Cross-lingual Evaluation of Multilingual Text Generation

Shamil Chollampatt **Minh-Quang Pham**
Sathish Reddy Indurthi **Marco Turchi**
Zoom Communications, Inc.
shamil.chollampatt@zoom.us

Abstract

Scaling automatic evaluation of multilingual text generation of LLMs to new tasks, domains, and languages remains a challenge. Traditional evaluation on benchmark datasets carries the risk of reference data leakage in LLM training or involves additional human annotation effort. The alternative strategy of using another LLM as a scorer also faces uncertainty about the ability of this LLM itself to score non-English text. To address these issues, we propose an annotation-free cross-lingual evaluation protocol for multilingual text generation. Given an LLM candidate to be evaluated and a set of non-English inputs for a particular text generation task, our method first generates English references from the translation of the non-English inputs into English. This is done by an LLM that excels in the equivalent English text generation task. The non-English text generated by the LLM candidate is compared against the generated English references using a cross-lingual evaluation metric to assess the ability of the candidate LLM on multilingual text generation. Our protocol shows a high correlation to the reference-based ROUGE metric in four languages on news text summarization. We also evaluate a diverse set of LLMs in over 90 languages with different prompting strategies to study their multilingual generative abilities.

1 Introduction

Large language models (LLMs) such as GPT-4 (OpenAI, 2023) have shown remarkable text generation capabilities and have been useful for tasks like text summarization (Pu et al., 2023; Goyal et al., 2022), question answering (Zhao et al., 2023), and text simplification (Feng et al., 2023). However, they have been predominantly trained on English corpora (Brown et al., 2020; Touvron et al., 2023) and benchmarked on English datasets. Generally, LLMs are not able to replicate similar success in other languages (Lai et al., 2023; Zhang et al., 2023; Ahuja et al., 2023). Moreover, comparative studies (Ahuja et al., 2024) also reveal a substantial performance gap between proprietary and open-source LLMs in other languages.

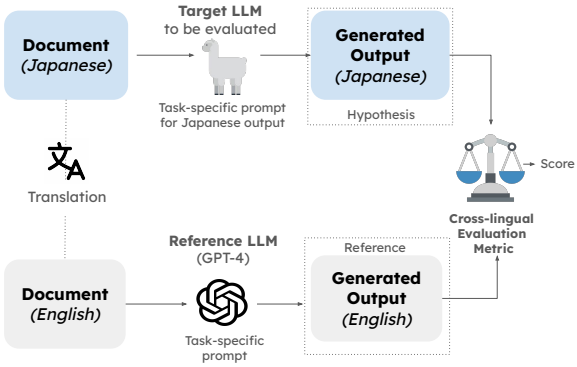

Figure 1: Proposed cross-lingual evaluation protocol.

LLMs are being increasingly trained on multilingual data and claim multilingual text generation capabilities (Jiang et al., 2024). This necessitates reliable methods of automatic evaluation. However, multilingual evaluation is often done using benchmarks dominated by language understanding tasks with limited representation of generation tasks (Lai et al., 2023; Liang et al., 2020; Asai et al., 2023) and often only on a handful of languages (Chen et al., 2022). Moreover, relying on such benchmarks leads to unfair comparisons due to data leakage (Zhou et al., 2023) since most LLMs are pre-trained on massive web-crawled corpora which may include the references from these benchmarks. Reference-based evaluation also hinders evaluation on newer tasks and domains due to annotation needs. The alternative approach is to use an LLM such as GPT-4 for scoring (Liu et al., 2023). In addition to the costs, the ability of GPT-4 to judge itself and other LLMs in other languages remains unclear. Moreover, LLM-based evaluation exhibits biases such as preferring longer text and their own outputs (Zheng et al., 2024; Shen et al., 2023).

Proceedings of the 31st International Conference on Computational Linguistics, pages 7766–7777
January 19–24, 2025.

To address these issues, we propose a cross-lingual evaluation protocol for evaluating multilingual text generation of LLMs without requiring annotated references or LLM-based scoring. Our protocol (shown in Figure 1) relies on translating the non-English input text into a reference language, typically English. An LLM (referred to as *reference* LLM) that is known to have a strong performance in English (e.g. GPT-4) is used to generate a reference in English given the English translation of the non-English input text. The English translation may be generated using machine translation (MT). For almost all language directions in WMT 2023 campaign (Kocmi et al., 2023), it was concluded that "MT systems produce outputs that cannot be identified as being worse than the manually produced references translations". Moreover, minor translation errors on the input, especially fluency errors, are unlikely to affect the reference quality given the robustness of the reference LLM.

The LLM candidate (referred to as *target LLM*) to be evaluated is used to generate the output text in a specific language given the original input text. We then make use of neural cross-lingual evaluation metrics to compare this output against the generated English reference to assess the quality.

We evaluate our protocol by computing correlations (Louis and Nenkova, 2013; Ellouze et al., 2013) against the de-facto reference-based multilingual ROUGE metric (Lin, 2004). We find that our protocol closely correlates to it, indicating that our low-cost, reference-free approach can be an alternative to reference-based evaluation of multilingual text generation. Furthermore, we use our method to rank several popular LLMs in several languages on the same task providing insights into their multilingual text generation capabilities.

2 Cross-lingual Evaluation Protocol

Our protocol assesses the text generation capability of the target LLM in a non-English language l for a given text generation task. In this task, the LLM is provided with an input d_l in language l. The goal is to generate the output text h (or *hypothesis*) in l based on d_l following the task instructions. Our evaluation protocol requires two key components: 1) the translation or equivalent of d_l in a reference language, which we consider to be English (d_{en}); and 2) the availability of a reference LLM, such as GPT-4, known for its strong text generation capabilities in English.

Given d_{en}, the reference LLM generates an English reference r_{en} given a suitable prompt. Finally, we use a neural cross-lingual evaluation (XE) metric to compare the hypothesis h generated by the target LLM against the English reference r_{en} generated by the reference LLM. We propose two XE metrics.

1. XE using Sentence Embeddings (XESE): XESE uses the formulation of the cross-lingual summarization metric LaSE introduced by Bhattacharjee et al. (2023), originally used to compare the system-generated summary in one language to that of the human-written reference in another language. The comparison is done using similarity measurement between sentence embeddings from a multilingual text representation model.

$$\text{XESE} = \text{SE}(h, r_{en}) \times \text{LP}(h, r_{en}) \times \text{LC}_l(h)$$

where $\text{SE}(h, r_{en})$ is inner-product between sentence embeddings of h and r_{en}. Length Penalty (LP) penalizes long hypotheses h compared to their references r_{en} and (2) Language Confidence (LC_l) is included for penalizing outputs that are not in the intended language l.

$$\text{LP}(h, r_{en}) = \begin{cases} 1 & |h| \leq |r_{en}| + \epsilon \\ \exp(1 - \frac{|h|}{|r_{en}|+\epsilon}) & \text{otherwise} \end{cases}$$

where ϵ is an offset to account for length differences (set to 6 based on Bhattacharjee et al. 2023), and LC_l is the language confidence that penalizes if h is not in the expected language l:

$$\text{LC}_l(h) = \begin{cases} 1 & \text{if } \text{argmax}_{l'} P_{\text{lid}}(l'|h) = l \\ P_{\text{lid}}(l|h), & \text{otherwise.} \end{cases}$$

where $P_{\text{lid}}(l|h)$ is the probability predicted by a language identification (LID) model that the hypothesis h is in the target language l.

The overall system score is the average score across all hypotheses scores.

2. XE by Translation Quality (XETQ): XETQ uses MT quality estimation (QE) as the backbone. QE measures the quality of translated output text given the original source text. Typically, QE is done using supervised models that gives a quality score given the source text and the MT output. In our protocol, we treat r_{en} as the source text and h

as the MT output for QE. LC_l is also incorporated in the XETQ:

$$\text{XETQ} = \text{QE}(h, r_{en}) \times \text{LC}_l(h)$$

LC_l is necessary since the underlying QE models we use are multilingual and the QE score itself does not penalize if the language of h is different from the target language. We avoid LP since QE models implicitly penalize length differences.

Our method does not rely on human-annotated references in either English or the target language. However, it requires English translations of the inputs which can be obtained automatically via good-quality MT systems (as shown in Section 4). This enables the online evaluation of deployed LLMs on real user inputs as well. The protocol can be applied to specialized tasks, domains, and a large number of languages. Moreover, not relying on pre-existing public benchmarks avoids inflated scores due to *benchmark data leakage* (Zhou et al., 2023) where the references in the benchmark test sets are included during the pre-training of LLMs. Also, we use the costly reference LLM only once to create one set of English references that can be used in evaluating the outputs from target LLMs in multiple languages. This is significantly cheaper compared to using an LLM-as-a-judge where typically one or more costly LLM inference calls are made to score each generated text for all target LLMs and languages leading to significantly higher costs.

3 Experiments

3.1 Evaluations

We conduct two kinds of evaluation: (1) meta-evaluation of the XE metrics and (2) comparison of LLMs using the proposed XE metrics. Both evaluations are done on the news text summarization task.

Meta-evaluation: Firstly, we conduct a *meta-evaluation* to assess if our reference-free protocol can substitute automatic text generation metrics for scaling multilingual text generation evaluation. For this purpose, we compute the *system-level* and *summary-level* correlations of XESE and XETQ against ROUGE-2 (Section 4.1). ROUGE-2 is the de-facto reference-based summarization metric and correlates highly with human scores for news summarization (Bhandari et al., 2020). Additional evaluation against BLEU (Papineni et al., 2002) is reported in Appendix B.6.

Two levels of correlation are employed: *system-level* and *summary-level*. At the system-level, the correlation of the ranking of the systems (i.e., target LLMs) by a metric $\mathcal{M}$ to that of the ranking produced by ROUGE-2 ($\mathcal{R}$) is computed via Spearman's rank correlation coefficient (ρ):

$$\rho = 1 - \frac{6 \sum_{i=1}^{m} (\text{rank}^i_{\mathcal{M}} - \text{rank}^i_{\mathcal{R}})^2}{n(n^2 - 1)}$$

where $\text{rank}^i_{\mathcal{M}}$ is the rank assigned for the ith system by metric $\mathcal{M}$ among the m systems.

On the other hand, summary-level correlation measures the concordance between metric $\mathcal{M}$ and ROUGE-2 ($\mathcal{R}$) in ranking individual summaries generated by the m systems for the same input j. We use the Kendall's rank correlation coefficient as the correlation statistic which also accounts for the ties assigned by the metrics. Summary-level correlation τ over n inputs is given by:

$$\tau = \frac{1}{n} \sum_{j=1}^{n} (\tau_j([\mathcal{M}^1_j \dots \mathcal{M}^m_j], [\mathcal{R}^1_j \dots \mathcal{R}^m_j]))$$

where $\mathcal{M}^i_j$ and $\mathcal{R}^i_j$ are the metric score and ROUGE-2 score for the output generated by the ith system on the jth input, respectively. τ_j is the Kendall's tau statistic for the jth input τ_j given by:

$$\tau_j = \frac{C_j - D_j}{\sqrt{(C_j + D_j + T^{\mathcal{M}}_j)(C_j + D_j + T^{\mathcal{R}}_j)}}$$

where C_j is the number of concordant pairs and D_j is the number of discordant pairs. A concordant pair is when the metrics $\mathcal{M}$ and $\mathcal{R}$ rank a pair of outputs (from system i and k, for example) similarly, i.e. if $\mathcal{M}^i_j < \mathcal{M}^k_j$ and $\mathcal{R}^i_j < \mathcal{R}^k_j$ or if $\mathcal{M}^i_j > \mathcal{M}^k_j$ and $\mathcal{R}^i_j > \mathcal{R}^k_j$. Similarly a discordant pair is when they rank dissimilarly. $T^{\mathcal{M}}_j$ and $T^{\mathcal{R}}_j$ are the number of ties assigned by $\mathcal{M}$ and $\mathcal{R}$, respectively for the jth input. Ties are neither considered in the discordant nor concordant pairs. The correlation statistics are computed using the `nlpstats` Python library.

LLM Comparison: Secondly, a comparison of the multilingual text generation capabilities of popular LLMs is conducted using our protocol. The results with XESE are provided in Section 4.2.

3.2 Datasets

For the meta-evaluation, we require a dataset with human-written reference summaries in the target language to compute ROUGE-2. We also require the corresponding articles in English for computing XESE and XETQ. Hence, we aligned 100 article-summary pairs from CrossSum (Bhattacharjee et al., 2023) in four languages: Arabic (ar), Spanish (es), Portuguese (pt), and Chinese (zh), with their corresponding human-written English articles. We also use translated articles from each of these four languages to English using NLLB-3.3B MT model (NLLB team et al., 2022) to additionally investigate the applicability of our method to scenarios where human translations are unavailable.

For the comparison of various LLMs with XE metrics, we do not require any human-written summaries. Hence, we use NTREX-128 (Federmann et al., 2022) dataset containing 123 news articles aligned across 128 languages including English. We primarily evaluate the LLMs on nine languages: Arabic (ar), German (de), Spanish (es), French (fr), Italian (it), Japanese (ja), Korean (ko), Portuguese (pt), and Traditional Chinese (zh). We further report evaluations on an additional set of 83 languages (in Appendix B.8).

3.3 Metrics

Proposed XE Metrics: For XESE, we experiment with both LaBSE (Feng et al., 2022) and SONAR (Duquenne et al., 2023) for computing sentence embeddings. They support over 100 and 200 languages, respectively. For XETQ, we experiment with CometKiwi (Rei et al., 2022, 2023) and xCOMET-XL (Guerreiro et al., 2023) as the QE model, both supporting over 90 languages. We use fasttext (Joulin et al., 2016) LID to compute language confidence. It can predict 176 languages. GPT-4 is used as the reference LLM for generating the English summaries given the corresponding English articles.

Baseline Metrics: We use two baseline metrics for comparison. (1) **G-Eval** (Liu et al., 2023) is a popular LLM-based scoring method that uses chain-of-thought prompting with GPT-4 to provide aspect-based scores of summaries. G-Eval scores four aspects (*coherence*, *consistency*, *fluency*, and *relevance*) of a summary on a 1 to 5 scale. We report the correlations of each aspect separately and the average correlation across all aspects. (2) **MT ROUGE$_{GPT}$** is the ROUGE-2 score of the English machine translation (MT) of the summaries generated by the target LLM against the corresponding GPT-4 generated English reference summaries.

	System-level (ρ)				Summary-level (τ)			
	ar	**es**	**pt**	**zh**	**ar**	**es**	**pt**	**zh**
G-Eval								
(Relevance)	.33	.21	.17	.00	.11	.10	-.01	-.01
(Coherence)	.43	.21	.19	-.10	.07	.10	.02	-.05
(Consistency)	.40	.24	.14	.12	.13	.19	.04	.08
(Fluency)	.31	.21	.19	.16	.08	.16	.06	.00
Avg. Correl	.37	.22	.17	.04	.10	.14	.03	.00
MT ROUGE$_{GPT}$	.64	.48	.43	-.29	.14	.17	.14	.03
XESE [SONAR]	**.98**	.90	.76	.88	**.50**	.53	.57	.62
XESE [LABSE]	.90	**.93**	**.83**	.95	.49	**.56**	**.60**	**.64**
XETQ [xCOMET]	**.98**	.74	.67	**.98**	.47	.46	.49	.57
XETQ [KIWI]	.88	.74	.71	**.98**	.37	.48	.47	.52
using machine-translated inputs								
XESE^{MT} [LABSE]	.93	.88	.86	.95	.46	.55	.59	.64
XETQ^{MT} [KIWI]	.91	.76	.67	.93	.42	.50	.47	.53

Table 1: Correlation of the metrics against ROUGE-2.

3.4 LLMs

We evaluate the following open-source instruction fine-tuned LLMs which are available on Huggingface Llama-2-7B-Chat-hf, Llama-2-13B-Chat-hf, Llama-2-70B-Chat-hf, Mistral-7B-Instruct-v0.2, TowerInstruct-v0.1, Gemma-7B-it, Mixtral-8x7B-Instruct-v0.1, BLOOMZ. Additionally, we evaluate two proprietary LLMs: OpenAI GPT-3.5 (gpt-3.5-turbo-1106), and GPT-4 (gpt-4-1106-preview). We evaluate all LLMs in a zero-shot setup using prompts with English instructions (P_{en}) and with instructions translated to the target language (P_l) (see Appendix A for prompts and parameters). For meta-evaluation, we use outputs of open-source LLMs with P_l.

4 Results and Discussion

4.1 Meta-Evaluation of Metrics

In Table 1, we report system-level (ρ) and summary-level (τ) correlations against ROUGE-2. We find that all variants of XETQ and XESE show strong system-level correlations against ROUGE compared to all aspects of G-Eval and also to MT ROUGE$_{GPT}$ which employs the same English GPT-4 generated reference used by XETQ and XESE. A high system-level correlation indicates that even without references, XETQ and XESE can closely match the ranking of LLMs produced by ROUGE. This shows that the proposed XE metrics are valuable alternatives to reference-based automatic multilingual evaluation measures. While MT

	ar		de		es		fr		it		ja		ko		pt		zh	
	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l
Llama-2-7B-Chat	.05	.48	.09	.42	.23	.49	.11	.39	.18	.45	.05	.30	.10	.49	.16	.50	.07	.30
Mistral-7b-Instruct-v0.2	.64	.60	.33	.35	.27	.27	.24	.30	.25	.31	.22	.21	.54	.46	.32	.33	.09	.07
Tower-Instruct-v0.1	.21	.54	.01	.27	.04	.24	.04	.19	.05	.23	.23	.32	.53	**.66**	.03	.25	.08	.21
Gemma-7B-it	.58	.53	**.66**	**.60**	**.60**	**.61**	**.58**	**.61**	**.58**	.57	**.46**	.44	**.65**	.62	**.61**	**.65**	**.41**	**.49**
Llama-2-13B-Chat	.36	.53	.50	.59	.57	.65	.53	.57	**.58**	**.58**	.41	**.47**	.51	.58	.53	.60	.35	.38
Llama-2-70B-Chat	.09	.56	.41	.53	.52	.58	.46	.59	.50	**.58**	.27	.45	.24	.61	.52	.57	.16	.37
Mixtral-8x7b-Instruct	**.69**	**.65**	.46	.48	.33	.36	.35	.31	.38	.38	.36	.27	.64	.62	.35	.41	.18	.11
BLOOMZ	.53	.51	.30	.36	.54	.55	.51	.51	.43	.49	.31	.28	.33	.19	.52	.53	**.41**	.41
GPT-3.5	.64	**.68**	**.69**	**.65**	**.70**	**.67**	**.67**	**.66**	**.70**	**.61**	**.53**	.51	.65	.67	**.67**	**.67**	**.50**	.22
GPT-4	**.72**	.61	.67	.50	.62	.56	.57	.46	.64	.51	.53	**.60**	**.70**	**.71**	.59	.56	.40	**.29**

Table 2: XESE scores using prompts with instructions in English (P_{en}) and in the target language (P_l).

$ROUGE_{GPT}$ is also annotation free, its correlation is lower compared to our XE metrics. This is possibly due to its reliance on the MT of the outputs and the discrete n-gram matches, leading to a pronounced coverage bias. On the other hand, XE metrics measure the proximity to the English GPT-4 generated summary in the continuous space and do not rely on MT.

Ranking at the summary-level is harder than aggregated system ranking and the correlation values tend to be lower (Novikova et al., 2017; Peyrard et al., 2017). Nonetheless, XE metrics perform better at the summary level also. XESE [LABSE] performs best on average compared to all other XE metrics in both system and summary-level correlations. XETQ [KIWI] performs on par to XETQ [xCOMET], except in Arabic. We use XESE [LABSE] for subsequent LLM comparisons (referred as XESE, henceforth).

We also compute correlation of the XE metrics when the reference summaries are generated by GPT-4 from machine-translated English articles ($XESE^{MT}$ and $XETQ^{MT}$ in final two rows of Table 1). Similar to the case with human-written English articles, we observe a strong correlation to ROUGE-2. Since MT is used only on the input side and as long as meaning is preserved, minor translation inaccuracies or lack of native-like fluency, if any, does not seem to impact the metrics' usefulness.

4.2 Comparison of LLMs

We evaluate several LLMs on nine languages using the XESE metric in Table 2 using prompts with English instruction (P_{en}) and with instructions in the corresponding target language (P_l). Additional evaluations on 83 other languages are in Appendix B. Gemma achieves the best performance (except on ar) among all open-source models despite being comparatively small (7B parameters). This can be attributed to its larger vocabulary that covers multiple languages despite not being trained with a notable quantity of multilingual corpora (Gemma Team, 2024). This finding holds even on the extended set of languages in Section B.8 (Tables 10 and 11). Unsurprisingly, high scores are also observed for Mixtral-8x7b (Jiang et al., 2024) which is a large mixture-of-experts model pre-trained with multilingual data. Despite the multilingual pre-training and size, BLOOMZ is not competitive with these LLMs. On ko with P_l, TowerInstruct-7B, an LLM fine-tuned for MT including ko, performs the best. Among the proprietary models, GPT-3.5 is competitive with GPT-4 and observes improvements, particularly in Latin-based languages. We also find that LLMs in general, except for GPT-4, tend to produce better scores with the P_l strategy compared to P_{en}. In P_l, the entire prompt is in a single language and this may suit the language modeling loss of next-word prediction better. This peculiarity of GPT-4 possibly indicates English instructions being used during fine-tuning on multilingual tasks.

5 Conclusion

We propose a simple protocol for multilingual text generation evaluation of LLMs by cross-lingually comparing the output generated by LLMs with a reference output generated by another LLM in English. Our approach can be extended to specialized text generation tasks, custom domains, and low-resource languages without annotations by humans. We propose two metrics for the cross-lingual evaluation, XETQ and XESE. Our study of their correlations to automatic metrics in multiple languages shows that the protocol can be used in place of reference-based evaluation. We also evaluate several popular LLMs on a large set of languages highlighting their capabilities.

Acknowledgment

The work presented in this paper is partially funded by the European Union's Horizon research and innovation programme under grant agreement no. 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People).

Limitations

We identify the following limitations in this work:

- XESE and XETQ metrics require the underlying model to support the target language.
- Due to cost, time, and license restrictions, we were unable to add evaluations of larger and more recent LLMs. The scope of the paper is limited to the applicability of the evaluation method which can be easily applied to evaluate text generation ability of any new LLM.
- Our approach has been only tested on the news summarization task. However, the summarization results are encouraging and show that our approach can be used as a replacement for reference-based metrics.
- We showed that our protocol can be used in place of reference-based automatic metrics like ROUGE. However, with human annotators in other languages, meta-evaluation can be done against human ratings instead.
- Our meta-evaluation is also limited to four languages due to computational costs. The baseline G-Eval is particularly costly since it involves running GPT-4 four times for a single example.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. *Preprint*, arXiv:2311.07463.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. *Preprint*, arXiv:2305.14857.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

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

Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. 2022. MTG: A benchmark suite for multilingual text generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations. *Preprint*, arXiv:2308.11466.

Samira Ellouze, Maher Jaoua, and Lamia Hadrich Belguith. 2013. An evaluation summary method based on a combination of content and linguistic metrics. In *Proceedings of the 2013 International Conference Recent Advances in Natural Language Processing*.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models. *Preprint*, arXiv:2302.11957.

Gemma Team. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *Preprint*, arXiv:2209.12356.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Preprint*, arXiv:2310.10482.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *Preprint*, arXiv:1612.03651.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop on Text Summarization Branches Out*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

NLLB team, Marta Ruiz Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *Preprint*, arXiv:2309.09558.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3Exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Advances in Neural Information Processing Systems*.

Kun Zhao, Bohao Yang, Chenghua Lin, Wenge Rong, Aline Villavicencio, and Xiaohui Cui. 2023. Evaluating open-domain dialogues in latent space with next sentence prediction and mutual information. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your LLM an evaluation benchmark cheater. *Preprint*, arXiv:2311.01964.

A Prompts and Parameters

We use the following English-instruction prompt for the summarization task in the respective language:

```
Summarize the text concisely using only the
{{language}} language:

TEXT: {{text}}

SUMMARY:
```

We translate the above prompt to the nine languages using a high-quality commercial MT engine and conduct evaluation with the translated prompts in the corresponding language.

For generating reference summaries in English given the translated English input, we use the following prompt with GPT-4.

```
Generate a concise summary in two sentences or
fewer for the text: {{text}}
Summary:
```

For generating the machine translated English inputs and for computing the baseline metric MT-ROUGE-G metric, we employ an MT pipeline with open-source tools and models. We split the article into sentences using PySBD[1] and translate using the NLLB 3.3B using the Huggingface *translator* pipeline. We set `length_penalty` to 1.0, `max_length` to 1024, and `num_beams` to 4.

For running the publicly available LLM models, we use vLLM. For all models, we set the sampling temperature to 1.0 and `top_p` to 1.

B Additional Results

B.1 XETQ Evaluation

We report scores of the LLMs using the XETQ metric (Table 3). According to XETQ, Mixtral-8x7b performs the best among open-source models with the exception of Chinese, Japanese, and Korean where Gemma seems to be superior. This is in contrast to the findings by XESE metric. Nonetheless, XESE metric can be considered to be more reliable due to its higher overall correlation to ROUGE-2.

B.2 Language Accuracy

Since these LLMs are predominantly trained on English corpora, they have a tendency to produce responses in English language. We specifically evaluate the language accuracy, i.e., the percentage of times the LLM accurately generated in the target language (Table 4). We show the results when using the prompts having instructions in English (P_{en}) and in the corresponding target language (P_l). As expected, GPT models generate the output in the correct language for majority of languages in both prompting strategies. Among the other models, Gemma-7B-it seems to be the most reliable in terms of language accuracy followed by Mixtral-8x7b. We also find that both Mistral models exhibit robustness especially in Latin-based languages and `ar`. The language accuracy also generally improves when the P_l strategy is used.

B.3 Contribution of Penalty Factors

We investigate the contribution of LP and LC in the XE metrics. The results are reported in Table 5. We find the correlation values drop sharply with the removal of these correction factors in the final metric. We observe a steeper drop when we ablate LC in both metrics. This indicates the necessity to integrate the validation of the language of the generated output when evaluating multilinguality of LLMs.

B.4 Reference LLM

To understand the generalization to another reference LLM instead from GPT-4, we apply our evalu-

[1] https://github.com/nipunsadvilkar/pySBD

	ar		de		es		fr		it		ja		ko		pt		zh	
	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l
LLaMa-7B-Chat	.02	.10	.04	.19	.16	.36	.06	.25	.11	.25	.04	.24	.05	.19	.10	.33	.05	.24
Mistral-7b-Instruct-v0.2	.10	.10	.24	.27	.30	.34	.21	.27	.28	.34	.14	.15	.16	.15	.27	.32	.11	.12
Tower-Instruct-v0.1	.04	.08	.01	.19	.05	.25	.03	.15	.07	.22	.15	.22	.18	.23	.04	.22	.10	.21
Gemma-7B-it	.20	.15	.26	.23	.37	.38	.29	.26	.37	.36	.39	.35	**.35**	**.35**	.35	.35	**.39**	**.32**
LLaMa-13B-Chat	.09	.13	.22	.26	**.39**	.37	.26	.32	.35	.34	.29	.34	.19	.23	.32	.33	.24	.27
LLaMa-70B-Chat	.03	.14	.20	.27	.37	.43	.28	.33	.33	.38	.24	**.36**	.11	.25	.32	**.39**	.12	**.32**
Mixtral-8x7b-Instruct	**.23**	**.19**	**.42**	**.42**	**.39**	**.39**	**.37**	**.34**	**.45**	**.40**	**.33**	.25	**.35**	.28	**.41**	**.39**	.34	.23
BLOOMZ	.12	.14	.01	.07	.12	.15	.06	.07	.06	.07	.08	.08	.05	.02	.09	.12	.15	.15
GPT-3.5	.33	.46	.42	.47	.44	.46	.39	.40	.46	.48	.42	.52	.41	**.48**	.45	.47	.44	.51
GPT-4	**.48**	**.53**	**.47**	**.52**	**.49**	**.51**	**.44**	**.44**	**.51**	**.53**	**.56**	**.50**	**.58**	.47	**.50**	**.52**	**.52**	**.54**

Table 3: XETQ scores using prompts with instructions in English (P_{en}) and in the target language (P_l).

	ar		de		es		fr		it		ja		ko		pt		zh	
	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l	P_{en}	P_l
Llama-2-7B-Chat-hf	.09	.87	.15	.76	.43	.91	.18	.86	.30	.79	.11	.73	.15	.76	.29	.92	.15	.76
Mistral-7b-Instruct	.98	.98	.96	.95	.95	.98	.91	.98	.94	.98	.76	.68	.90	.77	.94	.97	.59	.55
Tower-Instruct-13B	.37	.98	.07	.97	.15	**1.**	.13	.98	.24	.98	.66	.94	.78	.99	.14	.98	.41	.94
Gemma-7B-it	.98	.98	.99	.99	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	**.99**	**1.**	**1.**	**1.**	**1.**	**.98**	**.98**
Llama-2-13B-Chat-hf	.57	.89	.75	.93	.94	.98	.87	.98	.93	.94	.78	.93	.77	.90	.82	.97	.76	.85
Llama-2-70B-Chat-hf	.15	.89	.66	.85	.88	.98	.83	.99	.80	.95	.58	.94	.35	.89	.86	.96	.34	.90
Mixtral-8x7b-Instruct	**1.**	.98	**1.**	**1.**	**1.**	**1.**	**1.**	.98	**1.**	**1.**	**.97**	.89	**.97**	.95	.96	.97	.89	.80
BLOOMZ	**1.**	**.99**	.62	.85	.98	**1.**	.98	**1.**	.84	.95	.81	.87	.88	.85	.98	**1.**	.96	.97
GPT-3.5	.98	.98	**1.**	**1.**	**1.**	**1.**	.98	**1.**	.98	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**
GPT-4	.98	.98	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	**1.**	.99	**1.**

Table 4: Language accuracies using prompts with instructions in English (P_{en}) and in target language (P_l).

	System-level (ρ)				Summary-level (τ)			
	ar	**es**	**pt**	**zh**	**ar**	**es**	**pt**	**zh**
G-Eval	0.40	0.38	0.17	-0.05	.10	.12	.01	-.02
XESE	.90	.93	.83	.95	.49	.56	.60	.64
– LP	.90	.79	.81	.26	.45	.43	.49	.13
– LC_l	.55	.38	.21	.02	-.01	.10	.05	-.11
XETQ	.88	.74	.71	.98	.37	.48	.47	.52
– LC_l	-.43	.10	.14	.00	-.18	.04	.03	-.06

Table 5: Ablation of length penalty (LP) and language confidence (LC_l)

	System-level (ρ)				Summary-level (τ)			
	ar	**es**	**pt**	**zh**	**ar**	**es**	**pt**	**zh**
using GPT-4 reference								
XESE[LABSE]	.90	.90	.86	.98	.57	.62	.60	.66
XETQ[KIWI]	.88	.69	.69	.95	.40	.44	.43	.50
using Claude-3.5 reference								
XESE[LABSE]	90	.93	.83	.76	.51	.56	.60	.65
XETQ[KIWI]	.88	.74	.67	.98	.45	.48	.50	.53

Table 6: Correlation of the metrics when references are generated using Claude 3.5 instead of GPT-4

ation protocol using reference summaries generated by Claude 3.5[2]. Correlation results are reported in Table 6. We find that the correlations remain similarly high for the metrics (except for XESE with zh possibly due to the additional LP factor that needs to be tuned).

[2] https://www.anthropic.com/news/claude-3-5-sonnet

	System-level (ρ)				Summary-level (τ)			
	ar	**es**	**pt**	**zh**	**ar**	**es**	**pt**	**zh**
XESE[LABSE]	.90	.90	.86	.98	.57	.62	.60	.66
XETQ[KIWI]	.88	.69	.69	.95	.40	.44	.43	.50

Table 7: Correlation of the metrics with BLEU

B.5 Correlation against BLEU

In addition to ROUGE-2, we also study the correlation of XESE and XETQ against another popular reference-based text generation metric, BLEU (Papineni et al., 2002). Results are reported in Table 8. We use `sacrebleu`[3] with `flores200` tokenizer. BLEU was originally used for machine translation Similar to the conclusions observed from ROUGE, we observe high correlations for XESE (>85 on all languages). This indicates the ability of XESE to match abilities of diverse text generation metrics.

B.6 Correlation against G-Eval

We also assess the correlation of XESE and XETQ metrics against the four G-Eval aspects to understand if they correspond to or model any of them. We find that XETQ exhibits higher than usual correlation on es and pt, particularly on the content-based aspects (*relevance*, *coherence*, and *consistency*), against ROUGE-2. However, the correla-

[3] https://github.com/mjpost/sacrebleu

	XESE[LABSE]				XETQ[KIWI]			
	ar	**es**	**pt**	**zh**	**ar**	**es**	**pt**	**zh**
(Relevance)	.19	.12	.02	.19	.21	.69	.60	.14
(Coherence)	.26	.12	.00	.10	.24	.62	.55	.05
(Consistency)	.21	.19	.12	.19	.26	.76	.67	.33
(Fluency)	.05	.21	-.12	.31	.00	.55	.33	.23
Avg. Correl	.18	.16	.01	.20	.18	.65	.54	.19

Table 8: System-level correlation with G-Eval

	fr	**ja**	**ru**	**tr**
XESE[LABSE]	.98	.57	.74	.86
XETQ[KIWI]	.64	.83	.76	.79

Table 9: Correlation with ROUGE-2 on other languages

tion of our best performing XESE metric against G-eval is low on all aspects.

B.7 Correlation on Additional Languages

The correlation on four languages in Table 1 used multi-way aligned articles from CrossSum dataset with all languages sharing the same set of corresponding English articles leading to better comparability of correlation across languages. In this section, we conduct additional correlation analysis on four more languages: French (`fr`), Japanese (`ja`), Russian (`ru`), and Turkish (`tr`) where shared alignments to English were not available in CrossSum, but are independently aligned to their corresponding English articles. To compute the reference summaries in English for the test set in each language, we use GPT-4o (`gpt-4o-2024-05-13`).

We sample 100 article-summary pairs for each language along with their corresponding English article-summary pairs. The system-level correlation is reported in Table 9. We find that the metrics continue to exhibit a high positive correlation to ROUGE-2 in general. A lower correlation is observed on `ja` with XESE. This may also depend on the ability of the underlying model, i.e., LaBSE in this case, to work in that particular language.

B.8 Extending Evaluation of LLMs

We conduct evaluation on an extended set of languages using XESE scores and prompts with instructions in the target language (P_l). We evaluate on the remaining languages that are available in NTREX-128 and also supported by LaBSE and the language identification (LID) model that we use, resulting in 83 additional languages. The instructions are translated to corresponding language using NLLB-3.3B (NLLB team et al., 2022). We also use the LID model from NLLB for the extended language coverage. We report the results in Tables 10 and 11 with abbreviated LLM names (L2: Llama-2, M: Mistral, MX: Mixtral, TI: TowerInstruct, G: Gemma). We run the same set of open-source instruction fine-tuned LLMs used in earlier comparisons with the exception of BLOOMZ due to its size and computational runtime. For proprietary LLMs, we use the cheaper GPT-4o (`gpt-4o-2024-05-13`) in place of GPT-4.

As exhibited by previous experiments with XESE, Gemma performs the best among all open-source models on an overwhelming number of languages. Mixtral-8x7b comes second in the remaining languages with a few exceptions where the Llama models (13B or 70B) give a better result. This further signifies the importance of having an extended vocabulary in building the base LLM despite the pre-training of Gemma being predominantly done on English itself. Unsurprisingly, we also find that the GPT models mostly dominate over open-source models. However, most Indo-Aryan languages (Hindi, Bengali, Gujarati, Kannada, Punjabi, Sinhala, Tamil, Telugu, Urdu) are exceptions where open-source models outperform the GPT models.

Language	L2-7B	M-7B	TI-13B	G-7B	L2-13B	L2-70B	MX-8x7B	GPT-3.5	GPT-4o
Afrikaans	.16	.13	.13	**.56**	.37	.37	.27	**.67**	.60
Albanian	.21	.21	.16	.41	.22	.30	**.43**	.59	**.63**
Amharic	.13	.04	.05	.00	.12	**.16**	.11	.14	**.31**
Armenian	.11	.14	.16	**.24**	.17	.07	.20	.36	**.54**
Azerbaijani	.25	.40	.27	**.47**	.42	.35	.38	.64	**.66**
Basque	.20	.19	.19	**.49**	.25	.21	.22	**.59**	.56
Belarusian	.02	.21	.13	**.61**	.17	.13	.49	.68	**.71**
Bengali	.22	.33	.22	**.62**	.34	.26	.39	.47	**.48**
Bosnian	.33	.20	.18	**.43**	.41	.38	.24	.50	**.54**
Bulgarian	.37	.44	.33	**.65**	.48	.50	.49	**.73**	.64
Burmese	.07	.08	.04	**.17**	.06	.07	.13	.12	**.24**
Catalan	.60	.24	.18	.63	**.64**	.59	.40	**.71**	.57
Chinese (Simpl.)	.13	.02	.11	**.33**	.25	.20	.06	**.35**	.27
Croatian	.49	.37	.29	.54	**.63**	.58	.41	.65	**.66**
Czech	.43	.47	.36	**.66**	.58	.51	.52	**.73**	.67
Danish	.38	.24	.30	**.65**	.61	.55	.32	**.68**	.62
Dutch	.39	.30	.24	**.62**	.57	.55	.35	**.64**	.55
Estonian	.13	.35	.27	**.63**	.42	.49	.51	.72	**.74**
Finnish	.61	.35	.48	.68	**.69**	**.69**	.20	**.75**	**.75**
Galician	.06	.05	.01	**.64**	.14	.10	.36	**.69**	.53
Georgian	.21	.14	.24	**.44**	.34	.31	.37	.51	**.61**
Greek	.49	.58	.43	.59	.53	.56	**.66**	**.55**	.49
Gujarati	.11	.09	.03	.31	.07	.11	**.34**	.16	**.17**
Hausa	.10	.10	.09	**.36**	.13	.15	.12	.14	**.42**
Hebrew	.51	.56	.43	**.59**	.52	.43	.57	**.68**	**.68**
Hindi	.35	.46	.41	.36	.44	.45	**.47**	**.44**	.36
Hungarian	.48	.52	.40	**.61**	.54	.54	.48	**.70**	.69
Icelandic	.19	.28	.25	**.52**	.31	.30	.34	**.66**	.59
Igbo	.10	.18	.12	**.31**	.17	.18	.18	.27	**.44**
Indonesian	.58	.41	.37	**.66**	.62	.61	.53	**.72**	.68
Irish	.21	.01	.20	**.24**	.22	.22	.14	**.49**	.45
Kannada	.07	.11	.03	**.45**	.05	.09	.42	.16	**.18**
Kazakh	.09	.13	.21	**.42**	.09	.08	.28	.62	**.74**
Khmer	.12	.07	.06	**.59**	.13	.14	.17	.21	**.26**
Kinyarwanda	.23	.21	.11	**.31**	.24	.21	.16	.20	**.36**
Kirghiz	.14	.11	.04	.22	.14	.16	**.36**	.62	**.65**
Lao	.09	.12	.05	**.22**	.09	.07	.14	.10	**.19**
Latvian	.30	.26	.19	**.59**	.31	.29	.39	**.69**	**.69**
Lithuanian	.27	.45	.23	.48	.33	.23	**.54**	.70	**.75**
Luxembourgish	.01	.01	.02	**.43**	.10	.16	.22	.50	**.52**
Macedonian	.11	.29	.14	**.63**	.28	.28	.47	**.69**	.61
Malagasy	.18	.15	.12	**.34**	.18	.20	.19	.24	**.46**
Malay	.43	.30	.28	**.57**	**.57**	.54	.47	**.64**	.57
Malayalam	.05	.04	.06	.13	.11	.08	**.31**	.40	**.46**
Maltese	.19	.16	.19	**.38**	.20	.15	.27	**.49**	.44
Maori	.16	.04	.10	**.32**	.16	.19	.18	.30	**.43**
Marathi	.28	.31	.28	**.49**	.40	.35	.45	.52	**.58**
Mongolian	.06	.10	.16	**.38**	.13	.10	.19	.51	**.71**
Nepali	.27	.37	.31	.21	.39	.33	**.41**	.48	**.59**
Nyanja	.19	.18	.11	**.33**	.17	.24	.12	.45	**.48**
Persian	.50	.56	.48	.32	.56	.49	**.60**	**.57**	.47
Polish	.45	.53	.38	**.66**	.62	.56	.54	**.73**	.72
Punjabi	.10	.04	.03	.02	.10	.11	**.24**	.15	**.19**
Romanian	.51	.34	.27	.57	.58	**.59**	.36	**.63**	.56
Russian	.36	.51	.37	**.65**	.44	.53	.57	.67	**.75**

Table 10: XESE scores on an extended set of languages using prompts with instructions in target language.

Language	L2-7B	M-7B	TI-13B	G-7B	L2-13B	L2-70B	MX-8x7B	GPT-3.5	GPT-4o
Samoan	.10	.10	.06	**.29**	.14	.12	.11	.08	**.24**
Serbian	.04	.43	.31	.00	.12	.11	**.49**	.06	**.66**
Shona	.21	.19	.08	**.38**	.26	.25	.22	.43	**.61**
Sinhala	.14	.06	.06	**.45**	.12	.11	.24	.16	**.30**
Slovak	.20	.31	.26	**.67**	.45	.39	.39	**.72**	.64
Slovenian	.41	.40	.31	.56	.57	**.60**	.47	.60	**.65**
Somali	.20	.14	.15	**.33**	.19	.23	.13	.17	**.45**
Swahili	.20	.18	.08	**.62**	.35	.37	.29	**.69**	.63
Swedish	.33	.35	.30	**.67**	.63	.59	.45	**.62**	.56
Tagalog	.16	.10	.15	**.50**	.37	.23	.25	.50	**.54**
Tajik	.07	.15	.18	**.29**	.09	.08	.28	**.61**	.60
Tamil	.19	.22	.10	**.60**	.23	.21	.44	.32	**.42**
Tatar	.04	.08	.12	.10	.07	.06	**.23**	.42	**.68**
Telugu	.09	.09	.02	.22	.06	.09	**.30**	.15	**.18**
Thai	.37	.40	.37	**.61**	.41	.40	.52	.60	**.69**
Tibetan	.01	.12	.07	.02	.02	.01	**.16**	.00	**.13**
Turkish	.56	.46	.45	.52	**.61**	.59	.54	**.70**	.69
Turkmen	.19	.16	.16	**.40**	.29	.20	.30	.41	**.70**
Uighur	.06	.10	.11	.02	.09	.04	**.21**	.37	**.44**
Ukrainian	.36	.51	.36	**.67**	.50	.51	.57	.71	**.74**
Urdu	.43	.44	.41	.02	.43	.43	**.56**	**.46**	.33
Uzbek	.23	.26	.20	**.53**	.30	.27	.31	.61	**.66**
Vietnamese	.50	.17	.41	.50	**.55**	.49	.42	**.49**	.47
Welsh	.07	.04	.11	**.45**	.23	.19	.15	**.52**	.41
Wolof	.05	.04	.03	**.16**	.07	.05	.08	**.16**	.02
Xhosa	.20	.18	.06	**.45**	.22	.19	.23	.30	**.56**
Yoruba	.04	.17	.17	**.29**	.11	.09	.16	.28	**.32**
Zulu	.25	.21	.21	**.48**	.28	.23	.21	.45	**.60**

Table 11: XESE scores on an extended set of languages using prompts with instructions in target language.