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
<i>High-resource (44)</i>						
Algerian Arabic	50.00	60.00	46.67	40.00	36.67	60.00
Burushaski	40.00	43.33	40.00	26.67	40.00	40.00
Cantonese	70.00	60.00	70.00	60.00	46.67	66.67
Chinese (Simplified)	56.67	56.67	60.00	56.67	63.33	70.00
Chinese (Traditional)	56.67	50.00	53.33	40.00	56.67	53.33
Czech	73.33	63.33	73.33	56.67	40.00	70.00
Dutch	70.00	66.67	70.00	60.00	56.67	66.67
Egyptian Arabic	60.00	36.67	56.67	40.00	53.33	50.00
English	73.33	73.33	73.33	63.33	76.67	66.67
French (Canada)	86.67	80.00	90.00	76.67	50.00	80.00
French (France)	73.33	66.67	73.33	66.67	63.33	66.67
German	66.67	66.67	73.33	46.67	50.00	50.00
Greek	60.00	56.67	63.33	30.00	50.00	40.00
Gulf Arabic	40.00	66.67	36.67	63.33	43.33	70.00
Hungarian	83.33	83.33	83.33	83.33	80.00	86.67
Indonesian	83.33	86.67	86.67	76.67	73.33	83.33
Italian	76.67	60.00	76.67	60.00	43.33	70.00
Japanese	70.00	53.33	73.33	60.00	63.33	66.67
Korean	53.33	60.00	53.33	46.67	43.33	56.67
Mesopotamian Arabic	63.33	63.33	60.00	53.33	70.00	63.33
Modern Standard Arabic	63.33	60.00	63.33	53.33	60.00	46.67
Moroccan Arabic	50.00	53.33	50.00	40.00	66.67	56.67
Najdi Arabic	76.67	63.33	80.00	46.67	30.00	53.33
North Levantine Arabic (Jordan)	60.00	46.67	60.00	66.67	40.00	70.00
North Levantine Arabic (Lebanon)	73.33	70.00	73.33	50.00	66.67	70.00
North Levantine Arabic (Palestine)	56.67	56.67	56.67	46.67	50.00	63.33
North Levantine Arabic (Syria)	70.00	60.00	70.00	50.00	53.33	60.00
Polish	50.00	60.00	53.33	46.67	46.67	53.33
Portuguese (Brazil)	66.67	80.00	66.67	50.00	76.67	70.00
Portuguese (Portugal)	93.33	86.67	93.33	73.33	66.67	80.00
Romanian	90.00	86.67	93.33	86.67	73.33	93.33
Russian	80.00	56.67	83.33	60.00	56.67	66.67
Slovak	60.00	70.00	56.67	53.33	36.67	60.00
Slovak (SARI)	53.33	60.00	53.33	63.33	36.67	63.33
Spanish (Mexico)	76.67	76.67	76.67	76.67	80.00	90.00
Spanish (Peru)	90.00	83.33	90.00	83.33	70.00	90.00
Spanish (Spain)	66.67	90.00	63.33	66.67	66.67	70.00
Swedish	66.67	80.00	70.00	60.00	66.67	66.67
Ta'izzi-Adeni Arabic	60.00	66.67	63.33	63.33	53.33	73.33
Tunisian Arabic	63.33	63.33	63.33	56.67	53.33	56.67
Turkish	76.67	80.00	76.67	60.00	66.67	73.33
Ukrainian	80.00	76.67	80.00	56.67	56.67	73.33
Vietnamese	60.00	53.33	60.00	46.67	43.33	53.33
Western Persian	70.00	36.67	76.67	46.67	50.00	50.00
Avg	67.27	65.23	67.88	57.05	56.06	65.45
<i>Mid-resource (42)</i>						
Armenian	60.00	70.00	63.33	10.00	53.33	20.00
Belarusian	80.00	63.33	80.00	53.33	36.67	63.33
Bengali	80.00	80.00	83.33	30.00	53.33	43.33
Bengali (Romanized)	33.33	43.33	33.33	30.00	16.67	36.67
Bosnian	86.67	90.00	86.67	90.00	73.33	93.33
Bulgarian	86.67	73.33	86.67	56.67	53.33	73.33
Catalan	76.67	66.67	76.67	66.67	40.00	73.33
Croatian	83.33	70.00	83.33	66.67	63.33	76.67
Eastern Panjabi	63.33	63.33	63.33	43.33	56.67	56.67
Finnish	63.33	56.67	63.33	56.67	56.67	60.00
Galician	66.67	60.00	70.00	56.67	43.33	63.33

Continued on next page

Table 12 – continued

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
Georgian	56.67	63.33	56.67	60.00	46.67	56.67
Gujarati	90.00	76.67	90.00	40.00	36.67	46.67
Hebrew	60.00	60.00	63.33	46.67	33.33	53.33
Hindi	80.00	76.67	80.00	63.33	53.33	63.33
Icelandic	86.67	76.67	86.67	56.67	40.00	66.67
Kannada	76.67	70.00	83.33	36.67	56.67	50.00
Kazakh	50.00	60.00	50.00	23.33	33.33	30.00
Kyrgyz	60.00	63.33	60.00	63.33	20.00	63.33
Lithuanian	66.67	76.67	63.33	60.00	80.00	56.67
Macedonian	76.67	66.67	80.00	60.00	60.00	73.33
Malayalam	60.00	43.33	60.00	43.33	23.33	46.67
Marathi	90.00	60.00	90.00	60.00	43.33	66.67
Nepali	70.00	70.00	70.00	60.00	60.00	63.33
North Azerbaijani	76.67	50.00	76.67	36.67	50.00	46.67
Northern Uzbek	56.67	60.00	53.33	46.67	46.67	60.00
Norwegian Bokmål	80.00	60.00	80.00	76.67	60.00	76.67
Norwegian Nynorsk	83.33	63.33	86.67	43.33	53.33	46.67
Serbian (Cyrillic)	56.67	53.33	53.33	70.00	33.33	53.33
Serbian (Latin)	70.00	60.00	70.00	53.33	63.33	53.33
Sinhala	46.67	56.67	53.33	36.67	10.00	40.00
Slovenian	73.33	40.00	73.33	53.33	40.00	60.00
Slovenian (CERK)	43.33	46.67	53.33	36.67	50.00	40.00
Standard Malay	86.67	86.67	86.67	70.00	76.67	73.33
Swahili	76.67	66.67	76.67	60.00	56.67	63.33
Tagalog	76.67	60.00	76.67	60.00	50.00	50.00
Tamil	60.00	70.00	60.00	33.33	26.67	33.33
Telugu	90.00	66.67	90.00	40.00	33.33	36.67
Thai	83.33	76.67	83.33	46.67	56.67	60.00
Tosk Albanian	63.33	46.67	63.33	60.00	43.33	60.00
Urdu	86.67	73.33	86.67	60.00	60.00	73.33
Urdu (Romanized)	56.67	70.00	56.67	56.67	30.00	60.00
Avg	70.72	64.44	71.51	51.75	46.98	56.75
<i>Low-resource (30)</i>						
Amharic	60.00	50.00	56.67	40.00	26.67	53.33
Assamese	83.33	73.33	83.33	33.33	60.00	40.00
Bambara	43.33	53.33	43.33	36.67	30.00	50.00
Bhojpuri	56.67	70.00	63.33	46.67	43.33	53.33
Central Kurdish	56.67	53.33	56.67	33.33	50.00	43.33
Dhundari	80.00	46.67	80.00	50.00	56.67	63.33
Ekpeye	56.67	43.33	56.67	40.00	56.67	43.33
Estonian	53.33	53.33	50.00	36.67	46.67	40.00
Faroese	46.67	46.67	53.33	43.33	43.33	70.00
Hausa	46.67	43.33	46.67	46.67	50.00	56.67
Hawaiian	43.33	56.67	43.33	36.67	43.33	53.33
Idoma	40.00	40.00	40.00	30.00	40.00	50.00
Igbo	50.00	30.00	50.00	33.33	43.33	50.00
Isoko	50.00	70.00	53.33	50.00	36.67	53.33
Javanese	53.33	56.67	50.00	53.33	50.00	63.33
Kinyarwanda	63.33	46.67	66.67	43.33	26.67	50.00
Kumzari	53.33	36.67	53.33	36.67	46.67	56.67
Lingala	43.33	53.33	43.33	33.33	60.00	43.33
Luo	43.33	40.00	43.33	36.67	66.67	50.00
Marwari	73.33	66.67	73.33	60.00	56.67	60.00
Meitei (Bengali script)	66.67	40.00	70.00	40.00	33.33	50.00
Meitei (Meitei script)	43.33	50.00	46.67	50.00	43.33	53.33
Nagamese	43.33	46.67	40.00	43.33	53.33	66.67
Nigerian Pidgin	80.00	80.00	80.00	66.67	80.00	73.33
Sindhi	100.00	96.67	100.00	66.67	36.67	60.00
Sindhi (Devanagari)	63.33	56.67	66.67	56.67	60.00	43.33
Urhobo	46.67	53.33	43.33	36.67	40.00	50.00
Uyghur	46.67	46.67	43.33	50.00	53.33	53.33
Yoruba	43.33	50.00	43.33	40.00	36.67	46.67
Zulu	50.00	60.00	53.33	46.67	60.00	53.33

Continued on next page

Table 12 – continued

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
Avg	56.00	53.67	56.44	43.89	47.67	53.11
Overall (116)	65.60	61.95	66.24	51.72	50.60	59.11

Table 13: Per-language results on MGSM (accuracy) for Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. C_{std} = standard cascade; E2E = end-to-end; C_{ctx} = context-aware cascade. **Bold** marks the best system per row within each model.

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
<i>High-resource (7)</i>						
Chinese	38.00	58.00	50.00	37.00	38.00	37.00
English	36.00	66.00	56.00	53.00	56.00	53.00
French	46.00	52.00	56.00	49.00	43.00	48.00
German	46.00	61.00	54.00	46.00	37.00	41.00
Japanese	36.00	48.00	45.00	34.00	25.00	34.00
Russian	52.00	56.00	65.00	51.00	43.00	47.00
Spanish	49.00	58.00	50.00	46.00	33.00	45.00
Avg	43.29	57.00	53.71	45.14	39.29	43.57
<i>Mid-resource (4)</i>						
Bengali	29.00	40.00	35.00	24.00	6.00	23.00
Swahili	50.00	44.00	59.00	12.00	3.00	11.00
Telugu	30.00	35.00	39.00	7.00	4.00	6.00
Thai	30.00	52.00	36.00	29.00	19.00	29.00
Avg	34.75	42.75	42.25	18.00	8.00	17.25
Overall (11)	40.18	51.82	49.55	35.27	27.91	34.00

Table 14: Per-language results on MMath (accuracy) for Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. C_{std} = standard cascade; E2E = end-to-end; C_{ctx} = context-aware cascade. **Bold** marks the best system per row within each model.

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
<i>High-resource (9)</i>						
Arabic	18.00	8.00	21.00	7.00	1.00	7.00
Chinese	11.00	14.00	13.00	7.00	5.00	7.00
English	14.00	17.00	17.00	7.00	8.00	7.00
French	15.00	15.00	17.00	8.00	4.00	10.00
Japanese	12.00	15.00	15.00	3.00	6.00	3.00
Korean	9.00	15.00	11.00	4.00	1.00	3.00
Portuguese	14.00	15.00	20.00	5.00	6.00	6.00
Spanish	16.00	16.00	19.00	4.00	4.00	4.00
Vietnamese	15.00	10.00	18.00	4.00	2.00	5.00
Avg	13.78	13.89	16.78	5.44	4.11	5.78
<i>Mid-resource (1)</i>						
Thai	11.00	13.00	11.00	5.00	4.00	5.00
Avg	11.00	13.00	11.00	5.00	4.00	5.00
Overall (10)	13.50	13.80	16.20	5.40	4.10	5.70

Table 15: Per-language results on Aya (chrF) for Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. C_{std} = standard cascade; E2E = end-to-end; C_{ctx} = context-aware cascade. **Bold** marks the best system per row within each model.

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
<i>High-resource (4)</i>						
Arabic	8.67	9.43	13.63	9.25	9.18	11.21
Chinese	5.02	6.46	10.49	4.87	10.12	8.35
Portuguese	19.72	22.54	27.29	22.52	26.96	26.79
Turkish	14.33	15.64	20.94	15.41	17.67	18.02
Avg	11.94	13.52	18.09	13.01	15.98	16.09
<i>Mid-resource (1)</i>						
Telugu	7.93	11.01	11.09	3.36	8.19	5.09
Avg	7.93	11.01	11.09	3.36	8.19	5.09
<i>Low-resource (1)</i>						
Yoruba	1.66	2.55	2.11	2.86	3.51	2.50
Avg	1.66	2.55	2.11	2.86	3.51	2.50
Overall (6)	9.56	11.27	14.26	9.71	12.61	11.99

Table 16: Per-language results on BLENd (chrF) for Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. C_{std} = standard cascade; E2E = end-to-end; C_{ctx} = context-aware cascade. **Bold** marks the best system per row within each model.

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
<i>High-resource (9)</i>						
Arabic	11.82	12.13	12.82	10.47	5.23	10.80
Chinese	5.23	26.31	8.95	6.03	8.37	7.53
Greek	11.50	13.32	14.23	10.65	5.68	11.28
Indonesian	13.27	12.11	16.63	13.16	13.09	13.18
Korean (KP)	2.95	5.32	5.71	1.85	4.34	3.64
Korean (KR)	6.42	11.29	8.43	4.43	4.41	5.33
Persian	11.72	12.88	13.58	9.16	5.52	9.15
Spanish (ES)	14.97	13.79	18.01	15.81	17.41	16.52
Spanish (MX)	14.94	14.70	16.66	15.29	16.06	16.89
Avg	10.31	13.54	12.78	9.65	8.90	10.48
<i>Mid-resource (1)</i>						
Azerbaijani	10.20	13.04	15.07	8.81	6.83	9.31
Avg	10.20	13.04	15.07	8.81	6.83	9.31
<i>Low-resource (4)</i>						
Amharic	4.59	4.05	5.65	1.39	0.82	2.51
Assamese	9.08	11.17	11.79	6.16	2.49	7.74
Hausa	6.20	6.77	9.04	3.66	3.84	3.54
Javanese	9.74	9.93	13.45	8.95	5.86	9.24
Avg	7.40	7.98	9.98	5.04	3.25	5.76
Overall (14)	9.47	11.91	12.14	8.27	7.14	9.05

Table 17: Per-language results on Global-PIQA-OE (chrF) for Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. C_{std} = standard cascade; E2E = end-to-end; C_{ctx} = context-aware cascade. **Bold** marks the best system per row within each model.

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
<i>High-resource (44)</i>						
Algerian Arabic	10.56	9.06	9.92	12.25	5.37	12.00
Burushaski	2.66	5.23	4.32	2.92	3.64	5.28
Cantonese	2.56	3.53	3.94	1.79	1.66	3.45
Chinese (Simplified)	3.49	4.48	6.46	2.70	7.19	6.12
Chinese (Traditional)	3.84	4.98	8.71	2.87	9.21	7.00
Czech	11.08	8.77	13.65	14.27	14.49	15.87
Dutch	21.72	18.39	17.94	23.82	23.43	22.12
Egyptian Arabic	12.81	8.65	13.27	14.01	5.79	15.06
English	19.23	15.56	17.74	21.18	25.88	24.93
French (Canada)	18.74	16.82	21.85	24.10	22.63	24.52
French (France)	13.97	17.78	14.24	20.78	20.38	21.38
German	15.54	16.25	16.55	19.28	18.55	20.31
Greek	9.82	7.81	13.67	10.08	10.36	11.00
Gulf Arabic	11.01	8.45	10.23	11.47	5.65	12.41
Hungarian	16.63	14.42	17.92	21.76	17.07	21.96
Indonesian	21.30	17.89	18.78	22.96	21.92	22.88
Italian	12.81	24.10	21.37	18.80	21.24	22.88
Japanese	5.55	4.99	5.62	4.11	2.70	6.27
Korean	4.04	4.30	4.89	3.34	3.83	5.02
Mesopotamian Arabic	9.92	8.87	10.71	12.67	5.13	14.31
Modern Standard Arabic	12.88	9.04	12.10	12.31	7.82	12.29
Moroccan Arabic	10.54	9.08	10.61	12.14	5.77	12.48
Najdi Arabic	11.39	9.78	14.47	15.86	7.73	17.24
North Levantine Arabic (Jordan)	13.65	9.96	15.15	15.56	10.06	17.53
North Levantine Arabic (Lebanon)	7.90	6.79	10.36	12.53	5.47	12.77
North Levantine Arabic (Palestine)	7.28	6.80	9.64	11.06	5.06	11.55
North Levantine Arabic (Syria)	7.37	6.80	9.16	11.82	6.02	12.83
Polish	16.41	14.34	17.24	19.63	17.51	20.28
Portuguese (Brazil)	16.65	22.17	18.74	18.81	19.04	18.49
Portuguese (Portugal)	15.67	14.31	15.82	18.98	20.91	20.28
Romanian	16.10	13.90	16.51	19.73	18.21	20.20
Russian	15.59	14.60	16.62	20.71	22.50	21.98
Slovak	13.62	16.23	17.80	15.81	17.14	17.08
Slovak (SARI)	10.97	8.36	12.06	11.03	10.12	11.82
Spanish (Mexico)	16.55	19.11	17.06	25.11	25.92	26.49
Spanish (Peru)	18.86	18.13	17.86	23.18	22.98	23.79
Spanish (Spain)	18.43	17.09	19.28	22.80	23.07	23.87
Swedish	19.74	15.88	20.51	21.71	20.82	24.33
Ta'izzi-Adeni Arabic	10.25	8.82	12.54	13.65	5.27	16.13
Tunisian Arabic	11.46	9.70	10.20	12.79	5.48	13.66
Turkish	17.12	16.27	19.41	18.94	19.69	19.94
Ukrainian	14.13	9.81	14.46	16.47	14.80	17.49
Vietnamese	10.64	11.57	14.42	12.82	17.69	14.38
Western Persian	14.63	12.95	18.55	15.35	14.67	17.70
Avg	12.62	11.86	13.92	15.09	13.41	16.35
<i>Mid-resource (42)</i>						
Armenian	13.15	9.77	14.86	11.47	15.42	14.00
Belarusian	13.93	16.24	19.60	12.64	15.78	18.36
Bengali	11.25	12.81	13.61	8.86	4.61	10.40
Bengali (Romanized)	0.15	11.23	10.11	7.00	11.54	4.92
Bosnian	18.65	17.24	18.97	23.91	21.04	23.53
Bulgarian	16.58	13.83	19.13	23.91	20.38	24.04
Catalan	15.24	16.01	16.44	20.56	17.68	21.41
Croatian	13.69	13.96	14.34	16.26	16.06	15.29
Eastern Panjabi	12.34	14.37	15.27	8.53	7.45	8.69
Finnish	19.08	17.45	22.63	19.55	17.43	20.73
Galician	15.41	16.85	20.54	18.64	20.07	22.22

Continued on next page

Table 17 – continued

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
Georgian	13.82	13.33	16.76	10.30	11.52	11.56
Gujarati	13.36	15.61	15.95	6.41	8.53	9.64
Hebrew	12.72	8.70	15.07	13.78	9.33	16.00
Hindi	13.25	13.78	14.20	10.74	11.60	13.22
Icelandic	16.29	14.66	16.83	17.49	15.03	18.06
Kannada	16.99	19.60	20.65	11.79	17.25	13.51
Kazakh	14.95	18.14	20.10	11.74	11.81	13.33
Kyrgyz	14.99	17.10	16.85	11.23	9.47	10.99
Lithuanian	12.55	10.95	14.12	11.39	8.62	10.55
Macedonian	15.27	13.04	14.91	17.52	15.18	17.41
Malayalam	15.69	19.12	19.07	10.33	13.36	12.64
Marathi	10.20	14.47	15.49	8.45	6.67	12.54
Nepali	14.29	13.12	15.68	11.13	11.11	14.98
North Azerbaijani	14.08	13.25	13.72	14.75	11.99	15.97
Northern Uzbek	11.83	15.19	16.24	13.48	16.10	13.46
Norwegian Bokmål	17.86	17.96	17.34	20.31	19.91	20.62
Norwegian Nynorsk	16.63	15.79	17.48	18.70	17.78	19.92
Serbian (Cyrillic)	5.93	9.76	10.92	16.34	13.56	17.32
Serbian (Latin)	13.30	10.81	13.77	17.49	16.05	19.01
Sinhala	8.53	9.36	9.76	5.40	4.47	6.58
Slovenian	14.04	12.22	14.70	17.23	17.44	18.29
Slovenian (CERK)	9.77	7.95	9.96	11.76	9.61	11.52
Standard Malay	23.64	16.03	28.85	17.49	25.73	20.77
Swahili	14.80	13.55	12.76	10.75	13.24	13.03
Tagalog	21.37	19.60	24.72	25.77	26.88	27.56
Tamil	21.73	21.00	25.13	13.48	17.78	16.94
Telugu	17.07	18.59	20.48	7.86	11.38	10.11
Thai	12.84	8.68	15.19	11.83	12.36	13.47
Tosk Albanian	12.42	12.18	13.89	13.28	12.54	14.24
Urdu	12.89	12.65	15.78	13.11	13.27	16.07
Urdu (Romanized)	5.32	10.48	13.47	13.68	12.02	11.59
Avg	14.00	14.20	16.56	13.96	14.02	15.35
<i>Low-resource (30)</i>						
Amharic	4.04	6.45	7.72	1.66	1.29	2.58
Assamese	9.74	12.21	12.59	8.10	2.63	8.86
Bambara	5.32	13.25	10.29	5.91	9.78	9.91
Bhojpuri	11.72	11.57	12.82	8.57	5.98	10.07
Central Kurdish	10.00	9.50	16.01	8.41	4.87	12.70
Dhundari	10.69	11.51	11.36	7.45	4.92	10.25
Ekpeye	3.89	11.05	6.54	4.38	6.28	4.64
Estonian	14.39	11.24	18.64	14.14	9.94	14.62
Faroese	14.30	11.77	17.53	14.58	14.52	17.55
Hausa	11.76	8.45	13.13	9.90	10.45	9.48
Hawaiian	12.81	11.55	15.85	11.99	12.62	16.53
Idoma	3.89	10.80	7.71	5.59	7.95	7.90
Igbo	9.20	15.23	15.25	9.51	15.30	17.16
Isoko	4.08	12.30	6.66	5.40	8.30	10.06
Javanese	12.83	11.08	14.10	15.19	11.69	15.59
Kinyarwanda	8.80	11.21	12.86	9.55	11.07	12.44
Kumzari	1.19	2.98	1.51	3.11	4.55	1.17
Lingala	8.32	10.56	10.08	9.86	11.97	11.56
Luo	10.06	10.10	8.79	9.01	11.71	10.18
Marwari	9.68	10.13	11.20	9.21	7.15	11.17
Meitei (Bengali script)	9.19	11.90	10.91	2.65	1.89	2.88
Meitei (Meitei script)	0.06	13.26	14.47	0.01	3.32	10.38
Nagamese	7.55	9.09	11.15	12.73	10.05	13.14
Nigerian Pidgin	16.14	13.67	21.15	19.35	20.24	20.81
Sindhi	11.69	12.26	13.76	7.17	3.70	9.97
Sindhi (Devanagari)	2.20	9.03	10.11	3.91	4.61	7.68
Urhobo	5.18	13.18	10.42	6.32	10.76	15.49
Uyghur	6.64	10.83	11.13	8.23	3.55	10.40
Yoruba	4.15	9.25	7.94	4.81	7.31	8.38
Zulu	8.79	12.27	10.01	9.93	11.52	13.90

Continued on next page

Table 17 – continued

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
Avg	8.28	10.92	11.72	8.22	8.33	10.91
Overall (116)	11.99	12.46	14.31	12.90	12.32	14.58

Table 18: Per-language results on MKQA (chrF) for Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. C_{std} = standard cascade; E2E = end-to-end; C_{ctx} = context-aware cascade. **Bold** marks the best system per row within each model.

Language	Llama-3.1-8B-Instruct			Mistral-7B-Instruct-v0.3		
	C _{std}	E2E	C _{ctx}	C _{std}	E2E	C _{ctx}
<i>High-resource (19)</i>						
Arabic	11.73	10.47	12.47	10.03	5.79	10.55
Chinese (HK)	4.72	7.61	8.07	5.43	3.90	6.10
Chinese (Simplified)	7.91	9.31	10.88	3.96	3.04	4.82
Chinese (Traditional)	5.24	7.89	7.68	3.63	2.75	4.10
Danish	17.48	8.65	17.57	15.62	14.22	16.59
Dutch	15.01	12.11	15.20	14.06	16.62	14.54
French	16.79	17.05	18.38	15.33	15.93	16.32
German	16.41	14.44	17.44	16.36	15.34	17.47
Hungarian	13.82	14.32	15.36	14.59	14.04	15.24
Italian	16.91	15.35	17.25	16.42	15.67	17.27
Japanese	14.23	12.73	15.90	6.40	4.07	10.74
Korean	6.71	5.09	7.07	3.88	2.46	4.70
Polish	16.25	15.35	18.12	14.64	13.01	15.19
Portuguese	15.99	10.77	17.07	15.53	16.29	16.32
Russian	13.93	12.09	14.58	12.46	11.10	12.91
Spanish	16.39	14.52	16.93	14.98	17.72	15.63
Swedish	17.31	10.32	17.64	15.97	14.52	17.28
Turkish	14.33	10.95	13.32	13.52	8.84	13.52
Vietnamese	13.79	13.07	14.45	11.69	10.48	12.44
Avg	13.42	11.69	14.49	11.82	10.83	12.72
<i>Mid-resource (6)</i>						
Finnish	16.24	15.20	18.57	11.74	10.62	12.33
Hebrew	13.86	12.43	17.15	7.96	4.86	8.36
Khmer	3.50	3.51	4.46	1.87	2.58	2.13
Malay	16.02	10.30	15.31	13.38	11.44	13.61
Norwegian	16.47	9.79	17.28	14.98	13.37	15.72
Thai	13.04	14.39	17.29	7.84	5.55	8.48
Avg	13.19	10.94	15.01	9.63	8.07	10.11
Overall (25)	13.36	11.51	14.62	11.29	10.17	12.09

E Answer quality

Sampling and judging. Examples are selected independently of chrF scores and cascade performance using deterministic sampling. Within each dataset, language or locale groups and examples within each group are shuffled, followed by round-robin selection across groups to encourage language coverage. The same sampled examples are used across all three backbone models.

We use GPT-5.4-mini as the semantic judge. For each comparison, the judge receives the original target-language question, available reference answer(s), and the two candidate outputs, with model and cascade identities omitted. Candidate order is deterministically randomized and approximately balanced within each model–dataset combination. The judge selects the first candidate, the second candidate, or a tie based on factual correctness, relevance to the intended question, preservation of culturally and linguistically relevant meaning, and factual or translation errors. Purely stylistic or paraphrastic differences are treated as ties. Each judgment also includes a brief justification. Prompt is shown in [Figure 18](#).

Human validation. Two human annotators independently evaluate 48 randomly sampled comparisons using the same blinded inputs, labels, and evaluation criteria as the LLM judge. The validation subset is selected independently of the LLM judge’s prediction.

Pairwise agreement between the two human annotators is 0.65 Cohen’s κ . Agreement between GPT-5.4-mini and the two human annotators is 0.75 and 0.54 Cohen’s κ , respectively. Three-way agreement among the two human annotators and GPT-5.4-mini is 0.65 Fleiss’ κ . These agreement levels indicate reasonably consistent judgments overall, while also reflecting some annotator variation, and support using the LLM judge for aggregate answer quality evaluation.

F Translation quality

Taxonomy development. We develop the translation error taxonomy through manual inspection of MT1 outputs across datasets and backbone models, focusing on recurring failures that alter information relevant to the downstream task. We group these failures according to the aspect of the source meaning that is changed, yielding the five error categories described in [Section 4](#): structural/formatting error, referent/entity substitution, event/constraint distortion, cultural/local-term mistranslation, and hallucination/over-answering, in addition to the OK label for faithful translations. The resulting definitions and representative examples are incorporated into the judge prompt shown in [Figure 17](#).

Human validation. To validate the judge model, two human annotators independently labeled approximately 0.25% of the translated data (across all datasets and models) using the same label definitions. The agreement scores are computed only on this human-audited subset. Pairwise agreement between the two human annotators was 0.72 Cohen’s κ . Agreement between the judge and the two human annotators was 0.68 and 0.77 Cohen’s κ , respectively. The three-way agreement among the two human annotators and the judge was 0.72 Fleiss’ κ . These agreement scores suggest that the judge’s labels are reliable enough for aggregate analysis. After this validation step, we apply the judge to the full translated data.

G Detailed ablation results

G.1 Llama-3.1-8B-Instruct

Dataset	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>Open-ended generation (chrF)</i>					
Aya	9.56	14.26	13.48	11.08	11.73
BLEnD	9.47	12.14	11.88	10.59	8.76
Global-PIQA-OE	11.99	14.31	14.35	12.11	11.57
MKQA	13.36	14.62	16.48	15.30	12.77
<i>Multiple-choice (Acc.)</i>					
Global-MMLU	40.76	43.12	38.38	39.95	38.69
Belebele	56.89	58.18	57.08	56.53	52.98
Global-PIQA	65.60	66.24	66.84	58.51	62.93
MCSQA	64.62	65.12	50.38	64.25	63.62
<i>Math (Acc.)</i>					
MGSM	40.18	49.55	38.73	33.73	39.82
MMath	13.50	16.20	8.40	10.40	10.70

Table 19: Llama-3.1-8B-Instruct overall accuracy and chrF scores with varying context provided to MT2 on all datasets

Table 20: Llama-3.1-8B-Instruct overall accuracy with varying context provided to MT2 on Global-MMLU: High-resource (22), Mid-resource (11) and Low-resource (9).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (22)</i>					
Arabic	47.00	50.00	46.00	47.00	41.00
Chinese	54.00	54.00	50.00	54.00	54.00
Czech	48.00	49.00	37.00	48.00	48.00
Dutch	43.00	46.00	40.00	42.00	44.00
English	52.00	57.00	44.00	52.00	52.00
French	42.00	46.00	42.00	42.00	44.00
German	44.00	48.00	44.00	44.00	46.00
Greek	35.00	43.00	35.00	25.00	10.00
Indonesian	50.00	54.00	50.00	50.00	52.00
Italian	45.00	48.00	44.00	44.00	44.00
Japanese	42.00	46.00	41.00	42.00	45.00
Korean	39.00	41.00	38.00	39.00	42.00
Persian	42.00	42.00	41.00	42.00	37.00
Polish	47.00	47.00	37.00	47.00	47.00
Portuguese	40.00	50.00	36.00	41.00	45.00
Romanian	46.00	48.00	41.00	46.00	45.00
Russian	47.00	51.00	43.00	44.00	42.00
Spanish	43.00	45.00	38.00	43.00	45.00
Swedish	46.00	44.00	44.00	46.00	44.00
Turkish	43.00	45.00	40.00	43.00	38.00
Ukrainian	38.00	41.00	25.00	36.00	31.00
Vietnamese	36.00	40.00	35.00	36.00	38.00
Avg	44.05	47.05	40.50	43.32	42.45
<i>Mid-resource (11)</i>					
Bengali	42.00	46.00	41.00	41.00	35.00
Hebrew	40.00	38.00	40.00	37.00	31.00
Hindi	36.00	40.00	35.00	35.00	25.00
Kyrgyz	35.00	37.00	32.00	34.00	25.00
Lithuanian	42.00	44.00	39.00	41.00	42.00
Malay	41.00	44.00	40.00	41.00	42.00
Nepali	43.00	49.00	39.00	41.00	42.00
Serbian	41.00	43.00	37.00	41.00	42.00
Sinhala	37.00	39.00	37.00	36.00	37.00
Swahili	37.00	38.00	35.00	36.00	36.00

Continued on next page

Table 20 – continued

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
Telugu	41.00	43.00	41.00	40.00	39.00
Avg	39.55	41.91	37.82	38.45	36.00
<i>Low-resource (9)</i>					
Amharic	37.00	36.00	37.00	36.00	22.00
Chichewa	28.00	30.00	29.00	25.00	30.00
Filipino	49.00	52.00	48.00	48.00	51.00
Hausa	30.00	29.00	30.00	30.00	29.00
Igbo	31.00	32.00	32.00	31.00	31.00
Malagasy	36.00	34.00	33.00	36.00	34.00
Shona	30.00	33.00	29.00	29.00	31.00
Somali	34.00	31.00	34.00	35.00	33.00
Yoruba	33.00	38.00	33.00	32.00	34.00
Avg	34.22	35.00	33.89	33.56	32.78
Overall (42)	40.76	43.12	38.38	39.95	38.69

Table 21: Llama-3.1-8B-Instruct overall accuracy with varying context provided to MT2 on MCSQA: High-resource (8).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (8)</i>					
Chinese	60.00	60.00	52.00	60.00	60.00
Dutch	73.00	74.00	65.00	73.00	73.00
English	67.00	66.00	33.00	67.00	67.00
French	72.00	74.00	60.00	70.00	73.00
German	65.00	65.00	57.00	65.00	65.00
Japanese	72.00	73.00	61.00	72.00	73.00
Portuguese	73.00	72.00	49.00	73.00	69.00
Russian	35.00	37.00	26.00	34.00	29.00
Avg	64.62	65.12	50.38	64.25	63.62
Overall (8)	64.62	65.12	50.38	64.25	63.62

Table 22: Llama-3.1-8B-Instruct overall accuracy with varying context provided to MT2 on Belebele: High-resource (31), Mid-resource (42) and Low-resource (48).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (31)</i>					
Chinese (Simplified)	53.33	63.33	60.00	53.33	53.33
Chinese (Traditional)	63.33	70.00	63.33	63.33	63.33
Czech	76.67	76.67	66.67	76.67	76.67
Danish	76.67	76.67	76.67	76.67	76.67
Dutch	90.00	90.00	93.33	90.00	90.00
Egyptian Arabic	66.67	66.67	66.67	66.67	33.33
English	80.00	83.33	80.00	80.00	83.33
French	80.00	80.00	80.00	80.00	80.00
German	86.67	90.00	90.00	86.67	86.67
Greek	83.33	86.67	86.67	70.00	26.67
Hungarian	76.67	76.67	80.00	76.67	76.67
Indonesian	80.00	80.00	80.00	80.00	80.00
Italian	76.67	80.00	76.67	76.67	76.67
Japanese	63.33	63.33	63.33	63.33	63.33
Korean	76.67	76.67	76.67	76.67	76.67
Mesopotamian Arabic	66.67	63.33	66.67	66.67	53.33
Modern Standard Arabic	70.00	80.00	76.67	70.00	66.67
Modern Standard Arabic (Romanized)	20.00	20.00	20.00	20.00	20.00
Najdi Arabic	50.00	53.33	50.00	46.67	30.00
North Levantine Arabic	56.67	53.33	56.67	56.67	46.67
Polish	86.67	90.00	86.67	86.67	90.00

Continued on next page

Table 22 – continued

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
Portuguese	80.00	80.00	83.33	80.00	80.00
Romanian	70.00	73.33	70.00	70.00	73.33
Russian	66.67	66.67	73.33	66.67	40.00
Slovak	86.67	86.67	86.67	86.67	86.67
Spanish	93.33	93.33	93.33	93.33	86.67
Swedish	76.67	80.00	76.67	76.67	80.00
Turkish	66.67	66.67	70.00	66.67	63.33
Ukrainian	73.33	73.33	73.33	73.33	36.67
Vietnamese	83.33	86.67	86.67	83.33	83.33
Western Persian	76.67	76.67	76.67	76.67	66.67
Avg	72.69	74.30	73.76	72.15	66.02
<i>Mid-resource (42)</i>					
Afrikaans	86.67	86.67	83.33	86.67	86.67
Armenian	73.33	76.67	73.33	73.33	60.00
Basque	73.33	73.33	73.33	73.33	73.33
Bengali	53.33	53.33	53.33	53.33	40.00
Bengali (Romanized)	40.00	40.00	36.67	36.67	43.33
Bulgarian	80.00	80.00	80.00	80.00	66.67
Burmese	33.33	36.67	33.33	36.67	30.00
Catalan	76.67	76.67	76.67	76.67	76.67
Croatian	83.33	86.67	83.33	83.33	86.67
Eastern Panjabi	40.00	40.00	40.00	40.00	40.00
Finnish	66.67	70.00	66.67	70.00	66.67
Georgian	66.67	66.67	66.67	66.67	53.33
Gujarati	56.67	56.67	56.67	56.67	50.00
Hebrew	80.00	80.00	80.00	80.00	56.67
Hindi	66.67	63.33	66.67	63.33	33.33
Hindi (Romanized)	46.67	46.67	46.67	50.00	43.33
Icelandic	80.00	80.00	80.00	80.00	80.00
Kannada	60.00	63.33	60.00	60.00	40.00
Kazakh	60.00	60.00	60.00	60.00	40.00
Khmer	53.33	53.33	53.33	53.33	33.33
Kyrgyz	66.67	73.33	66.67	66.67	53.33
Lithuanian	66.67	70.00	66.67	66.67	66.67
Macedonian	60.00	70.00	60.00	60.00	60.00
Malayalam	46.67	50.00	46.67	46.67	50.00
Marathi	73.33	73.33	73.33	73.33	66.67
Nepali	40.00	40.00	40.00	40.00	26.67
North Azerbaijani	60.00	60.00	60.00	60.00	60.00
Northern Uzbek	70.00	73.33	70.00	70.00	63.33
Norwegian Bokmål	83.33	83.33	83.33	83.33	83.33
Serbian	76.67	80.00	76.67	76.67	80.00
Sinhala	46.67	50.00	46.67	46.67	50.00
Sinhala (Romanized)	30.00	33.33	30.00	30.00	30.00
Slovenian	70.00	76.67	70.00	70.00	76.67
Standard Malay	90.00	90.00	90.00	90.00	90.00
Swahili	56.67	66.67	56.67	56.67	60.00
Tagalog	53.33	56.67	53.33	53.33	56.67
Tamil	50.00	50.00	50.00	50.00	46.67
Telugu	40.00	40.00	40.00	40.00	36.67
Thai	60.00	66.67	60.00	60.00	60.00
Tosk Albanian	86.67	90.00	86.67	86.67	86.67
Urdu	90.00	90.00	90.00	90.00	90.00
Urdu (Romanized)	33.33	33.33	33.33	33.33	26.67
Avg	62.54	64.44	62.38	62.62	57.62
<i>Low-resource (48)</i>					
Amharic	33.33	33.33	33.33	26.67	0.00
Assamese	50.00	53.33	50.00	50.00	33.33
Bambara	26.67	23.33	26.67	23.33	26.67
Cebuano	60.00	56.67	60.00	60.00	56.67
Central Kurdish	73.33	70.00	73.33	73.33	76.67
Estonian	70.00	76.67	66.67	70.00	73.33
Ganda	33.33	26.67	33.33	33.33	26.67
Guarani	30.00	30.00	30.00	30.00	30.00

Continued on next page

Table 22 – continued

Language	C _{std}	C _{ctx}	q _t + a _e	q _e + a _e	r _e + a _e
Haitian Creole	63.33	63.33	63.33	63.33	63.33
Halh Mongolian	46.67	50.00	46.67	46.67	36.67
Hausa	43.33	50.00	43.33	43.33	46.67
Igbo	40.00	43.33	40.00	40.00	43.33
Ilocano	46.67	46.67	46.67	46.67	50.00
Javanese	66.67	66.67	66.67	66.67	66.67
Jingpho	20.00	23.33	20.00	20.00	23.33
Kabuverdianu	63.33	63.33	63.33	63.33	60.00
Kinyarwanda	23.33	23.33	23.33	23.33	20.00
Lao	30.00	36.67	30.00	30.00	40.00
Lingala	23.33	23.33	23.33	23.33	23.33
Luo	23.33	23.33	23.33	23.33	23.33
Maltese	66.67	70.00	66.67	66.67	70.00
Maori	60.00	60.00	60.00	60.00	60.00
Nepali (Romanized)	46.67	50.00	46.67	40.00	50.00
Nigerian Fulfulde	36.67	40.00	36.67	36.67	40.00
Northern Sotho	30.00	30.00	30.00	33.33	30.00
Nyanja	33.33	33.33	33.33	33.33	33.33
Odia	53.33	56.67	53.33	50.00	56.67
Plateau Malagasy	26.67	26.67	26.67	26.67	26.67
Shan	26.67	23.33	26.67	26.67	23.33
Shona	36.67	36.67	36.67	33.33	36.67
Sindhi	53.33	56.67	53.33	53.33	56.67
Somali	16.67	16.67	16.67	16.67	16.67
Southern Pashto	56.67	60.00	56.67	56.67	56.67
Southern Sotho	26.67	30.00	26.67	26.67	26.67
Standard Latvian	60.00	56.67	60.00	60.00	50.00
Standard Tibetan	36.67	33.33	36.67	33.33	33.33
Sundanese	53.33	53.33	53.33	53.33	53.33
Swati	46.67	46.67	46.67	46.67	50.00
Tajik	66.67	70.00	66.67	66.67	70.00
Tigrinya	23.33	26.67	23.33	20.00	23.33
Tsonga	30.00	33.33	30.00	30.00	33.33
Tswana	30.00	26.67	30.00	30.00	26.67
Waray	56.67	56.67	56.67	56.67	56.67
West Central Oromo	43.33	43.33	43.33	43.33	33.33
Wolof	36.67	33.33	36.67	36.67	33.33
Xhosa	23.33	23.33	23.33	23.33	23.33
Yoruba	26.67	20.00	26.67	23.33	20.00
Zulu	33.33	33.33	33.33	33.33	33.33
Avg	41.74	42.29	41.67	41.11	40.49
Overall (121)	56.89	58.18	57.08	56.53	52.98

Table 23: Llama-3.1-8B-Instruct overall accuracy with varying context provided to MT2 on Global-PIQA: High-resource (44), Mid-resource (42) and Low-resource (30).

Language	C _{std}	C _{ctx}	q _t + a _e	q _e + a _e	r _e + a _e
<i>High-resource (44)</i>					
Algerian Arabic	50.00	46.67	56.67	50.00	60.00
Burushaski	40.00	40.00	36.67	26.67	23.33
Cantonese	70.00	70.00	73.33	70.00	66.67
Chinese (Simplified)	56.67	60.00	53.33	53.33	53.33
Chinese (Traditional)	56.67	53.33	70.00	70.00	66.67
Czech	73.33	73.33	73.33	66.67	73.33
Dutch	70.00	70.00	70.00	66.67	70.00
Egyptian Arabic	60.00	56.67	66.67	33.33	60.00
English	73.33	73.33	73.33	76.67	73.33
French (Canada)	86.67	90.00	80.00	70.00	76.67
French (France)	73.33	73.33	73.33	70.00	76.67
German	66.67	73.33	76.67	73.33	76.67
Greek	60.00	63.33	73.33	6.67	10.00
Gulf Arabic	40.00	36.67	63.33	46.67	56.67

Continued on next page

Table 23 – continued

Language	C _{std}	C _{ctx}	q _t + a _e	q _e + a _e	r _e + a _e
Hungarian	83.33	83.33	83.33	83.33	83.33
Indonesian	83.33	86.67	80.00	76.67	83.33
Italian	76.67	76.67	76.67	70.00	76.67
Japanese	70.00	73.33	70.00	63.33	73.33
Korean	53.33	53.33	53.33	53.33	63.33
Mesopotamian Arabic	63.33	60.00	73.33	73.33	70.00
Modern Standard Arabic	63.33	63.33	60.00	43.33	60.00
Moroccan Arabic	50.00	50.00	63.33	63.33	63.33
Najdi Arabic	76.67	80.00	76.67	70.00	66.67
North Levantine Arabic (Jordan)	60.00	60.00	56.67	46.67	53.33
North Levantine Arabic (Lebanon)	73.33	73.33	76.67	70.00	66.67
North Levantine Arabic (Palestine)	56.67	56.67	60.00	26.67	53.33
North Levantine Arabic (Syria)	70.00	70.00	53.33	33.33	46.67
Polish	50.00	53.33	63.33	60.00	63.33
Portuguese (Brazil)	66.67	66.67	73.33	66.67	73.33
Portuguese (Portugal)	93.33	93.33	86.67	50.00	83.33
Romanian	90.00	93.33	90.00	90.00	90.00
Russian	80.00	83.33	66.67	26.67	53.33
Slovak	60.00	56.67	73.33	70.00	73.33
Slovak (SARI)	53.33	53.33	56.67	46.67	56.67
Spanish (Mexico)	76.67	76.67	80.00	60.00	83.33
Spanish (Peru)	90.00	90.00	93.33	76.67	93.33
Spanish (Spain)	66.67	63.33	80.00	66.67	80.00
Swedish	66.67	70.00	63.33	50.00	66.67
Ta'izzi-Adeni Arabic	60.00	63.33	56.67	56.67	53.33
Tunisian Arabic	63.33	63.33	53.33	50.00	50.00
Turkish	76.67	76.67	80.00	80.00	80.00
Ukrainian	80.00	80.00	80.00	16.67	56.67
Vietnamese	60.00	60.00	53.33	53.33	50.00
Western Persian	70.00	76.67	56.67	56.67	60.00
Avg	67.27	67.88	68.86	57.50	65.23
<i>Mid-resource (42)</i>					
Armenian	60.00	63.33	66.67	60.00	50.00
Belarusian	80.00	80.00	73.33	66.67	73.33
Bengali	80.00	83.33	76.67	76.67	63.33
Bengali (Romanized)	33.33	33.33	46.67	40.00	43.33
Bosnian	86.67	86.67	96.67	93.33	96.67
Bulgarian	86.67	86.67	83.33	53.33	73.33
Catalan	76.67	76.67	73.33	60.00	73.33
Croatian	83.33	83.33	83.33	76.67	83.33
Eastern Panjabi	63.33	63.33	70.00	63.33	60.00
Finnish	63.33	63.33	56.67	56.67	56.67
Galician	66.67	70.00	66.67	63.33	70.00
Georgian	56.67	56.67	53.33	50.00	53.33
Gujarati	90.00	90.00	86.67	86.67	80.00
Hebrew	60.00	63.33	66.67	30.00	56.67
Hindi	80.00	80.00	83.33	60.00	76.67
Icelandic	86.67	86.67	83.33	83.33	83.33
Kannada	76.67	83.33	73.33	76.67	53.33
Kazakh	50.00	50.00	46.67	40.00	36.67
Kyrgyz	60.00	60.00	63.33	60.00	36.67
Lithuanian	66.67	63.33	66.67	63.33	66.67
Macedonian	76.67	80.00	63.33	63.33	63.33
Malayalam	60.00	60.00	56.67	53.33	53.33
Marathi	90.00	90.00	86.67	76.67	83.33
Nepali	70.00	70.00	70.00	63.33	50.00
North Azerbaijani	76.67	76.67	66.67	63.33	66.67
Northern Uzbek	56.67	53.33	60.00	53.33	53.33
Norwegian Bokmål	80.00	80.00	73.33	76.67	73.33
Norwegian Nynorsk	83.33	86.67	66.67	63.33	66.67
Serbian (Cyrillic)	56.67	53.33	63.33	33.33	36.67
Serbian (Latin)	70.00	70.00	70.00	66.67	73.33
Sinhala	46.67	53.33	43.33	46.67	50.00
Slovenian	73.33	73.33	63.33	63.33	66.67
Slovenian (CERK)	43.33	53.33	53.33	43.33	56.67

Continued on next page

Table 23 – continued

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
Standard Malay	86.67	86.67	96.67	90.00	93.33
Swahili	76.67	76.67	73.33	66.67	70.00
Tagalog	76.67	76.67	80.00	80.00	80.00
Tamil	60.00	60.00	76.67	73.33	73.33
Telugu	90.00	90.00	76.67	66.67	73.33
Thai	83.33	83.33	76.67	76.67	80.00
Tosk Albanian	63.33	63.33	73.33	56.67	66.67
Urdu	86.67	86.67	80.00	76.67	76.67
Urdu (Romanized)	56.67	56.67	66.67	56.67	70.00
Avg	70.72	71.51	70.32	63.57	65.79
<i>Low-resource (30)</i>					
Amharic	60.00	56.67	60.00	46.67	33.33
Assamese	83.33	83.33	76.67	66.67	76.67
Bambara	43.33	43.33	63.33	40.00	53.33
Bhojpuri	56.67	63.33	63.33	60.00	60.00
Central Kurdish	56.67	56.67	56.67	56.67	60.00
Dhundari	80.00	80.00	66.67	60.00	66.67
Ekpeye	56.67	56.67	40.00	33.33	40.00
Estonian	53.33	50.00	53.33	56.67	53.33
Faroese	46.67	53.33	50.00	50.00	46.67
Hausa	46.67	46.67	46.67	50.00	46.67
Hawaiian	43.33	43.33	53.33	46.67	53.33
Idoma	40.00	40.00	70.00	63.33	60.00
Igbo	50.00	50.00	53.33	53.33	46.67
Isoko	50.00	53.33	46.67	46.67	46.67
Javanese	53.33	50.00	63.33	56.67	63.33
Kinyarwanda	63.33	66.67	63.33	50.00	60.00
Kumzari	53.33	53.33	53.33	53.33	56.67
Lingala	43.33	43.33	43.33	40.00	50.00
Luo	43.33	43.33	33.33	26.67	30.00
Marwari	73.33	73.33	83.33	83.33	76.67
Meitei (Bengali script)	66.67	70.00	46.67	36.67	46.67
Meitei (Meitei script)	43.33	46.67	36.67	30.00	30.00
Nagamese	43.33	40.00	60.00	53.33	60.00
Nigerian Pidgin	80.00	80.00	90.00	56.67	93.33
Sindhi	100.00	100.00	93.33	86.67	93.33
Sindhi (Devanagari)	63.33	66.67	76.67	70.00	53.33
Urhobo	46.67	43.33	53.33	46.67	46.67
Uyghur	46.67	43.33	53.33	56.67	43.33
Yoruba	43.33	43.33	56.67	56.67	50.00
Zulu	50.00	53.33	63.33	53.33	70.00
Avg	56.00	56.44	59.00	52.89	55.56
Overall (116)	65.60	66.24	66.84	58.51	62.93

Table 24: Llama-3.1-8B-Instruct overall accuracy with varying context provided to MT2 on MGSM: High-resource (7) and Mid-resource (4).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (7)</i>					
Chinese	38.00	50.00	38.00	31.00	38.00
English	36.00	56.00	38.00	36.00	44.00
French	46.00	56.00	43.00	40.00	34.00
German	46.00	54.00	40.00	42.00	52.00
Japanese	36.00	45.00	36.00	32.00	40.00
Russian	52.00	65.00	48.00	41.00	49.00
Spanish	49.00	50.00	43.00	40.00	39.00
Avg	43.29	53.71	40.86	37.43	42.29
<i>Mid-resource (4)</i>					
Bengali	29.00	35.00	29.00	20.00	28.00
Swahili	50.00	59.00	49.00	39.00	53.00

Continued on next page

Table 24 – continued

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
Telugu	30.00	39.00	31.00	24.00	26.00
Thai	30.00	36.00	31.00	26.00	35.00
Avg	34.75	42.25	35.00	27.25	35.50
Overall (11)	40.18	49.55	38.73	33.73	39.82

Table 25: Llama-3.1-8B-Instruct overall accuracy with varying context provided to MT2 on MMath: High-resource (9) and Mid-resource (1).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (9)</i>					
Arabic	18.00	21.00	16.00	16.00	19.00
Chinese	11.00	13.00	10.00	10.00	9.00
English	14.00	17.00	14.00	14.00	15.00
French	15.00	17.00	15.00	12.00	13.00
Japanese	12.00	15.00	13.00	9.00	11.00
Korean	9.00	11.00	8.00	9.00	10.00
Portuguese	14.00	20.00	14.00	13.00	17.00
Spanish	16.00	19.00	17.00	16.00	15.00
Vietnamese	15.00	18.00	12.00	14.00	14.00
Avg	13.78	16.78	13.22	12.56	13.67
<i>Mid-resource (1)</i>					
Thai	11.00	11.00	11.00	11.00	12.00
Avg	11.00	11.00	11.00	11.00	12.00
Overall (10)	13.50	16.20	13.00	12.40	13.50

Table 26: Llama-3.1-8B-Instruct chrF scores with varying context provided to MT2 on Aya: High-resource (4), Mid-resource (1) and Low-resource (1).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (4)</i>					
Arabic	8.67	13.63	12.76	10.69	11.28
Chinese	5.02	10.49	11.51	7.11	6.84
Portuguese	19.72	27.29	25.75	22.71	22.86
Turkish	14.33	20.94	18.47	16.44	19.70
Avg	11.93	18.09	17.12	14.24	15.17
<i>Mid-resource (1)</i>					
Telugu	7.93	11.09	10.27	7.79	7.90
Avg	7.93	11.09	10.27	7.79	7.90
<i>Low-resource (1)</i>					
Yoruba	1.66	2.11	2.10	1.76	1.80
Avg	1.66	2.11	2.10	1.76	1.80
Overall (6)	9.56	14.26	13.48	11.08	11.73

Table 27: Llama-3.1-8B-Instruct chrF scores with varying context provided to MT2 on BLENd: High-resource (9), Mid-resource (1) and Low-resource (4).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (9)</i>					
Arabic	11.82	12.82	12.73	11.12	9.82
Chinese	5.23	8.95	10.07	10.32	5.41
Greek	11.50	14.23	13.58	13.15	11.72

Continued on next page

Table 27 – continued

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
Indonesian	13.27	16.63	17.07	16.01	12.58
Korean (KP)	2.95	5.71	6.42	3.98	3.01
Korean (KR)	6.42	8.43	8.29	7.21	4.35
Persian	11.72	13.58	14.36	12.81	8.68
Spanish (ES)	14.97	18.01	14.99	13.99	13.76
Spanish (MX)	14.94	16.66	14.76	14.04	12.90
Avg	10.31	12.78	12.47	11.40	9.14
<i>Mid-resource (1)</i>					
Azerbaijani	10.20	15.07	14.05	12.34	10.30
Avg	10.20	15.07	14.05	12.34	10.30
<i>Low-resource (4)</i>					
Amharic	4.59	5.65	6.22	4.79	4.96
Assamese	9.08	11.79	12.41	9.61	8.56
Hausa	6.20	9.04	7.87	7.63	6.87
Javanese	9.74	13.45	13.49	11.32	9.65
Avg	7.40	9.98	10.00	8.34	7.51
Overall (14)	9.47	12.14	11.88	10.59	8.76

Table 28: Llama-3.1-8B-Instruct chrF scores with varying context provided to MT2 on Global-PIQA-OE: High-resource (44), Mid-resource (42) and Low-resource (30).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (44)</i>					
Algerian Arabic	10.56	9.92	10.32	10.83	9.34
Burushaski	2.66	4.32	7.24	3.77	3.66
Cantonese	2.56	3.94	3.33	3.03	2.63
Chinese (Simplified)	3.49	6.46	5.54	3.41	3.12
Chinese (Traditional)	3.84	8.71	6.93	4.27	3.50
Czech	11.08	13.65	12.50	11.97	13.66
Dutch	21.72	17.94	21.21	19.85	17.40
Egyptian Arabic	12.81	13.27	12.33	11.23	10.53
English	19.23	17.74	19.45	18.62	19.03
French (Canada)	18.74	21.85	19.28	20.01	17.34
French (France)	13.97	14.24	17.46	13.82	12.27
German	15.54	16.55	16.66	15.86	15.47
Greek	9.82	13.67	11.83	9.67	11.11
Gulf Arabic	11.01	10.23	10.42	11.30	9.02
Hungarian	16.63	17.92	16.78	16.99	15.69
Indonesian	21.30	18.78	21.13	20.74	16.73
Italian	12.81	21.37	19.86	18.17	20.84
Japanese	5.55	5.62	6.73	5.18	3.96
Korean	4.04	4.89	5.35	4.16	3.40
Mesopotamian Arabic	9.92	10.71	11.21	10.82	9.87
Modern Standard Arabic	12.88	12.10	13.56	12.13	10.53
Moroccan Arabic	10.54	10.61	10.87	10.50	9.13
Najdi Arabic	11.39	14.47	13.80	12.92	12.85
North Levantine Arabic (Jordan)	13.65	15.15	13.65	11.24	9.01
North Levantine Arabic (Lebanon)	7.90	10.36	10.00	6.30	3.79
North Levantine Arabic (Palestine)	7.28	9.64	10.20	8.39	5.25
North Levantine Arabic (Syria)	7.37	9.16	8.13	7.43	4.38
Polish	16.41	17.24	18.59	16.87	15.72
Portuguese (Brazil)	16.65	18.74	16.87	17.12	12.74
Portuguese (Portugal)	15.67	15.82	17.02	16.54	14.43
Romanian	16.10	16.51	17.25	16.70	15.31
Russian	15.59	16.62	19.93	17.33	17.18
Slovak	13.62	17.80	17.65	15.21	14.75
Slovak (SARI)	10.97	12.06	9.89	10.66	10.38
Spanish (Mexico)	16.55	17.06	20.36	16.77	15.15
Spanish (Peru)	18.86	17.86	19.70	18.48	16.25
Spanish (Spain)	18.43	19.28	19.34	18.70	18.39

Continued on next page

Table 28 – continued

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
Swedish	19.74	20.51	18.43	18.90	18.98
Ta'izzi-Adeni Arabic	10.25	12.54	11.37	11.70	11.11
Tunisian Arabic	11.46	10.20	11.40	10.49	9.00
Turkish	17.12	19.41	19.38	17.38	15.90
Ukrainian	14.13	14.46	14.66	14.20	14.71
Vietnamese	10.64	14.42	13.84	12.52	12.78
Western Persian	14.63	18.55	19.77	16.08	15.08
Avg	12.62	13.92	14.12	12.91	11.85
<i>Mid-resource (42)</i>					
Armenian	13.15	14.86	15.41	14.83	13.77
Belarusian	13.93	19.60	17.37	14.11	12.34
Bengali	11.25	13.61	14.05	12.04	10.86
Bengali (Romanized)	0.15	10.11	11.79	1.61	3.72
Bosnian	18.65	18.97	19.91	18.46	17.79
Bulgarian	16.58	19.13	18.57	17.88	16.92
Catalan	15.24	16.44	17.22	17.65	16.52
Croatian	13.69	14.34	12.95	13.88	11.37
Eastern Panjabi	12.34	15.27	15.56	13.35	13.12
Finnish	19.08	22.63	18.82	20.59	19.88
Galician	15.41	20.54	20.57	17.76	20.05
Georgian	13.82	16.76	16.33	13.69	13.39
Gujarati	13.36	15.95	16.52	13.72	14.23
Hebrew	12.72	15.07	13.73	12.49	14.13
Hindi	13.25	14.20	16.02	13.56	11.60
Icelandic	16.29	16.83	17.84	15.29	14.75
Kannada	16.99	20.65	18.95	18.06	17.60
Kazakh	14.95	20.10	20.14	16.31	15.94
Kyrgyz	14.99	16.85	16.50	13.41	13.85
Lithuanian	12.55	14.12	12.55	12.18	11.08
Macedonian	15.27	14.91	16.00	14.22	13.30
Malayalam	15.69	19.07	19.82	14.58	16.07
Marathi	10.20	15.49	15.87	10.37	9.22
Nepali	14.29	15.68	15.28	13.09	14.18
North Azerbaijani	14.08	13.72	14.39	13.85	11.87
Northern Uzbek	11.83	16.24	13.76	11.91	12.60
Norwegian Bokmål	17.86	17.34	18.78	17.88	15.69
Norwegian Nynorsk	16.63	17.48	17.62	16.87	16.32
Serbian (Cyrillic)	5.93	10.92	8.30	0.79	1.29
Serbian (Latin)	13.30	13.77	12.26	13.62	13.30
Sinhala	8.53	9.76	10.99	6.51	7.40
Slovenian	14.04	14.70	15.12	14.94	14.98
Slovenian (CERK)	9.77	9.96	10.46	10.82	8.71
Standard Malay	23.64	28.85	26.73	23.52	23.57
Swahili	14.80	12.76	14.53	12.98	11.71
Tagalog	21.37	24.72	21.71	22.54	20.99
Tamil	21.73	25.13	23.21	22.78	22.75
Telugu	17.07	20.48	19.89	16.75	15.56
Thai	12.84	15.19	16.04	12.03	12.51
Tosk Albanian	12.42	13.89	13.70	13.48	14.03
Urdu	12.89	15.78	15.75	14.65	12.88
Urdu (Romanized)	5.32	13.47	13.82	3.28	4.29
Avg	14.00	16.56	16.31	14.10	13.72
<i>Low-resource (30)</i>					
Amharic	4.04	7.72	7.96	4.68	4.38
Assamese	9.74	12.59	12.62	10.15	9.28
Bambara	5.32	10.29	11.16	4.49	4.31
Bhojpuri	11.72	12.82	12.11	10.24	10.95
Central Kurdish	10.00	16.01	14.73	11.66	11.44
Dhundari	10.69	11.36	11.58	10.86	10.70
Ekpeye	3.89	6.54	7.15	3.25	3.78
Estonian	14.39	18.64	15.89	15.66	16.98
Faroese	14.30	17.53	15.82	14.25	12.24
Hausa	11.76	13.13	11.87	10.05	10.54
Hawaiian	12.81	15.85	14.67	10.25	10.64

Continued on next page

Table 28 – continued

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
Idoma	3.89	7.71	7.88	3.90	4.26
Igbo	9.20	15.25	16.41	9.60	9.60
Isoko	4.08	6.66	9.67	2.93	2.63
Javanese	12.83	14.10	14.66	12.44	11.05
Kinyarwanda	8.80	12.86	10.38	8.07	8.13
Kumzari	1.19	1.51	2.48	2.06	1.87
Lingala	8.32	10.08	9.46	8.53	7.55
Luo	10.06	8.79	8.60	8.55	9.61
Marwari	9.68	11.20	12.50	11.36	9.60
Meitei (Bengali script)	9.19	10.91	13.47	4.98	7.31
Meitei (Meitei script)	0.06	14.47	15.72	0.12	0.17
Nagamese	7.55	11.15	10.04	6.19	7.42
Nigerian Pidgin	16.14	21.15	18.70	18.12	18.36
Sindhi	11.69	13.76	14.91	10.83	10.60
Sindhi (Devanagari)	2.20	10.11	9.99	5.59	6.66
Urhobo	5.18	10.42	16.89	5.19	5.08
Uyghur	6.64	11.13	10.13	7.37	7.07
Yoruba	4.15	7.94	8.92	3.72	3.87
Zulu	8.79	10.01	12.67	8.80	8.18
Avg	8.28	11.72	11.97	8.13	8.14
Overall (116)	11.99	14.31	14.35	12.11	11.57

Table 29: Llama-3.1-8B-Instruct chrF scores with varying context provided to MT2 on MKQA: High-resource (19) and Mid-resource (6).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (19)</i>					
Arabic	11.73	12.47	14.15	12.98	11.02
Chinese (HK)	4.72	8.07	7.92	7.46	6.01
Chinese (Simplified)	7.91	10.88	9.76	10.22	8.42
Chinese (Traditional)	5.24	7.68	7.53	7.24	6.16
Danish	17.48	17.57	20.17	18.71	16.25
Dutch	15.01	15.20	18.77	16.89	13.56
French	16.79	18.38	20.40	18.46	16.28
German	16.41	17.44	20.90	20.15	15.73
Hungarian	13.82	15.36	16.33	15.74	13.59
Italian	16.91	17.25	19.39	18.91	16.16
Japanese	14.23	15.90	18.50	16.19	12.76
Korean	6.71	7.07	8.14	7.29	6.06
Polish	16.25	18.12	20.37	19.32	16.39
Portuguese	15.99	17.07	19.39	17.77	14.57
Russian	13.93	14.58	16.69	15.46	12.67
Spanish	16.39	16.93	19.80	17.73	15.21
Swedish	17.31	17.64	21.04	19.40	16.15
Turkish	14.33	13.32	17.47	15.49	11.30
Vietnamese	13.79	14.45	16.62	15.77	11.42
Avg	13.42	14.49	16.49	15.33	12.62
<i>Mid-resource (6)</i>					
Finnish	16.24	18.57	21.14	19.35	16.44
Hebrew	13.86	17.15	17.20	16.60	15.30
Khmer	3.50	4.46	4.25	3.76	3.68
Malay	16.02	15.31	18.72	17.58	14.21
Norwegian	16.47	17.28	19.73	18.07	15.64
Thai	13.04	17.29	17.72	15.99	14.24
Avg	13.19	15.01	16.46	15.22	13.25
Overall (25)	13.36	14.62	16.48	15.30	12.77

G.2 Mistral-7B-Instruct-v0.3

Dataset	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>Open-ended generation (chrF)</i>					
Aya	9.71	11.99	11.73	10.02	10.72
BLEnD	8.27	9.05	10.23	8.99	8.09
Global-PIQA-OE	12.90	14.58	15.37	12.69	12.80
MKQA	11.29	12.09	13.19	11.96	11.13
<i>Multiple-choice (Acc.)</i>					
Global-MMLU	26.29	32.07	30.31	25.36	31.07
Belebele	24.60	25.92	26.75	23.47	26.91
Global-PIQA	51.72	59.11	58.68	37.64	52.39
MCSQA	57.12	59.12	62.38	50.75	56.50
<i>Math (Acc.)</i>					
MGSM	35.27	34.00	27.00	29.45	26.82
MMath	5.40	5.70	4.40	5.70	4.90

Table 30: Mistral-7B-Instruct-v0.3 overall accuracy and chrF scores with varying context provided to MT2 on all datasets.

Table 31: Mistral-7B-Instruct-v0.3 accuracy with varying context provided to MT2 on Global-MMLU: High-resource (22), Mid-resource (11) and Low-resource (9).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (22)</i>					
Arabic	23.00	28.00	27.00	21.00	39.00
Chinese	27.00	31.00	28.00	26.00	32.00
Czech	33.00	39.00	39.00	37.00	39.00
Dutch	34.00	42.00	42.00	31.00	39.00
English	32.00	40.00	43.00	30.00	39.00
French	33.00	42.00	42.00	28.00	38.00
German	31.00	35.00	37.00	35.00	36.00
Greek	25.00	29.00	27.00	22.00	29.00
Indonesian	27.00	36.00	31.00	28.00	29.00
Italian	31.00	45.00	39.00	29.00	42.00
Japanese	30.00	42.00	33.00	26.00	35.00
Korean	28.00	31.00	27.00	29.00	28.00
Persian	36.00	38.00	39.00	31.00	40.00
Polish	38.00	43.00	42.00	36.00	43.00
Portuguese	26.00	36.00	34.00	25.00	32.00
Romanian	34.00	37.00	36.00	33.00	39.00
Russian	30.00	39.00	35.00	27.00	29.00
Spanish	34.00	41.00	40.00	29.00	42.00
Swedish	26.00	37.00	29.00	26.00	34.00
Turkish	32.00	37.00	35.00	29.00	34.00
Ukrainian	34.00	43.00	36.00	24.00	39.00
Vietnamese	27.00	29.00	31.00	27.00	32.00
Avg	30.50	37.27	35.09	28.59	35.86
<i>Mid-resource (11)</i>					
Bengali	20.00	29.00	26.00	26.00	27.00
Hebrew	28.00	32.00	30.00	23.00	33.00
Hindi	25.00	32.00	29.00	23.00	31.00
Kyrgyz	19.00	21.00	23.00	23.00	29.00
Lithuanian	23.00	26.00	27.00	21.00	35.00
Malay	31.00	38.00	31.00	34.00	41.00
Nepali	21.00	26.00	20.00	22.00	25.00
Serbian	28.00	35.00	29.00	21.00	12.00
Sinhala	20.00	23.00	22.00	17.00	24.00
Swahili	22.00	29.00	27.00	24.00	28.00
Telugu	14.00	21.00	20.00	19.00	22.00
Avg	22.82	28.36	25.82	23.00	27.91
<i>Low-resource (9)</i>					
Amharic	29.00	31.00	32.00	24.00	26.00

Continued on next page

Table 31 – continued

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
Chichewa	15.00	15.00	16.00	14.00	23.00
Filipino	28.00	34.00	32.00	31.00	31.00
Hausa	13.00	22.00	19.00	19.00	20.00
Igbo	15.00	20.00	17.00	15.00	18.00
Malagasy	25.00	28.00	31.00	22.00	27.00
Shona	22.00	25.00	22.00	22.00	25.00
Somali	19.00	21.00	24.00	14.00	20.00
Yoruba	16.00	19.00	24.00	22.00	19.00
Avg	20.22	23.89	24.11	20.33	23.22
Overall (42)	26.29	32.07	30.31	25.36	31.07

Table 32: Mistral-7B-Instruct-v0.3 overall accuracy with varying context provided to MT2 on MCSQA: High-resource (8).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (8)</i>					
Chinese	55.00	59.00	61.00	50.00	56.00
Dutch	69.00	75.00	75.00	55.00	69.00
English	65.00	69.00	74.00	66.00	70.00
French	54.00	58.00	59.00	49.00	52.00
German	58.00	61.00	65.00	53.00	61.00
Japanese	66.00	64.00	67.00	57.00	55.00
Portuguese	57.00	56.00	60.00	48.00	58.00
Russian	33.00	31.00	38.00	28.00	31.00
Avg	57.12	59.12	62.38	50.75	56.50
Overall (8)	57.12	59.12	62.38	50.75	56.50

Table 33: Mistral-7B-Instruct-v0.3 accuracy with varying context provided to MT2 on Belebele: High-resource (31), Mid-resource (42) and Low-resource (48).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (31)</i>					
Chinese (Simplified)	30.00	26.67	26.67	30.00	20.00
Chinese (Traditional)	33.33	40.00	36.67	26.67	33.33
Czech	20.00	16.67	20.00	23.33	23.33
Danish	23.33	26.67	30.00	23.33	23.33
Dutch	26.67	46.67	43.33	30.00	33.33
Egyptian Arabic	33.33	23.33	33.33	20.00	33.33
English	56.67	56.67	70.00	60.00	63.33
French	33.33	30.00	43.33	36.67	30.00
German	26.67	36.67	43.33	33.33	26.67
Greek	20.00	26.67	23.33	23.33	20.00
Hungarian	16.67	23.33	26.67	20.00	16.67
Indonesian	20.00	26.67	30.00	10.00	20.00
Italian	40.00	56.67	53.33	40.00	36.67
Japanese	36.67	36.67	30.00	33.33	36.67
Korean	16.67	20.00	26.67	13.33	20.00
Mesopotamian Arabic	20.00	23.33	16.67	16.67	26.67
Modern Standard Arabic	50.00	33.33	50.00	23.33	43.33
Modern Standard Arabic (Romanized)	23.33	16.67	23.33	26.67	33.33
Najdi Arabic	20.00	20.00	23.33	36.67	30.00
North Levantine Arabic	20.00	23.33	23.33	20.00	20.00
Polish	26.67	30.00	30.00	33.33	33.33
Portuguese	23.33	33.33	40.00	30.00	26.67
Romanian	16.67	23.33	23.33	26.67	30.00
Russian	33.33	30.00	46.67	30.00	23.33
Slovak	16.67	20.00	20.00	13.33	13.33
Spanish	40.00	40.00	50.00	30.00	40.00

Continued on next page

Table 33 – continued

Language	C _{std}	C _{ctx}	q _t + a _e	q _e + a _e	r _e + a _e
Swedish	26.67	26.67	36.67	26.67	36.67
Turkish	36.67	36.67	40.00	36.67	36.67
Ukrainian	43.33	50.00	53.33	36.67	43.33
Vietnamese	26.67	30.00	26.67	23.33	20.00
Western Persian	43.33	43.33	46.67	40.00	46.67
Avg	29.03	31.40	35.05	28.17	30.32
<i>Mid-resource (42)</i>					
Afrikaans	30.00	30.00	50.00	23.33	30.00
Armenian	16.67	16.67	13.33	13.33	23.33
Basque	30.00	26.67	30.00	26.67	33.33
Bengali	30.00	26.67	26.67	26.67	26.67
Bengali (Romanized)	23.33	20.00	16.67	30.00	26.67
Bulgarian	26.67	26.67	30.00	26.67	23.33
Burmese	23.33	23.33	16.67	23.33	30.00
Catalan	20.00	23.33	23.33	26.67	26.67
Croatian	30.00	33.33	43.33	16.67	33.33
Eastern Panjabi	20.00	16.67	20.00	13.33	23.33
Finnish	26.67	23.33	36.67	16.67	33.33
Georgian	16.67	13.33	23.33	26.67	20.00
Gujarati	26.67	26.67	23.33	16.67	26.67
Hebrew	33.33	43.33	36.67	40.00	36.67
Hindi	30.00	40.00	33.33	30.00	36.67
Hindi (Romanized)	20.00	16.67	16.67	23.33	26.67
Icelandic	23.33	36.67	26.67	36.67	23.33
Kannada	10.00	16.67	13.33	20.00	13.33
Kazakh	10.00	13.33	10.00	10.00	13.33
Khmer	13.33	26.67	16.67	16.67	30.00
Kyrgyz	16.67	20.00	16.67	16.67	20.00
Lithuanian	33.33	40.00	33.33	40.00	33.33
Macedonian	40.00	36.67	40.00	30.00	33.33
Malayalam	10.00	13.33	3.33	13.33	13.33
Marathi	26.67	30.00	26.67	30.00	36.67
Nepali	13.33	20.00	13.33	10.00	10.00
North Azerbaijani	43.33	43.33	40.00	40.00	43.33
Northern Uzbek	36.67	40.00	36.67	33.33	40.00
Norwegian Bokmål	30.00	13.33	36.67	26.67	23.33
Serbian	36.67	36.67	36.67	30.00	13.33
Sinhala	20.00	16.67	23.33	10.00	20.00
Sinhala (Romanized)	20.00	23.33	20.00	20.00	23.33
Slovenian	26.67	23.33	36.67	16.67	13.33
Standard Malay	30.00	33.33	40.00	26.67	40.00
Swahili	13.33	16.67	13.33	13.33	16.67
Tagalog	26.67	23.33	33.33	23.33	30.00
Tamil	20.00	20.00	20.00	13.33	26.67
Telugu	26.67	20.00	26.67	26.67	33.33
Thai	23.33	23.33	30.00	20.00	16.67
Tosk Albanian	26.67	26.67	26.67	30.00	33.33
Urdu	46.67	53.33	50.00	46.67	50.00
Urdu (Romanized)	36.67	26.67	30.00	36.67	46.67
Avg	25.32	26.19	27.14	24.21	27.46
<i>Low-resource (48)</i>					
Amharic	30.00	16.67	16.67	16.67	23.33
Assamese	33.33	43.33	33.33	23.33	36.67
Bambara	13.33	13.33	13.33	13.33	20.00
Cebuano	36.67	36.67	36.67	33.33	36.67
Central Kurdish	30.00	16.67	26.67	23.33	30.00
Estonian	23.33	26.67	26.67	26.67	30.00
Ganda	3.33	6.67	3.33	6.67	6.67
Guarani	20.00	20.00	20.00	20.00	16.67
Haitian Creole	23.33	26.67	23.33	16.67	26.67
Halh Mongolian	30.00	30.00	30.00	23.33	33.33
Hausa	23.33	23.33	23.33	20.00	26.67
Igbo	13.33	16.67	13.33	6.67	16.67
Ilocano	33.33	36.67	33.33	33.33	30.00

Continued on next page

Table 33 – continued

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
Javanese	16.67	16.67	20.00	16.67	16.67
Jingpho	10.00	10.00	10.00	23.33	20.00
Kabuverdianu	20.00	23.33	20.00	20.00	23.33
Kinyarwanda	20.00	16.67	23.33	26.67	26.67
Lao	26.67	33.33	20.00	23.33	36.67
Lingala	13.33	13.33	16.67	20.00	13.33
Luo	16.67	20.00	16.67	16.67	33.33
Maltese	30.00	36.67	33.33	16.67	33.33
Maori	26.67	26.67	26.67	30.00	26.67
Nepali (Romanized)	20.00	20.00	20.00	30.00	10.00
Nigerian Fulfulde	23.33	23.33	26.67	10.00	20.00
Northern Sotho	23.33	23.33	23.33	23.33	30.00
Nyanja	16.67	23.33	20.00	16.67	26.67
Odia	20.00	16.67	10.00	30.00	26.67
Plateau Malagasy	13.33	13.33	13.33	13.33	16.67
Shan	23.33	16.67	20.00	23.33	23.33
Shona	20.00	16.67	23.33	16.67	20.00
Sindhi	23.33	26.67	20.00	23.33	26.67
Somali	10.00	10.00	10.00	10.00	13.33
Southern Pashto	16.67	16.67	16.67	16.67	16.67
Southern Sotho	20.00	23.33	23.33	23.33	26.67
Standard Latvian	30.00	23.33	30.00	23.33	33.33
Standard Tibetan	13.33	13.33	10.00	10.00	16.67
Sundanese	23.33	30.00	23.33	23.33	20.00
Swati	26.67	26.67	30.00	26.67	26.67
Tajik	20.00	23.33	20.00	13.33	20.00
Tigrinya	6.67	13.33	6.67	13.33	23.33
Tsonga	26.67	30.00	20.00	13.33	26.67
Tswana	23.33	23.33	23.33	23.33	23.33
Waray	30.00	30.00	30.00	26.67	40.00
West Central Oromo	6.67	10.00	10.00	3.33	10.00
Wolof	33.33	33.33	33.33	36.67	36.67
Xhosa	13.33	23.33	13.33	13.33	26.67
Yoruba	16.67	23.33	26.67	13.33	16.67
Zulu	20.00	20.00	20.00	16.67	23.33
Avg	21.11	22.15	21.04	19.79	24.24
Overall (121)	24.60	25.92	26.75	23.47	26.91

Table 34: Mistral-7B-Instruct-v0.3 accuracy with varying context provided to MT2 on Global-PIQA: High-resource (44), Mid-resource (42) and Low-resource (30).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (44)</i>					
Algerian Arabic	40.00	60.00	63.33	16.67	46.67
Burushaski	26.67	40.00	36.67	16.67	36.67
Cantonese	60.00	66.67	66.67	50.00	43.33
Chinese (Simplified)	56.67	70.00	63.33	63.33	63.33
Chinese (Traditional)	40.00	53.33	46.67	40.00	43.33
Czech	56.67	70.00	60.00	50.00	63.33
Dutch	60.00	66.67	63.33	40.00	60.00
Egyptian Arabic	40.00	50.00	46.67	13.33	40.00
English	63.33	66.67	66.67	53.33	70.00
French (Canada)	76.67	80.00	80.00	63.33	80.00
French (France)	66.67	66.67	66.67	33.33	63.33
German	46.67	50.00	50.00	50.00	43.33
Greek	30.00	40.00	46.67	20.00	23.33
Gulf Arabic	63.33	70.00	66.67	40.00	63.33
Hungarian	83.33	86.67	83.33	60.00	53.33
Indonesian	76.67	83.33	76.67	53.33	76.67
Italian	60.00	70.00	70.00	50.00	66.67
Japanese	60.00	66.67	66.67	53.33	60.00
Korean	46.67	56.67	50.00	46.67	56.67

Continued on next page

Table 34 – continued

Language	C _{std}	C _{ctx}	q _t + a _e	q _e + a _e	r _e + a _e
Mesopotamian Arabic	53.33	63.33	60.00	26.67	53.33
Modern Standard Arabic	53.33	46.67	60.00	23.33	50.00
Moroccan Arabic	40.00	56.67	53.33	26.67	53.33
Najdi Arabic	46.67	53.33	50.00	30.00	50.00
North Levantine Arabic (Jordan)	66.67	70.00	70.00	13.33	60.00
North Levantine Arabic (Lebanon)	50.00	70.00	63.33	20.00	60.00
North Levantine Arabic (Palestine)	46.67	63.33	50.00	10.00	46.67
North Levantine Arabic (Syria)	50.00	60.00	63.33	20.00	46.67
Polish	46.67	53.33	66.67	33.33	60.00
Portuguese (Brazil)	50.00	70.00	56.67	40.00	60.00
Portuguese (Portugal)	73.33	80.00	80.00	50.00	70.00
Romanian	86.67	93.33	90.00	73.33	90.00
Russian	60.00	66.67	66.67	30.00	36.67
Slovak	53.33	60.00	60.00	50.00	50.00
Slovak (SARI)	63.33	63.33	63.33	36.67	56.67
Spanish (Mexico)	76.67	90.00	90.00	56.67	90.00
Spanish (Peru)	83.33	90.00	90.00	60.00	90.00
Spanish (Spain)	66.67	70.00	76.67	50.00	70.00
Swedish	60.00	66.67	66.67	46.67	50.00
Ta'izzi-Adeni Arabic	63.33	73.33	63.33	10.00	56.67
Tunisian Arabic	56.67	56.67	56.67	23.33	56.67
Turkish	60.00	73.33	63.33	43.33	63.33
Ukrainian	56.67	73.33	70.00	43.33	60.00
Vietnamese	46.67	53.33	56.67	33.33	53.33
Western Persian	46.67	50.00	53.33	40.00	50.00
Avg	57.05	65.45	63.86	38.71	57.65
<i>Mid-resource (42)</i>					
Armenian	10.00	20.00	23.33	23.33	23.33
Belarusian	53.33	63.33	60.00	30.00	53.33
Bengali	30.00	43.33	43.33	33.33	33.33
Bengali (Romanized)	30.00	36.67	40.00	30.00	40.00
Bosnian	90.00	93.33	93.33	36.67	23.33
Bulgarian	56.67	73.33	76.67	50.00	40.00
Catalan	66.67	73.33	73.33	46.67	63.33
Croatian	66.67	76.67	70.00	50.00	43.33
Eastern Panjabi	43.33	56.67	56.67	50.00	53.33
Finnish	56.67	60.00	53.33	43.33	56.67
Galician	56.67	63.33	73.33	46.67	66.67
Georgian	60.00	56.67	60.00	30.00	46.67
Gujarati	40.00	46.67	53.33	36.67	43.33
Hebrew	46.67	53.33	50.00	40.00	60.00
Hindi	63.33	63.33	66.67	36.67	60.00
Icelandic	56.67	66.67	66.67	50.00	63.33
Kannada	36.67	50.00	46.67	33.33	40.00
Kazakh	23.33	30.00	33.33	20.00	33.33
Kyrgyz	63.33	63.33	63.33	36.67	60.00
Lithuanian	60.00	56.67	60.00	43.33	56.67
Macedonian	60.00	73.33	70.00	50.00	63.33
Malayalam	43.33	46.67	46.67	30.00	40.00
Marathi	60.00	66.67	63.33	43.33	53.33
Nepali	60.00	63.33	70.00	50.00	60.00
North Azerbaijani	36.67	46.67	40.00	33.33	50.00
Northern Uzbek	46.67	60.00	60.00	43.33	60.00
Norwegian Bokmål	76.67	76.67	76.67	60.00	76.67
Norwegian Nynorsk	43.33	46.67	56.67	36.67	50.00
Serbian (Cyrillic)	70.00	53.33	70.00	20.00	3.33
Serbian (Latin)	53.33	53.33	56.67	43.33	53.33
Sinhala	36.67	40.00	40.00	26.67	43.33
Slovenian	53.33	60.00	53.33	36.67	30.00
Slovenian (CERK)	36.67	40.00	56.67	36.67	33.33
Standard Malay	70.00	73.33	76.67	36.67	70.00
Swahili	60.00	63.33	70.00	36.67	56.67
Tagalog	60.00	50.00	60.00	33.33	50.00
Tamil	33.33	33.33	46.67	30.00	40.00
Telugu	40.00	36.67	40.00	23.33	40.00

Continued on next page

Table 34 – continued

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
Thai	46.67	60.00	70.00	40.00	70.00
Tosk Albanian	60.00	60.00	56.67	46.67	63.33
Urdu	60.00	73.33	56.67	43.33	70.00
Urdu (Romanized)	56.67	60.00	53.33	33.33	60.00
Avg	51.75	56.75	58.41	38.10	49.92
<i>Low-resource (30)</i>					
Amharic	40.00	53.33	50.00	23.33	23.33
Assamese	33.33	40.00	40.00	36.67	36.67
Bambara	36.67	50.00	50.00	26.67	40.00
Bhojpuri	46.67	53.33	63.33	43.33	53.33
Central Kurdish	33.33	43.33	36.67	33.33	46.67
Dhundari	50.00	63.33	53.33	43.33	60.00
Ekpeye	40.00	43.33	50.00	30.00	50.00
Estonian	36.67	40.00	40.00	26.67	36.67
Faroese	43.33	70.00	60.00	43.33	53.33
Hausa	46.67	56.67	56.67	36.67	46.67
Hawaiian	36.67	53.33	46.67	33.33	56.67
Idoma	30.00	50.00	43.33	50.00	50.00
Igbo	33.33	50.00	40.00	23.33	40.00
Isoko	50.00	53.33	53.33	36.67	53.33
Javanese	53.33	63.33	63.33	30.00	60.00
Kinyarwanda	43.33	50.00	40.00	30.00	43.33
Kumzari	36.67	56.67	46.67	46.67	46.67
Lingala	33.33	43.33	40.00	33.33	46.67
Luo	36.67	50.00	53.33	36.67	43.33
Marwari	60.00	60.00	73.33	40.00	63.33
Meitei (Bengali script)	40.00	50.00	43.33	26.67	36.67
Meitei (Meitei script)	50.00	53.33	53.33	50.00	46.67
Nagamese	43.33	66.67	56.67	40.00	43.33
Nigerian Pidgin	66.67	73.33	73.33	46.67	70.00
Sindhi	66.67	60.00	66.67	56.67	60.00
Sindhi (Devanagari)	56.67	43.33	46.67	23.33	53.33
Urhobo	36.67	50.00	43.33	26.67	40.00
Uyghur	50.00	53.33	53.33	33.33	53.33
Yoruba	40.00	46.67	60.00	26.67	50.00
Zulu	46.67	53.33	46.67	30.00	40.00
Avg	43.89	53.11	51.44	35.44	48.11
Overall (116)	51.72	59.11	58.68	37.64	52.39

Table 35: Mistral-7B-Instruct-v0.3 accuracy with varying context provided to MT2 on MGSM: High-resource (7) and Mid-resource (4).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (7)</i>					
Chinese	37.00	37.00	20.00	30.00	21.00
English	53.00	53.00	53.00	45.00	51.00
French	49.00	48.00	40.00	43.00	37.00
German	46.00	41.00	39.00	40.00	35.00
Japanese	34.00	34.00	19.00	31.00	25.00
Russian	51.00	47.00	41.00	42.00	39.00
Spanish	46.00	45.00	34.00	40.00	41.00
Avg	45.14	43.57	35.14	38.71	35.57
<i>Mid-resource (4)</i>					
Bengali	24.00	23.00	20.00	17.00	17.00
Swahili	12.00	11.00	6.00	7.00	11.00
Telugu	7.00	6.00	3.00	3.00	2.00
Thai	29.00	29.00	22.00	26.00	16.00
Avg	18.00	17.25	12.75	13.25	11.50
Overall (11)	35.27	34.00	27.00	29.45	26.82

Table 36: Mistral-7B-Instruct-v0.3 accuracy with varying context provided to MT2 on MMath: High-resource (9) and Mid-resource (1).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (9)</i>					
Arabic	7.00	7.00	5.00	8.00	7.00
Chinese	7.00	7.00	5.00	7.00	4.00
English	7.00	7.00	6.00	7.00	7.00
French	8.00	10.00	9.00	10.00	9.00
Japanese	3.00	3.00	4.00	3.00	3.00
Korean	4.00	3.00	3.00	4.00	3.00
Portuguese	5.00	6.00	3.00	4.00	6.00
Spanish	4.00	4.00	2.00	6.00	4.00
Vietnamese	4.00	5.00	3.00	3.00	4.00
Avg	5.44	5.78	4.44	5.78	5.22
<i>Mid-resource (1)</i>					
Thai	5.00	5.00	4.00	5.00	2.00
Avg	5.00	5.00	4.00	5.00	2.00
Overall (10)	5.40	5.70	4.40	5.70	4.90

Table 37: Mistral-7B-Instruct-v0.3 chrF scores with varying context provided to MT2 on Aya: High-resource (4), Mid-resource (1) and Low-resource (1).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (4)</i>					
Arabic	9.25	11.21	10.50	9.18	9.57
Chinese	4.87	8.35	8.76	5.97	6.05
Portuguese	22.52	26.79	26.11	22.77	23.72
Turkish	15.41	18.02	16.13	15.75	18.21
Avg	13.01	16.09	15.38	13.42	14.39
<i>Mid-resource (1)</i>					
Telugu	3.36	5.09	6.28	3.94	4.19
Avg	3.36	5.09	6.28	3.94	4.19
<i>Low-resource (1)</i>					
Yoruba	2.86	2.50	2.61	2.52	2.59
Avg	2.86	2.50	2.61	2.52	2.59
Overall (6)	9.71	11.99	11.73	10.02	10.72

Table 38: Mistral-7B-Instruct-v0.3 chrF scores with varying context provided to MT2 on BLEnD: High-resource (9), Mid-resource (1) and Low-resource (4).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (9)</i>					
Arabic	10.47	10.80	12.00	11.25	10.08
Chinese	6.03	7.53	10.30	9.49	7.58
Greek	10.65	11.28	11.84	10.85	10.35
Indonesian	13.16	13.18	15.43	14.40	13.38
Korean (KP)	1.85	3.64	3.90	3.09	2.43
Korean (KR)	4.43	5.33	6.98	5.34	4.61
Persian	9.16	9.15	10.10	9.03	8.37
Spanish (ES)	15.81	16.52	19.11	16.18	15.27
Spanish (MX)	15.29	16.89	19.27	15.64	14.21
Avg	9.65	10.48	12.10	10.59	9.59
<i>Mid-resource (1)</i>					
Azerbaijani	8.81	9.31	9.87	8.83	7.57

Continued on next page

Table 38 – continued

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
Avg	8.81	9.31	9.87	8.83	7.57
<i>Low-resource (4)</i>					
Amharic	1.39	2.51	2.76	1.60	1.43
Assamese	6.16	7.74	8.33	6.67	6.02
Hausa	3.66	3.54	4.01	3.52	3.17
Javanese	8.95	9.24	9.30	10.02	8.79
Avg	5.04	5.76	6.10	5.45	4.85
Overall (14)	8.27	9.05	10.23	8.99	8.09

Table 39: Mistral-7B-Instruct-v0.3 chrF scores with varying context provided to MT2 on Global-PIQA-OE: High-resource (44), Mid-resource (42) and Low-resource (30).

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (44)</i>					
Algerian Arabic	12.25	12.00	13.41	12.57	11.46
Burushaski	2.92	5.28	5.35	2.75	2.86
Cantonese	1.79	3.45	3.41	2.58	2.61
Chinese (Simplified)	2.70	6.12	6.50	3.75	3.71
Chinese (Traditional)	2.87	7.00	7.41	3.94	3.99
Czech	14.27	15.87	15.35	13.90	15.54
Dutch	23.82	22.12	25.44	23.86	21.41
Egyptian Arabic	14.01	15.06	14.91	14.69	14.61
English	21.18	24.93	25.55	23.76	23.98
French (Canada)	24.10	24.52	24.42	24.99	24.30
French (France)	20.78	21.38	21.88	20.86	20.35
German	19.28	20.31	20.29	19.50	19.58
Greek	10.08	11.00	12.09	11.11	11.11
Gulf Arabic	11.47	12.41	12.84	11.38	11.74
Hungarian	21.76	21.96	22.07	21.04	20.20
Indonesian	22.96	22.88	24.97	22.77	21.88
Italian	18.80	22.88	17.38	16.28	20.37
Japanese	4.11	6.27	5.95	3.97	4.91
Korean	3.34	5.02	6.05	4.60	4.55
Mesopotamian Arabic	12.67	14.31	13.74	12.28	13.07
Modern Standard Arabic	12.31	12.29	13.70	11.32	11.68
Moroccan Arabic	12.14	12.48	13.26	12.28	12.02
Najdi Arabic	15.86	17.24	16.69	16.15	15.77
North Levantine Arabic (Jordan)	15.56	17.53	16.92	15.81	16.19
North Levantine Arabic (Lebanon)	12.53	12.77	14.16	12.15	11.66
North Levantine Arabic (Palestine)	11.06	11.55	12.24	11.30	10.90
North Levantine Arabic (Syria)	11.82	12.83	12.14	11.39	11.47
Polish	19.63	20.28	20.14	18.85	18.70
Portuguese (Brazil)	18.81	18.49	21.39	19.34	19.19
Portuguese (Portugal)	18.98	20.28	21.01	18.46	18.92
Romanian	19.73	20.20	20.73	19.21	18.69
Russian	20.71	21.98	22.30	20.00	21.24
Slovak	15.81	17.08	18.13	17.60	17.72
Slovak (SARI)	11.03	11.82	12.26	10.76	11.70
Spanish (Mexico)	25.11	26.49	28.63	24.41	24.23
Spanish (Peru)	23.18	23.79	23.99	23.24	23.18
Spanish (Spain)	22.80	23.87	23.97	22.97	22.49
Swedish	21.71	24.33	21.94	21.63	21.67
Ta'izzi-Adeni Arabic	13.65	16.13	15.54	13.55	14.67
Tunisian Arabic	12.79	13.66	13.80	13.57	13.19
Turkish	18.94	19.94	20.82	18.71	19.59
Ukrainian	16.47	17.49	17.78	15.35	17.58
Vietnamese	12.82	14.38	13.88	12.73	13.18
Western Persian	15.35	17.70	18.60	16.05	16.75
Avg	15.09	16.35	16.66	15.17	15.33
<i>Mid-resource (42)</i>					

Continued on next page

Table 39 – continued

Language	C _{std}	C _{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
Armenian	11.47	14.00	16.10	12.46	10.98
Belarusian	12.64	18.36	21.34	12.99	11.89
Bengali	8.86	10.40	11.59	9.67	9.65
Bengali (Romanized)	7.00	4.92	8.46	1.75	0.93
Bosnian	23.91	23.53	25.31	22.98	22.44
Bulgarian	23.91	24.04	25.40	23.39	23.16
Catalan	20.56	21.41	20.13	19.22	21.30
Croatian	16.26	15.29	17.40	16.31	15.30
Eastern Panjabi	8.53	8.69	10.90	8.78	8.66
Finnish	19.55	20.73	21.60	18.56	19.36
Galician	18.64	22.22	20.17	19.20	21.62
Georgian	10.30	11.56	11.85	9.74	9.57
Gujarati	6.41	9.64	11.11	7.41	7.86
Hebrew	13.78	16.00	15.88	13.88	13.65
Hindi	10.74	13.22	14.98	12.78	12.78
Icelandic	17.49	18.06	19.09	15.39	16.26
Kannada	11.79	13.51	15.73	11.40	11.25
Kazakh	11.74	13.33	14.39	11.73	11.89
Kyrgyz	11.23	10.99	12.53	9.70	10.70
Lithuanian	11.39	10.55	11.96	10.02	9.50
Macedonian	17.52	17.41	17.43	17.01	15.77
Malayalam	10.33	12.64	14.62	10.38	11.77
Marathi	8.45	12.54	13.35	9.83	9.65
Nepali	11.13	14.98	15.25	12.89	12.95
North Azerbaijani	14.75	15.97	16.38	13.61	12.66
Northern Uzbek	13.48	13.46	13.81	12.79	14.19
Norwegian Bokmål	20.31	20.62	21.73	19.41	19.17
Norwegian Nynorsk	18.70	19.92	20.07	19.39	19.30
Serbian (Cyrillic)	16.34	17.32	18.06	17.88	17.22
Serbian (Latin)	17.49	19.01	18.11	18.36	17.82
Sinhala	5.40	6.58	7.96	6.04	5.72
Slovenian	17.23	18.29	18.81	17.17	16.84
Slovenian (CERK)	11.76	11.52	12.09	11.47	11.53
Standard Malay	17.49	20.77	22.65	18.41	20.41
Swahili	10.75	13.03	13.42	10.63	11.31
Tagalog	25.77	27.56	29.20	26.22	25.65
Tamil	13.48	16.94	20.04	15.36	12.75
Telugu	7.86	10.11	13.73	9.23	9.38
Thai	11.83	13.47	17.08	12.79	14.25
Tosk Albanian	13.28	14.24	14.24	12.88	12.89
Urdu	13.11	16.07	16.69	14.05	13.75
Urdu (Romanized)	13.68	11.59	10.78	4.53	4.93
Avg	13.96	15.35	16.46	13.75	13.78
<i>Low-resource (30)</i>					
Amharic	1.66	2.58	3.95	1.66	1.70
Assamese	8.10	8.86	10.35	7.80	7.84
Bambara	5.91	9.91	10.79	4.98	5.10
Bhojpuri	8.57	10.07	12.71	10.79	10.48
Central Kurdish	8.41	12.70	13.48	9.10	8.42
Dhundari	7.45	10.25	11.78	6.16	7.53
Ekpeye	4.38	4.64	6.18	3.63	3.30
Estonian	14.14	14.62	15.49	14.21	17.09
Faroese	14.58	17.55	17.75	14.21	14.21
Hausa	9.90	9.48	10.67	8.83	8.83
Hawaiian	11.99	16.53	16.99	10.16	11.55
Idoma	5.59	7.90	8.62	3.55	3.68
Igbo	9.51	17.16	19.10	9.08	8.58
Isoko	5.40	10.06	10.68	4.09	4.18
Javanese	15.19	15.59	15.44	13.98	14.61
Kinyarwanda	9.55	12.44	12.83	8.91	9.44
Kumzari	3.11	1.17	3.93	1.23	0.10
Lingala	9.86	11.56	11.62	8.55	8.26
Luo	9.01	10.18	9.51	6.26	6.58
Marwari	9.21	11.17	12.32	9.84	10.72
Meitei (Bengali script)	2.65	2.88	7.95	2.62	2.04

Continued on next page

Table 39 – continued

Language	C_{std}	C_{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
Meitei (Meitei script)	0.01	10.38	9.58	0.09	0.14
Nagamese	12.73	13.14	13.67	11.33	11.73
Nigerian Pidgin	19.35	20.81	21.20	19.35	18.75
Sindhi	7.17	9.97	10.56	7.19	7.12
Sindhi (Devanagari)	3.91	7.68	8.21	2.74	3.06
Urhobo	6.32	15.49	16.73	4.63	4.67
Uyghur	8.23	10.40	11.56	7.84	8.42
Yoruba	4.81	8.38	10.54	4.27	4.44
Zulu	9.93	13.90	14.04	9.96	8.94
Avg	8.22	10.91	11.94	7.57	7.72
Overall (116)	12.90	14.58	15.37	12.69	12.80

Table 40: Mistral-7B-Instruct-v0.3 chrF scores with varying context provided to MT2 on MKQA: High-resource (19) and Mid-resource (6).

Language	C_{std}	C_{ctx}	$q_t + a_e$	$q_e + a_e$	$r_e + a_e$
<i>High-resource (19)</i>					
Arabic	10.03	10.55	11.32	10.43	9.69
Chinese (HK)	5.43	6.10	6.37	6.22	5.57
Chinese (Simplified)	3.96	4.82	5.09	4.51	4.41
Chinese (Traditional)	3.63	4.10	4.37	3.96	3.71
Danish	15.62	16.59	18.28	17.02	15.81
Dutch	14.06	14.54	17.00	14.62	13.79
French	15.33	16.32	18.56	15.97	15.29
German	16.36	17.47	18.96	17.41	16.25
Hungarian	14.59	15.24	16.08	15.19	14.20
Italian	16.42	17.27	18.83	17.13	16.21
Japanese	6.40	10.74	10.92	9.01	8.91
Korean	3.88	4.70	5.01	4.60	4.09
Polish	14.64	15.19	16.53	15.54	14.45
Portuguese	15.53	16.32	18.13	16.19	15.24
Russian	12.46	12.91	14.75	13.12	12.21
Spanish	14.98	15.63	18.51	15.63	14.69
Swedish	15.97	17.28	18.46	17.14	16.18
Turkish	13.52	13.52	14.75	13.74	11.82
Vietnamese	11.69	12.44	13.05	12.09	11.01
Avg	11.82	12.72	13.95	12.61	11.76
<i>Mid-resource (6)</i>					
Finnish	11.74	12.33	12.84	11.87	10.57
Hebrew	7.96	8.36	8.78	8.01	7.34
Khmer	1.87	2.13	2.19	1.70	1.59
Malay	13.38	13.61	15.17	13.89	12.96
Norwegian	14.98	15.72	16.63	15.73	14.85
Thai	7.84	8.48	9.17	8.29	7.40
Avg	9.63	10.10	10.80	9.92	9.12
Overall (25)	11.29	12.09	13.19	11.96	11.13

H Significance testing

We now assess whether the context-aware cascade (C_{ctx}) consistently improves over the standard cascade (C_{std}) at the language level. For each model and dataset, we treat languages as the analysis units and perform two paired tests. First, for the earlier win-rate analysis (Table 3), we apply an exact binomial sign test after discarding tied languages, testing whether C_{ctx} or C_{std} beats the other more often than expected under a 50/50 null. Second, for the magnitude analysis (Table 2), we apply a two-sided Wilcoxon signed rank test (Wilcoxon, 1945) to the per-language differences $C_{\text{ctx}} - C_{\text{std}}$, testing whether the distribution of language-level gains is shifted above or below 0.

Because we evaluate multiple datasets per model, we report Benjamini-Hochberg (BH) corrected p -values (Benjamini and Hochberg, 1995), with correction applied separately within each model across

Dataset	Wilcoxon signed-rank test p (two-sided)		
	Llama-3.1-8B-Instruct	Mistral-7B-Instruct-v0.3	GPT-4o-mini
<i>Open-ended generation (chrF)</i>			
Aya	0.0391	0.0781	0.1125
BLEnD	3.05e-4	0.0017	0.8042
Global-PIQA-OE	6.62e-15	6.32e-15	3.39e-4
MKQA	1.06e-4	4.55e-5	0.0026
<i>Multiple-choice (accuracy)</i>			
Global-MMLU	2.40e-5	7.61e-8	0.1172
Belebele	7.66e-4	0.0045	3.52e-11
Global-PIQA	0.0481	2.88e-14	5.99e-9
MCSQA	0.3438	0.1128	1.0
<i>Math (accuracy)</i>			
MGSM	0.0016	0.0223	0.6429
MMath	0.0056	0.5000	—

Table 41: BH-corrected p -values for the Wilcoxon signed-rank test (two-sided) over language-level differences $C_{\text{ctx}} - C_{\text{std}}$. Green indicates a significant result in the direction $C_{\text{ctx}} > C_{\text{std}}$, red indicates a significant result in the direction $C_{\text{std}} > C_{\text{ctx}}$, and uncolored cells are not significant. BH correction is applied separately within each model across the ten datasets, and we use $p \leq 0.05$.

the ten benchmarks. These tests treat languages as approximately independent analysis units, which we acknowledge is a simplifying assumption given genealogical and areal correlations.

Table 41 and Table 42 show that the statistical evidence for improvements from C_{ctx} depends strongly on model strength. For Llama-3.1-8B-Instruct, the effect is highly consistent: 9 of 10 datasets remain positively significant after the BH correction under the Wilcoxon signed-rank test and 8 of 10 under the binomial sign test. Mistral-7B-Instruct-v0.3 is less consistent overall, but still shows broad evidence of improvement: 6 of 10 datasets were positively significant under both tests. By contrast, GPT-4o-mini passes far fewer tests: only 3 of 10 datasets were positively significant under both tests. These patterns are consistent with the discussion in § 4 about stronger improvements from context for weaker models.

Binomial sign test p (two-sided)				
Dataset	Lang.	Llama-3.1-8B-Instruct	Mistral-7B-Instruct-v0.3	GPT-4o-mini
Open-ended generation (chrF)				
Aya	6	0.0391	0.2734	0.3281
BLEnD	14	2.44e-4	0.0216	0.8893
Global-PIQA-OE	116	1.51e-17	2.48e-15	4.14e-5
MKQA	25	4.86e-5	2.98e-7	3.52e-4
Multiple-choice (accuracy)				
Global-MMLU	42	4.77e-5	3.03e-12	0.1969
Belebele	121	4.77e-5	0.0081	8.72e-13
Global-PIQA	116	0.0601	8.08e-20	4.95e-11
MCSQA	8	0.6875	0.7266	1.0
Math (accuracy)				
MGSM	11	0.0016	0.0223	0.6429
MMath	10	0.0056	0.6944	—

Table 42: BH-corrected p -values for the exact binomial sign test (two-sided) over non-tied languages. Green indicates a significant result in the direction $C_{\text{ctx}} > C_{\text{std}}$, red indicates a significant result in the direction $C_{\text{std}} > C_{\text{ctx}}$, and uncolored cells are not significant. BH correction is applied separately within each model across the ten datasets, and we use $p \leq 0.05$.