ChatGPT MT: Competitive for High- (but not Low-) Resource Languages

Nathaniel R. Robinson[1,2*] **Perez Ogayo**[1*] **David R. Mortensen**[1] **Graham Neubig**[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Center for Language and Speech Processing, Johns Hopkins University
nrobin38@jhu.edu, {aogayo, dmortens, gneubig}@cs.cmu.edu
* Authors contributed equally

Abstract

Large language models (LLMs) implicitly learn to perform a range of language tasks, including machine translation (MT). Previous studies explore aspects of LLMs' MT capabilities. However, there exist a wide variety of languages for which recent LLM MT performance has never before been evaluated. Without published experimental evidence on the matter, it is difficult for speakers of the world's diverse languages to know how and whether they can use LLMs for their languages. We present the first experimental evidence for an expansive set of 204 languages, along with MT cost analysis, using the FLORES-200 benchmark. Trends reveal that GPT models approach or exceed traditional MT model performance for some high-resource languages (HRLs) but consistently lag for low-resource languages (LRLs), under-performing traditional MT for 84.1% of languages we covered. Our analysis reveals that a language's resource level is the most important feature in determining ChatGPT's relative ability to translate it, and suggests that ChatGPT is especially disadvantaged for LRLs and African languages.

1 Introduction

Despite the majority of the world's languages being low-resource, current MT systems still perform poorly on them or do not include them at all. Some commercial systems like Google Translate[1] support a number of LRLs, but many systems do not support any, and in either case the majority of LRLs are largely neglected in language technologies.

In recent years, generative LLMs have shown increasingly impressive translation abilities (Radford et al., 2019; Brown et al., 2020). Even more recently, LLM tools like ChatGPT have become popular and accessible to end users. This marks an important shift, since a majority of LLM users are now consumers rather than researchers. The prospect of LLM translation is exciting, since theoretically, generative LLMs could support more languages than commercial systems like Google's.[2] But only beginning steps have been made to test this hypothesis. While some studies outlined in §4 have evaluated MT with recent LLMs, evaluation is still lacking for many languages. This brings up important questions, such as: *Can end users in need of MT for a variety of languages use ChatGPT? Are ChatGPT and other LLMs reliable translators? For which languages are they reliable?* Initially we hypothesize that LLMs translate HRLs better than LRLs. But due to limited information about the training data and methods for powerful LLMs like ChatGPT (**GPT-3.5** and variants) and GPT-4, hypotheses like this must be experimentally verified.

We significantly expand experimental verification for such hypotheses by testing ChatGPT's performance on the FLORES-200 benchmark (NLLB Team et al., 2022), containing 204 language varieties. We emphasize that, rather than optimizing LLM MT for a few languages, we focus on helping end users of various language communities know how and when to use LLM MT. We expect that our contributions may benefit both direct end users, such as LRL speakers in need of translation, and indirect users, such as researchers of LRL translation considering ChatGPT to enhance specialized MT systems. In summary, we contribute:

1. MT scores on 203 languages for ChatGPT and comparisons with GPT-4, Google Translate, and NLLB (NLLB Team et al., 2022)
2. Evidence that LLMs are competitive with traditional MT models for many HRLs but lag for LRLs (with baselines outperforming ChatGPT on 84.1% of languages evaluated)
3. Evidence that few-shot prompts offer

[1]https://translate.google.com

[2]Google Translate currently supports only 133 languages with systems deemed high enough quality for deployment.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 392–418
December 6–7, 2023.

marginal benefits for LLM translation

4. A decision tree analysis of language features' correlation with LLM effectiveness in MT, suggesting ChatGPT is especially disadvantaged for LRLs and African languages

5. A cost comparison across MT systems

Our experiments are motivated by the interests of LLM users speaking a variety of languages. In addition to evaluating a large language set (§3), we chose to analyse language features (§3.4), to draw generalizations for even more LRL speakers. We compare MT costs because they impact end users (§3.7). We keep ChatGPT central to our analyses because of its current popularity among consumers.

2 Methodology

We used data for 204 language varieties from FLORES-200 (NLLB Team et al., 2022). We used the 1012 *devtest* sentences for our main experiments and the 997 *dev* sentences for follow-up experiments. We queried the OpenAI API[3] to translate our test set from English into the target languages. We explored ENG→X translation only because the FLORES-200 English data was taken from Wikipedia. Thus OpenAI's GPT models were likely trained on those exact English sentences, making fair X→ENG evaluation infeasible.

2.1 Experimental setup

We evaluated ChatGPT's (`gpt-3.5-turbo`) MT for our full language set. We compared with NLLB-MOE (NLLB Team et al., 2022) as our baseline, as it is the current state-of-the-art open-source MT model that covers such a wide variety of languages. NLLB is a discriminative transformer trained on supervised bi-text data (the traditional MT paradigm). We obtained scores for NLLB outputs of ENG→X translation into 201 of the language varieties in our set (as reported by NLLB Team et al. (2022)).

We used both zero- and five-shot prompts for ChatGPT MT. (See §2.3.) Previous studies (Hendy et al., 2023; Gao et al., 2023; Moslem et al., 2023; Brown et al., 2020; Zhu et al., 2023) suggest that few-shot prompts produce slightly (albeit not consistently) better translations. But zero-shot prompts are more convenient and affordable for users.

We also compare with results for subsets of our selected languages from two other MT engines. Google Translate API was an important baseline for our analysis because it is popular among end users. We also included it to represent commercial MT systems in our study. Because Google's API does not support all 204 of the FLORES-200 languages, we obtained results only for the 115 non-English languages it supports.

Lastly, we obtained MT results from GPT-4, since it is a popular LLM and has been shown to outperform ChatGPT on MT (Jiao et al., 2023; Wang et al., 2023). Because the cost of GPT-4 use exceeds that of ChatGPT by 1900%, our resources did not permit its evaluation on all 203 non-English languages. Instead we selected a 20-language subset by picking approximately every 10th language, with languages sorted by chrF++ differentials between ChatGPT and NLLB ($chrf_{GPT} - chrf_{NLLB}$). We chose this criterion in order to have 20 languages with a range of relative ChatGPT performance and a variety of resource levels. We used only five-shot prompts for GPT-4.

2.2 Implementation details

We conducted all LLM experiments with `gpt-3.5-turbo` (ChatGPT) and `gpt-4-0613` (GPT-4). We used `top_p` 1, `temperature` 0.3, `context_length` −1, and `max_tokens`[4] 500.

To evaluate the outputs, we used:[5]

spBLEU: BLEU (Papineni et al., 2002) is standard in MT evaluation. We find spBLEU scores (Goyal et al., 2022) via sacreBLEU (Post, 2018) with the SPM-200 tokenizer (NLLB Team et al., 2022).

chrF2++: We use sacreBLEU's implementation of chrF++ (Popović, 2017). We adopt it as our main metric, as it overcomes some of BLEU's weaknesses, and refer to it as *chrF* for brevity.

2.3 Zero- and few-shot prompts

Previous works (Gao et al., 2023; Jiao et al., 2023) investigated LLM prompting to optimize MT performance. We adopt Gao et al. (2023)'s recommended prompts for both zero- and few-shot MT (Table 1). We are interested in multiple n-shot prompt settings because, as mentioned in §2.1, they

[3] `https://platform.openai.com`

[4] Although some languages had higher token counts than others (see §3.4), we found that adjusting `max_tokens` had a minimal effect on MT performance. We thus decided to maintain the same value of `max_tokens` across all languages for experimental consistency.

[5] We excluded learned MT metrics like COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020), since they do not support many LRLs.

Shot	Prompt
zero	This is an English to [TGT] translation, please provide the [TGT] translation for this sentence. Do not provide any explanations or text apart from the translation. [SRC]: [src-sentence] [TGT]:
five	This is an English to [TGT] translation, please provide the [TGT] translation for these sentences: [SRC]: [src-sentence] [TGT]: [tgt-sentence] [SRC]: [src-sentence] [TGT]: [tgt-sentence] [SRC]: [src-sentence] [TGT]: [tgt-sentence] [SRC]: [src-sentence] [TGT]: [tgt-sentence] [SRC]: [src-sentence] [TGT]: [tgt-sentence] Please provide the translation for the following sentence. Do not provide any explanations or text apart from the translation. [SRC]: [src-sentence] [TGT]:

Table 1: Prompts used for zero- and five-shot settings

present different benefits to LLM users. We explored zero-shot (no in-context example), one-shot (1 example), and five-shot (5 examples). We employed both zero- and five-shot prompts in our main experiments over 203 languages, and we analyzed all three n-shot settings for a subset of languages on FLORES-200 *dev* sets.

The languages in FLORES-200 represent 22 language families. To experiment with multiple n-shot settings, we selected one language from each of the 12 families containing at least two members in the set. We chose four HRLs ($\geq$1M Wikipedia pages[6]), four LRLs (25K-1M pages), and four extremely LRLs ($\leq$25K pages). These languages also employ a variety of scripts. See Table 2.

Language	Code	Family	Script	Wiki. #
French	fra	Indo-European	Latn	12.7M
Chinese	zho	Sino-Tibetan	Hans	7.48M
Turkish	tur	Turkic	Latn	2.48M
Finnish	fin	Uralic	Latn	1.46M
Tamil	tam	Dravidian	Taml	496K
Tagalog	tgl	Austronesian	Latn	239K
Kiswahili	swh	Niger-Congo	Latn	167K
Amharic	amh	Afroasiatic	Ethi	46.2K
Santali	sat	Austroasiatic	Olck	20.0K
Lao	lao	Kra-Dai	Laoo	14.0K
Papiamento	pap	Creole	Latn	6.84K
Luo	luo	Nilo-Saharan	Latn	0

Table 2: Diverse subset of languages experiments with few-shot settings. **Wiki. #** is the number of Wikipedia pages in the language.

[6]Throughout the paper we use the "Total pages" count from https://en.wikipedia.org/wiki/List_of_Wikipedias, accessed 7 August 2023, as a proxy for the resource level of a language.

	#langs.	avg. chrF	avg. BLEU
ChatGPT (0-shot)	203	32.3	16.7
ChatGPT (5-shot)	203	33.1	17.3
GPT-4	20	44.6	24.6
NLLB	201	45.3	27.1
Google	115	**52.2**	**34.6**

Table 3: Languages evaluated, average chrF, and average BLEU for each MT system. Best scores are **bold**.

3 Results and Analysis

3.1 Traditional MT generally beats LLMs

Table 3 shows the number of languages we evaluated for each MT system, as noted in §2.1, with average chrF and BLEU scores across those languages. The best performing model on average was (1) Google, then (2) NLLB, (3) GPT-4, and (4) ChatGPT. Unabridged results are in Table 11 in Appendix A. Supplementary materials can also be browsed on our repository.[7] (Also see the interactive score visualizer on our Zeno browser.[8])

Table 4 shows chrF for the 20 languages evaluated on both LLM systems. Of the 11 languages evaluated on all four systems, Google performed best for 9 of them. Notably, GPT-4 surpassed NLLB in five languages and Google in one[9] (Mesopotamian Arabic, acm_Arab).

On the 20 languages for which we tested it, GPT-4 improved over ChatGPT by 6.5 chrF on average. The standard deviation of performance difference with NLLB ($chrF_{GPT} - chrF_{NLLB}$) was 8.6 for GPT-4, compared with ChatGPT's 12.7 for the same languages, suggesting a more consistent advantage across language directions. GPT-4 offered larger improvements for LRLs, whereas HRL performance plateaued between the LLMs. Previous studies have found GPT-4 improving multilingual capabilities over ChatGPT on a range of tasks (Xu et al., 2023; Zhang et al., 2023; OpenAI, 2023). This may account for its superior MT performance.

Google Translate outperformed all other systems in chrF on 100 of the 115 languages for which we evaluated it, with an average improvement of 2.0 chrF points over the next best system for each language. (See Appendix A for unabridged results.)

[7]https://github.com/cmu-llab/gpt_mt_benchmark

[8]https://hub.zenoml.com/project/cabreraalex/GPT%20MT%20Benchmark

[9]Our language identification analysis in §3.6 and manual inspection suggest that GPT models only output one Arabic variety: Modern Standard Arabic (MSA). It seems the LLMs' high performance on some Arabic varieties is due simply to incidental high token overlap with MSA targets.

Lang.	GPT-4	ChatGPT	Google	NLLB
ssw_Latn	24.1	6.7	-	43.3
sna_Latn	29.2	16.3	**44.4**	43.4
ckb_Arab	33.1	24.8	**47.7**	47.2
mag_Deva	44.6	39.9	-	58.5
ibo_Latn	27.7	16.3	**43.5**	41.4
hau_Latn	40.3	22.4	53.2	**53.5**
pbt_Arab	26.7	21.1	-	39.4
tam_Taml	42.7	34.5	**55.8**	53.7
kat_Geor	41.4	33.5	**51.4**	48.1
gle_Latn	53.0	47.5	**60.1**	58.0
kmr_Latn	34.3	27.4	**40.0**	39.3
war_Latn	54.0	49.5	-	57.4
ajp_Arab	48.4	47.5	-	51.3
lim_Latn	45.1	42.7	-	47.9
ukr_Cyrl	56.3	55.4	**58.6**	56.3
fra_Latn	71.7	71.3	**72.7**	69.7
lvs_Latn	57.3	55.2	-	54.8
ron_Latn	**65.3**	64.2	65.0	61.3
tpi_Latn	**49.5**	39.2	-	41.6
acm_Arab	**46.5**	46.1	-	31.9

Table 4: chrF (↑) scores across models for all languages we used to evaluate GPT-4. Best scores are **bold**. ChatGPT scores here are 5-shot, to compare with GPT-4.

Google's was the best performing MT system overall, though NLLB has broader language coverage.

NLLB outperformed ChatGPT in chrF on 169 (84.1%) of the 201 languages for which we obtained scores for both, with NLLB scoring an average of 11.9 chrF points higher than the better n-shot ChatGPT setting for each language. This trend is corroborated by Zhu et al. (2023). Table 5 has both BLEU and chrF scores from both systems for the five languages with the most negative chrF deltas ($chrF_{GPT} - chrF_{NLLB}$) on top, followed by the five languages with the highest positive deltas on bottom. For many of the subsequent sections of this paper we focus on comparing ChatGPT and NLLB, since we evaluted them on the most languages.

	ChatGPT		NLLB	
Lang.	**BLEU**	**chrF**	**BLEU**	**chrF**
srp_Cyrl	1.36	3.26	**43.4**	**59.7**
kon_Latn	0.94	8.50	**18.9**	**45.3**
tso_Latn	2.92	15.0	**26.7**	**50.0**
kac_Latn	0.04	2.95	**14.3**	**37.5**
nso_Latn	3.69	16.7	**26.5**	**50.8**
jpn_Jpan	**28.4**	**32.9**	20.1	27.9
nno_Latn	**37.1**	**58.7**	33.4	53.6
zho_Hans	**36.3**	**31.0**	26.6	22.8
zho_Hant	**26.0**	**24.4**	12.4	14.0
acm_Arab	**28.2**	**44.7**	11.8	31.9

Table 5: Lowest (top) and highest (bottom) chrF differences between zero-shot ChatGPT and NLLB. Best scores for each metric in **bold** (with BLEU **blue**).

3.2 ChatGPT under-performs for LRL

Using NLLB Team et al.'s (2022) resource categorization, we find that ChatGPT performs worse on LRLs than HRLs, corroborating findings of previous works (Jiao et al., 2023; Zhu et al., 2023). There is a strong positive correlation between ChatGPT and NLLB chrF scores, but the correlation is higher for HRLs (ρ=0.85) than LRLs (ρ=0.78), indicating that ChatGPT struggles to keep up with NLLB for LRLs.

Figure 1 shows scatter plots where dots represent languages, with ChatGPT's (positive or negative) *relative improvement* over NLLB chrF ($\frac{chrf_{GPT} - chrf_{NLLB}}{chrf_{NLLB}}$) on the y-axis. When languages are grouped by family or script, some trends are apparent (in part because we ordered groups by descending average scores). For example, ChatGPT fairs better with Uralic and Indo-European languages and clearly worse with Niger-Congo and Nilo-Saharan languages. However, the clearest natural correlation appears when languages are grouped by resource level, approximated by number of Wikipedia pages (Figure 1, bottom). Note the *relative improvement* (y-axis) is typically negative since ChatGPT rarely outperformed NLLB.

In the five-shot setting, ChatGPT outperformed NLLB on 47% of the HRLs designated by NLLB Team et al. (2022), but only on 6% of the LRLs. These findings contrast with what is commonly observed in multilingual MT models (Liu et al., 2020; Fan et al., 2020; Siddhant et al., 2022; Bapna et al., 2022; NLLB Team et al., 2022), where LRLs benefit the most. This highlights the need to investigate how decoder-only models may catch up with encoder-decoder models in low-resource applications. It underscores the importance of MT-specialized models when larger multitask models cannot overcome low-resource challenges.

3.3 Few-shot prompts offer marginal improvement

Our main experiments suggested that n-shot setting had only a modest effect on MT performance. We conducted a more concentrated study of n-shot prompts using dev sets for the 12 languages in Table 2. Results in Table 6 show five-shot prompts performing best. For some LRLs, this was simply a result of ChatGPT's failure to model the language. In Santali's case, for example, zero-shot ChatGPT was unable to produce the Ol Chiki script at all. In the five-shot setting, it was able to imitate the script

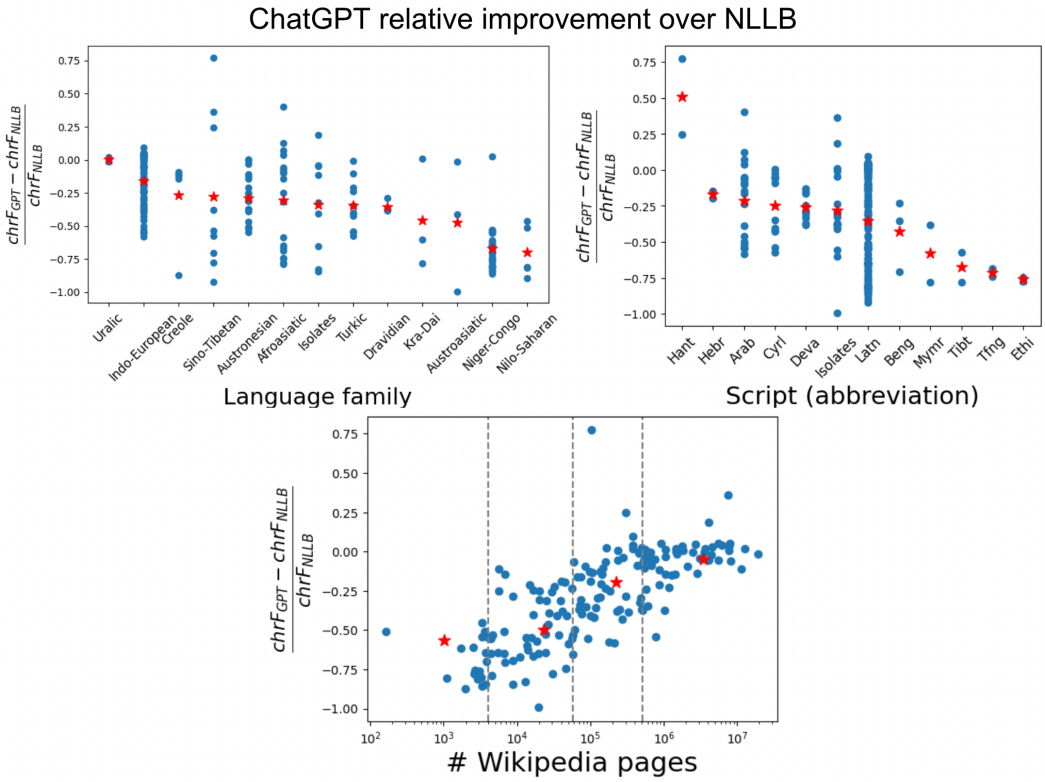

Figure 1: ChatGPT *relative improvement* over NLLB chrF, with languages organized by family, script, and number of Wikipedia pages. Red stars represent averages per group. In the bottom plot, languages are grouped into quartiles of equal size (with dotted lines at the Q1, median, and Q3). More expansive visualizations with language labels for each value can be found in Appendix C.

characters from the context, but without any coherence or accuracy. Excepting Santali as an outlier, five-shot settings offered generally marginal improvements over zero-shot (the most cost-effective of the settings), with an average improvement of only 1.41 chrF across all 12 languages (0.31 if we exclude Santali). Zero-shot prompts actually produced the best chrF score for six of the 12 languages. The one-shot setting performed worst. We noted this trend of few-shot contexts offering only meager and inconsistent improvements throughout our experiments, with five-shot MT improving on zero-shot by only 0.88 average chrF across all 203 language directions. (See Appendix A.)

3.4 Importance of language features

We were interested in which language features determined LLMs' effectiveness compared to traditional MT. Analyzing this may reveal trends helpful to end users deciding which MT system to use, especially if their language is not represented here but shares some of the features we consider. In this section we focus on comparing ChatGPT and NLLB, since we evaluated the most languages with them. We focus on zero-shot ChatGPT, as it is the most common and convenient setting for end users.

	0-shot		1-shot		5-shot	
	BLEU	chrF	BLEU	chrF	BLEU	chrF
fra	55.4	**71.3**	50.4	70.3	55.4	71.2
zho	30.0	29.9	28.2	30.8	**30.7**	**31.1**
fin	34.6	56.6	31.7	56.3	34.6	**56.7**
tur	38.2	**58.6**	34.8	57.6	**38.3**	58.6
tgl	35.9	**60.2**	35.2	59.6	**36.1**	60.1
tam	**13.8**	**35.3**	11.7	34.3	11.9	34.6
swh	39.7	**60.6**	36.0	59.5	**40.0**	60.5
amh	3.4	10.1	3.2	9.6	**3.9**	**10.6**
pap	26.6	51.5	29.3	54.1	**34.8**	**56.1**
lao	4.8	21.6	4.4	20.8	**5.3**	**22.1**
luo	**0.8**	**7.6**	0.2	4.6	0.2	5.2
sat	0.0	0.3	2.2	11.3	**3.0**	**13.8**

Table 6: Three n-shot settings for 12 diverse languages

We encoded each of the 203 languages in our set as a *feature vector*. In these language *feature vectors* we included **four numerical features**: number of Wikipedia pages in the language (`wiki_ct`), size of the language's bi-text corpus in the Oscar MT database[10] (`oscar_ct`) (Abadji et al., 2022), percentage of ASCII characters[11] in the FLORES-

[10] https://oscar-project.org

[11] Percentage of characters with an encoding between 0 and

200 *dev* set for the language (`ascii_percentage`), and average number of tokens per *dev* set sentence in FLORES-200 with ChatGPT's tokenizer (`token_ct`). We also included **two categorical features**: language family (`family`) and script the language was written in (`script`); and **one binary feature**: the FLORES resource designation of the language—with 1 for high-resource and 0 for low-resource (`hi/lo`). Before analysis, we one-hot encoded the two **categorical features** into 48 binary features like `family_Niger-Congo` and `script_Latn`.

We selected `token_ct` as a feature because we observed languages in low-resource scripts having many tokens. For example, ChatGPT's tokenizer encodes multiple tokens for every character in Ol Chiki script. This tendency for GPT models with low-resource scripts has been noted in previous studies (Ahia et al., 2023).

We fit a decision tree with these *feature vectors* to regress on ChatGPT's *relative improvement* over NLLB in chrF ($\frac{chrf_{GPT} - chrf_{NLLB}}{chrf_{NLLB}}$), for each of the 201 languages with NLLB scores. When we used `max_depth` 3, the tree in Figure 2 was learned. Languages are delimited first by `wiki_ct`; then LRLs are separated into Niger-Congo languages and others, while HRLs are delimited by `token_ct`. The only group where ChatGPT beat NLLB is of languages with more than 58,344 Wikipedia pages, fewer than 86 tokens per average sentence, and less than 15.5% ASCII characters. This group contains some East Asian HRLs. The group where ChatGPT was least advantaged contains Niger-Congo languages with fewer than 3,707 Wikipedia pages.

We also fit a random forest regressor with the same features and labels to find feature importance values. Only ten features had importance ≥ 0.01, shown in Table 7. The most important feature by far was `wiki_ct`. (This feature correlates strongly with ChatGPT's *relative improvement*, $\rho = 0.68$.) `family_Niger-Congo` was much more important than any other family feature. No script feature had an importance exceeding 0.01. In general, features for resource level and tokenization were more important than family or script.

ChatGPT has a blind spot not only for Niger-Congo languages, but for African languages in general. Figure 1 shows ChatGPT is least advantaged for the two exclusively African families, Niger-Congo and Nilo-Saharan; and the two exclusively African scripts, Tifinagh (`Tfng`) and Ge'ez (`Ethi`).

feature	importance
`wiki_ct`	0.514
`token_ct`	0.157
`ascii_percentage`	0.104
`family_Niger-Congo`	0.054
`oscar_ct`	0.040
`family_Afroasiatic`	0.025
`family_Indo-European`	0.025
`family_Sino-Tibetan`	0.022
`family_Creole`	0.012
`family_Nilo-Saharan`	0.011

Table 7: Ten most important language features to predict ChatGPT's effectiveness relative to NLLB

3.5 Impact of script

Prior research suggests that ChatGPT output quality is sensitive to language script (Bang et al., 2023). Our own analysis in §3.4 actually suggests that script is the least important language feature in predicting ChatGPT's MT effectiveness. However, differences in performance are clear when comparing scripts used for the same language. Table 8 shows one script typically outperforming the other, by an average of 14.3 chrF points for zero-shot. Five-shot contexts narrowed the gap slightly to 12.0. Although transliteration is a deterministic process for many languages, these performance gaps suggest that ChatGPT has not implicitly learned it as part of a translation task. We hypothesize that ChatGPT's observed sensitivity to script in earlier studies may be particular to the languages and tasks evaluated.

	BLEU		chrF	
Lang.	0-shot	5-shot	0-shot	5-shot
ace_Arab	1.27	2.26	8.41	9.75
ace_Latn	**4.98**	4.35	**19.82**	17.96
arb_Arab	37.60	**37.85**	53.79	**53.81**
arb_Latn	5.33	8.38	22.79	26.92
bjn_Arab	1.96	3.05	10.43	13.24
bjn_Latn	10.96	**12.29**	35.92	**37.98**
kas_Arab	**3.99**	3.30	**15.51**	14.33
kas_Deva	2.31	2.68	12.91	13.91
knc_Arab	0.51	1.06	5.26	4.67
knc_Latn	**2.61**	0.91	**13.38**	8.11
min_Arab	1.56	3.49	10.06	14.88
min_Latn	11.51	**13.07**	36.99	**38.43**
taq_Latn	0.82	0.28	8.18	6.24
taq_Tfng	0.62	**1.37**	5.23	**8.31**
zho_Hans	36.33	**36.51**	31.03	**31.89**
zho_Hant	29.30	30.38	24.82	26.02

Table 8: ChatGPT performance on languages with multiple scripts. Each better scoring script is **bold**.

128, inclusive, using the Python built-in `ord` function

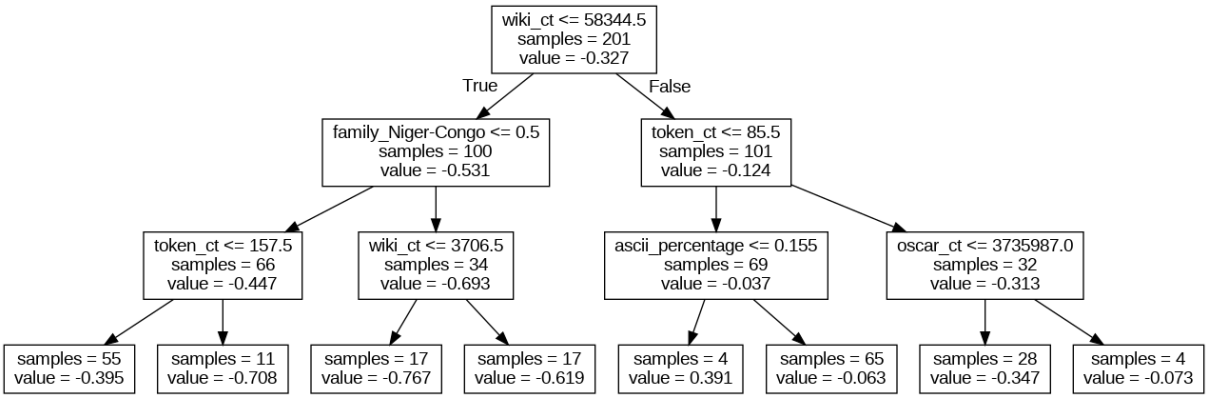

Figure 2: Decision tree predicting ChatGPT *relative improvement* over NLLB chrF, from language features.

3.6 LLMs often get the language wrong

LLMs' performing worse than NLLB may be due in large part to their translating into the wrong language. Using FLORES-200's *dev* data, we trained a logistic regression language identifier for 100 epochs. Language identification accuracies for four of the models we evaluated are in Table 9. Zero-shot ChatGPT only translated on target 72% of the time. This expectedly improved with five-shot prompts, and GPT-4 performed even better, still just shy of NLLB. LLMs' tendency to translate off target is corroborated by Zhu et al. (2023).

model	lang. ID acc.
ChatGPT (0-shot)	72%
ChatGPT (5-shot)	83%
GPT-4 (5-shot)	90%
NLLB	91%

Table 9: Proportion of the time each model translated into the correct target language

3.7 Cost comparison

Our results suggest that GPT-4 is a better translator than ChatGPT. However in considering the needs of MT end users, it would be remiss not to consider the respective costs of the systems evaluated. GPT-4's high cost (roughly 2000% that of ChatGPT's) prohibited us from evaluating it on all FLORES-200 languages. In general, using few-shot prompts for LLMs is more costly than zero-shot prompts, since users are charged for both input and output tokens. And for this same reason, some languages are more costly than others in LLM MT. Previous work has found that Google Translate has associated costs comparable to those of five-shot ChatGPT (Neubig and He, 2023). NLLB is the least expensive system we evaluated.

We estimated cost values for each MT system and language: the expense, in USD, of translating the full FLORES-200 *devtest* English set into the language. We estimated GPT model costs using the prompts employed in our experiments, the tiktoken tokenizer[12] used by both models, and inference prices posted by OpenAI.[13] Conveniently, Google Translate costs nothing for the first 500K input characters. But since frequent MT users may have already expended this allowance, we calculated costs from their rates beyond the first 500K.[14] As the NLLB-MOE model (54.5B parameters) is difficult to run on standard computing devices, NLLB Team et al. (2022) also provided a version with only 3.3B parameters that achieves similar performance. Since users commonly opt for the smaller model, and since the performance difference does not impact our estimates significantly, we estimated the costs to run the 3.3B-parameter NLLB model using a single GPU on Google Colab. Details of our estimation method are in Appendix B.1. Table 10 contains the average cost for each system across the languages we evaluated with it.

model	cost
NLLB	$0.09
ChatGPT (0-shot)	$0.35
ChatGPT (5-shot)	$1.32
Google	$2.66
GPT-4 (5-shot)	$25.93

Table 10: Estimated cost in USD to translate FLORES-200 *devtest* ENG→X with each system, averaged across all languages we evaluated with each

Figure 3 displays chrF scores for the 11 languages on which we evaluated all four MT sys-

[12] https://github.com/openai/tiktoken
[13] https://openai.com/pricing
[14] https://cloud.google.com/translate/pricing

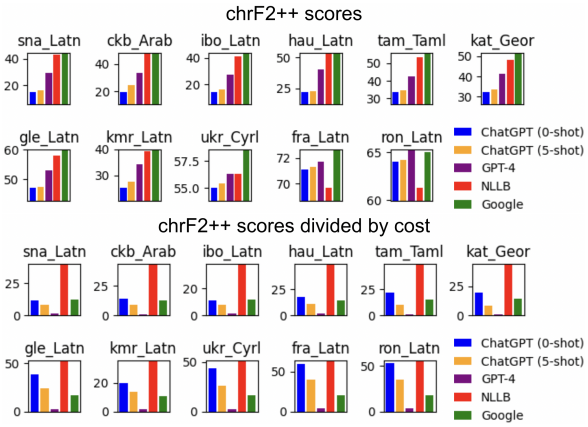

Figure 3: chrF scores for the 11 languages on which we evaluted all MT systems (top), followed by the same scores divided by the estimated cost of each system for each language (bottom)

tems (top), and the same scores divided by the approximate cost of each model (bottom). Bars for GPT-4 drop significantly in the bottom chart because of its high cost. Note from the top chart that Google Translate scores the best, but the bottom chart shows that NLLB has the best scores for its price. Zero-shot ChatGPT also tops five-shot in the bottom chart, suggesting that while few-shot prompts provide modest score improvements, they may not be worth the extra cost. See Appendix B for fuller visualizations with all 203 languages.

4 Related Work

We are not the first researchers to explore LLM MT. However, most existing studies do not provide benchmarks for a large number or languages. Wang et al. (2023) studied GPT model discourse MT, but only for four languages. Gao et al. (2023) studied prompt engineering for GPT model MT, a helpful precursor to our work, but only for three languages. Moslem et al. (2023) probed the abilities of GPT models for adaptive and domain-appropriate MT and term extraction, only including six languages in five directions. Jiao et al. (2023) produced MT benchmarks for ChatGPT and GPT-4, but only for five languages, none of them LRLs.[15] They corroborated our findings that GPT models lag behind traditional MT models, but that GPT-4 outperforms ChatGPT. Hendy et al. (2023) explored 18 language pairs in a similar study, including four LRLs, but they focused more on MT performance across text domains, in-context learning, and reasoning than on multilingual benchmarks.

In all the heretofore mentioned works combined, researchers explored only 18 languages, including five LRLs. This few-language approach does not address the needs of LLM users seeking to translate any languages other than the small few represented. In a work most comparable to our own, Zhu et al. (2023) attempted to address this issue. They provided benchmarks comparing LLMs and traditional MT models across 102 languages, including 68 LRLs. Their results corroborate our own conclusions that LLMs lag behind traditional MT models, especially for LRLs. However, their analysis focuses primarily on few-shot learning and prompt engineering, including some topics somewhat removed from end user needs (such as the viability of nonsensical prompts in few-shot settings). Our work differs from existing studies in our focus on end users. We include more languages than any existing work (**204** languages, including **168** LRLs), to address the needs of various LRL communities. Our analysis suggests which language features predict LLM effectiveness, to help end users make hypotheses even about languages not represented in our study. We evaluate monetary costs, since they are a concern for LLM users.

5 Conclusion

We provide benchmarks for LLM ENG→X MT performance across 203 languages, with comparisons to state-of-the-art commercial and open-source MT models. For many HRLs, LLMs like ChatGPT perform competitively with these traditional models. But for LRLs, traditional MT remains dominant, despite LLMs' increased parameter size. Our decision-tree analysis reveals language features that predict ChatGPT's translation effectiveness relative to NLLB, finding that ChatGPT is especially disadvantaged for LRLs and African languages, and that the number of Wikipedia pages a language has is a strong predictor of ChatGPT's effectiveness in it. We present evidence that few-shot learning offers generally marginal improvements for ENG→X MT, which may not justify its additional cost. We provide MT users with scores and cost estimates for four LLM and traditional MT systems, to help them determine which to use for their languages.

Future work may include more translation directions (X→ENG and non-English-centric), document-level MT, and human evaluation of LLM

[15] In this section, we define LRLs as languages having fewer than 1M Wikipedia pages.

outputs to reveal trends along fluency and accuracy dimensions. We open-source software and outputs of the models we evaluated on our repository.

Limitations

We acknowledge limitations of using ChatGPT models for research. Since they are closed-source models, there is much we do not know about their architectural and training details, which can impact our understanding of their capabilities and biases. For instance, OpenAI's implementation of mechanisms to prevent the generation of harmful or toxic content may inadvertently impact the quality of the model's output. This can be a concern when evaluating the reliability and accuracy of the results. OpenAI continuously updates and deprecates models behind the ChatGPT API, so our assessment may not be immaculate for future versions. Future work may mitigate these concerns by evaluating white-box LLMs, such as BLOOM (Scao et al., 2022) or MPT (Team, 2023), or LLMs not tuned for instruction, like GPT-3 (Brown et al., 2020).

While FLORES-200 is large and diverse, it is likely not representative of the vast array of languages worldwide. Some low-resource sets within FLORES-200 may contain noisy or corrupted data, potentially affecting the validity of the automatic metrics we employ in our reporting of scores. Additionally, FLORES-200 sets were translated from English Wikipedia. We avoided any X→ENG translation directions, since it is likely that GPT models were trained on English Wikipedia. However, the semantic proximity of the other language sets to the original English source could potentially provide an advantage to these models in generating them. We also acknowledge the absence of non-English-centric translation directions from this study; we leave this for future work.

Lastly, the unavailability of semantic MT evaluation techniques like COMET (Rei et al., 2020) or BLEURT (Sellam et al., 2020) for LRLs hinders our ability to conduct comprehensive semantic evaluations and may leave some aspects of the translation quality unexplored. Future researchers may gain additional insights by evaluating LLM COMET scores for the target languages in which they are available. Human evaluation (which we leave for future work) may also reveal much in this area. These limitations surrounding model transparency, representative data, and evaluation should be taken into account when interpreting the findings of this work. Future studies may benefit from addressing these challenges to enhance the robustness and reliability of MT conclusions.

Ethics Statement

The new prominence of LLMs in language technologies has numerous ethical implications. This study makes it apparent that even powerful LLMs like ChatGPT have significant limitations, such as an inability to translate a large number of low-resource languages. It also suggests that although these LLMs are trained on large and diverse data sets, they still have implicit biases, such as a clear disadvantage in MT for African languages. We hope to stress the importance of acknowledging and publicizing the limits and biases of these LLMs. This is especially relevant because a majority of LLM users may not be familiar or experienced with artificial intelligence (AI) engineering practices, and the commercial entities providing LLMs often have a monetary incentive to deliberately downplay the models' limitations. This can lead to unethical exploitation of users, who may attempt to use LLMs in applications where their limitations and biases can cause harm. Part of our goal in this work is to bring these discussions to the forefront of AI research. Ethical considerations like these should be a top concern for AI researchers, especially when many recent AI advancements are piloted by powerful commercial corporations.

We hope also to acknowledge some of the ethical considerations involved in our own research. As we strive to develop improved open-source and accessible translation systems, it is essential to acknowledge that some language communities may have reservations about having their languages translated. Another crucial point is that utilizing the FLORES-200 test set in this research may inadvertently contribute to its incorporation into OpenAI's training data. OpenAI's current position is that API requests are not used for training (Schade, 2023), but if this position were altered or disregarded, it could compromise the reliability of this test set for future GPT iterations. (This is a consideration for many commercial LLMs, though we only used OpenAI's in the current work.) This scenario has a potential negative impact on the MT community, since many researchers depend on FLORES-200 and other MT benchmarks for large, diverse, high-quality data to conduct system comparisons.

Acknowledgements

We thank Simran Khanuja for her help in running our Google Translate baseline and her general support. We also thank Alex Cabrera for his help developing our Zeno browser. This material is based on research sponsored in part by the Air Force Research Laboratory under agreement number FA8750-19-2-0200. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government. This work was also supported in part by the National Science Foundation under grant #2040926, a grant from the Singapore Defence Science and Technology Agency.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, page arXiv:2201.06642.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study. *arXiv e-prints*, pages arXiv–2304.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Wenxiang Jiao, WX Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.

Graham Neubig and Zhiwei He. 2023. Zeno GPT Machine Translation Report.

NLLB Team, Marta R. 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, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers,

Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI. 2023. Gpt-4 technical report.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Michael Schade. 2023. How your data is used to improve model performance.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *CoRR*, abs/2201.03110.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis.

A Unabridged Result Table

In Table 11 we report full results for 203 target languages in ENG→X translation directions, across four MT systems: two LLMs (ChatGPT and GPT-4, with two n-shot settings for ChatGPT), one open-source encoder-decoder MT model (NLLB), and one commercial system (Google). We order in them in increasing order of performance, with zero-shot ChatGPT performing the worst and Google performing the best overall. We obtained scores for 203 target languages with ChatGPT, 201 with NLLB, 115 with Google Translate, and 20 with GPT-4. Our scores are spBLEU (Goyal et al., 2022) using the SPM-200 tokenizer (NLLB Team et al., 2022) and chrF2++ (Popović, 2017). All results are also available on our repository, and interactive visualizations and histograms can be browsed on our Zeno browser.

B Unabridged Bar Charts and Cost Estimation

See Figures 4 and 5 for chrF and BLEU scores across all MT systems and languages. Google Translate and NLLB are generally the best performers in both metrics, though GPT-4 and ChatGPT are occasionally best. An "x" indicates where we did not evaluate one of the systems for a language. Figures 6 and 7 display chrF and BLEU scores divided by the estimated cost of each MT system. The cost value is measured as the amount in USD that it would cost to translate the entire FLORES-200 *devtest* set for each language.

These visualizations are also available on our repository. (Also see our Zeno browser for interactive visualizations of our results.) We also include cost estimates and scores divided thereby for all languages and MT systems in Table 14. We exclude cost estimates by language for NLLB and Google because there is very little variation between languages. Our estimated cost of translating FLORES-200 *devtest* ENG→ is approximately $0.09 for every target language. And the respective estimate for Google Translate is roughly $2.66 regardless of the target language, since Google's API only charges for input characters.

B.1 Details about estimating NLLB cost

To estimate the cost of running NLLB's 3.3B-parameter model for translation, we used one GPU from Google Colab to translate the full FLORES-200 *devtest* set from English into six languages representing six high- and low-resource scripts–Burmese (`mya_Mymr`), Simplified Chinese (`zho_Hans`), Standard Arabic (`arb_Arab`), Hindi (`hin_Deva`), Armenian (`hye_Armn`), and French (`fra_Latn`)–and measured the time for each. We assumed that runtime t is determined by an equation with unknown coefficients x_1, x_2, and x_3:

$$t = x_1 n_{input} + x_2 n_{output} + x_3 \quad (1)$$

where n_{input} represents the number of input tokens and n_{output} is the number of output tokens. In this case, x_1 represents the rate at which the encoder processes input tokens, x_2 represents the rate at which the decoder undergoes inference, and x_3 is the amount of time to perform all other computations, independent of the number of tokens. We estimated x_1, x_2, and x_3 via a least-squares solution to the linear system defined by the six languages for which we obtained runtime t:

$$\begin{bmatrix} n_{input} & n_{output}(\texttt{mya}) & 1 \\ n_{input} & n_{output}(\texttt{zho}) & 1 \\ n_{input} & n_{output}(\texttt{arb}) & 1 \\ n_{input} & n_{output}(\texttt{hin}) & 1 \\ n_{input} & n_{output}(\texttt{hye}) & 1 \\ n_{input} & n_{output}(\texttt{fra}) & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} t_{\texttt{mya}} \\ t_{\texttt{zho}} \\ t_{\texttt{arb}} \\ t_{\texttt{hin}} \\ t_{\texttt{hye}} \\ t_{\texttt{fra}} \end{bmatrix}$$

where n_{input} is the number of tokens in the English *devtest* set, and n_{output} for each language is the number of tokens in the NLLB-MOE model output provided by NLLB Team et al. (2022). (We used the same tokenizer that we had used for GPT model cost estimation, for simplicity.) After estimating x_1, x_2, and x_3, we used them in Equation 1 to estimate t values for all 201 languages for which we obtained NLLB MT scores. We then used Google Colab's estimated rate of $0.35/hour for use of one GPU to estimate costs for each language.

C Visualizations Comparing ChatGPT and NLLB

See Figures 8 and 9. They are also posted on our repository. (Also see our Zeno browser for interactive visualizations of our results.)

D Estimating Wikipedia Page Counts

As mentioned in §2.3, we used the "Total pages" count from `https://en.wikipedia.org/wiki/List_of_Wikipedias`, accessed 7 August 2023,

as a proxy for the resource level of a language (refered to as `wiki_ct` in §3.4). We had to make some decisions regarding macrolanguage and microlanguage matches when making these estimates. Many of the languages in FLORES-200 (NLLB Team et al., 2022) are in fact microlanguages of a macrolangauge not included in the dataset. In some cases this microlanguage was did not have a listed Wikipedia page count, so we used the macrolanguage page count instead. Table 12 lists all the languages for which we used the Wikipedia page count of a macrolanguage (with a different ISO 639-3 code), based on our best judgment. In every case this was because the FLORES-200 microlanguage was not listed.

There were also cases where we decided to list zero for a microlanguage's `wiki_ct`, even if its macrolanguage was listed with a nonzero number of pages. This was in cases where we could reasonably assume that the macrolanguage's Wikipedia pages were likely (either all or predominantly) in another microlanguage or dialect. We list the languages that we considered in this manner in Table 13.

We also made some decisions regarding `wiki_ct` assignment based on the script of a language. We recorded zero Wikipedia pages for `kas_Deva` and 13,210 for `kas_Arab` (all of the Kashmiri pages) because a majority of Kashmiri pages seem to be in Perso-Arabic script. (There may be a few in Devanagari, but we simplify by assuming none are.) We also recorded zero pages for `mni_Beng` because, although Wikipedia has pages in Meitei, they appear to be in the Meitei `Mtei` script, not Bengali `Beng`. Lastly, we assigned Wikipedia's count for 'Classical Chinese' (`zh-classical`) to `zho_Hant` and its count for 'Chinese' to `zho_Hans` (though it is possible that some of the 'Chinese' pages may be in the Traditional Chinese (`Hant`) script).

In all other cases, if a language did not have a listed number of Wikipedia pages, we took this to mean it had zero.

Table 11: BLEU and chrF results on ENG→X directions. “0-shot" and “5-shot" are ChatGPT with zero- and five-shot settings, respectively. “NLLB" is the NLLB-MOE model, and “Google" is Google Translate. We used five-shot settings only for GPT-4. Models are listed in order of their effectiveness in MT (with zero-shot ChatGPT performing the worst and Google Translate performing the best).

Language	**spBLEU200**					**chrF2++**				
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google
ace_Arab	1.3	2.3	–	5.5	–	8.4	9.8	–	17.4	–
ace_Latn	5.0	4.3	–	11.6	–	19.8	18.0	–	37.1	–
acm_Arab	28.2	29.6	29.5	11.8	–	44.7	46.1	46.5	31.9	–
acq_Arab	30.9	31.9	–	26.9	–	47.5	48.1	–	42.2	–
aeb_Arab	24.2	24.7	–	19.9	–	41.0	41.3	–	38.2	–
afr_Latn	47.2	46.7	–	44.4	48.7	67.0	66.7	–	64.3	67.8
ajp_Arab	31.5	32.2	32.2	36.3	–	47.1	47.5	48.4	51.3	–
aka_Latn	3.2	3.1	–	11.7	–	13.3	13.8	–	34.5	–
als_Latn	33.6	34.2	–	39.4	–	56.0	56.3	–	58.3	–
amh_Ethi	3.5	3.7	–	31.6	34.1	10.0	10.6	–	39.4	42.0
apc_Arab	30.4	30.9	–	36.7	–	45.5	45.8	–	50.6	–
arb_Arab	37.6	37.9	–	43.0	48.6	53.8	53.8	–	57.1	62.6
arb_Latn	5.3	8.4	–	–	7.9	22.8	26.9	–	–	35.4
ars_Arab	35.9	37.2	–	36.7	–	52.4	53.1	–	50.5	–
ary_Arab	19.3	19.6	–	23.3	–	36.3	36.7	–	38.9	–
arz_Arab	26.2	26.6	–	32.1	–	42.3	42.7	–	46.8	–
asm_Beng	8.2	10.6	–	22.5	23.2	23.2	26.1	–	35.9	37.4
ast_Latn	31.3	32.3	–	34.5	–	53.8	54.5	–	56.8	–
awa_Deva	15.6	16.6	–	27.6	–	35.4	36.3	–	47.1	–
ayr_Latn	0.2	0.1	–	7.6	7.2	4.7	3.8	–	29.7	31.5
azb_Arab	3.5	3.6	–	5.4	–	17.9	18.5	–	23.5	–
azj_Latn	16.6	17.7	–	24.6	–	38.4	40.3	–	42.9	–
bak_Cyrl	5.5	5.7	–	30.3	–	20.1	20.7	–	47.3	–
bam_Latn	0.5	0.7	–	9.3	9.5	6.1	6.9	–	30.5	32.6
ban_Latn	10.9	9.0	–	19.4	–	30.7	27.4	–	44.6	–
bel_Cyrl	19.5	20.5	–	27.3	30.1	38.3	39.1	–	42.0	44.4
bem_Latn	1.6	1.1	–	13.6	–	10.3	9.1	–	37.9	–
ben_Beng	21.8	22.1	–	36.0	37.6	38.5	39.0	–	50.0	51.4
bho_Deva	11.9	12.5	–	23.6	21.0	29.7	30.7	–	42.8	40.0
bjn_Arab	2.0	3.0	–	5.8	–	10.4	13.2	–	17.1	–
bjn_Latn	11.0	12.3	–	21.9	–	35.9	38.0	–	48.2	–
bod_Tibt	0.2	0.4	–	8.5	–	12.7	14.7	–	29.7	–
bos_Latn	40.0	40.6	–	40.7	44.0	59.9	60.1	–	58.8	61.8
bug_Latn	5.2	2.7	–	9.1	–	23.3	16.4	–	33.7	–
bul_Cyrl	44.1	44.4	–	50.0	53.1	61.6	61.9	–	64.8	67.9
cat_Latn	47.8	47.9	–	48.9	51.1	65.4	65.3	–	65.0	67.2
ceb_Latn	28.0	29.1	–	34.5	40.2	51.0	52.9	–	57.3	62.2
ces_Latn	40.8	40.8	–	42.4	46.0	57.6	57.4	–	57.4	60.3
cjk_Latn	0.2	0.1	–	4.0	–	4.4	4.5	–	24.3	–
ckb_Arab	4.7	6.5	11.2	26.8	25.8	19.7	24.8	33.1	47.2	47.7
crh_Latn	6.0	6.8	–	27.4	–	27.8	29.0	–	47.0	–
cym_Latn	48.0	48.5	–	58.4	63.6	64.7	64.9	–	70.8	74.5
dan_Latn	52.3	52.5	–	50.0	55.3	69.7	69.7	–	66.4	70.3
deu_Latn	47.7	47.9	–	46.6	51.2	65.4	65.4	–	62.8	66.5
dik_Latn	0.2	0.1	–	6.1	–	4.6	4.4	–	24.2	–
dyu_Latn	0.5	0.1	–	2.7	–	4.5	4.3	–	17.7	–
dzo_Tibt	0.1	0.7	–	13.3	–	7.7	15.9	–	34.7	–
ell_Grek	35.8	35.8	–	38.7	40.1	51.6	51.6	–	52.0	53.6
epo_Latn	37.9	38.5	–	42.8	40.4	58.5	58.8	–	61.4	60.1
est_Latn	35.3	35.8	–	36.5	41.4	56.8	56.9	–	56.1	59.9
eus_Latn	19.1	19.5	–	29.0	33.9	44.2	43.9	–	50.0	54.5
ewe_Latn	0.6	0.7	–	17.2	17.0	6.0	6.1	–	39.0	39.9
fao_Latn	18.1	19.2	–	31.6	–	40.5	41.5	–	49.8	–
fij_Latn	5.5	4.8	–	23.6	–	22.9	21.3	–	46.7	–
fin_Latn	35.8	36.1	–	36.6	39.2	56.2	56.4	–	55.3	58.0
fon_Latn	0.2	0.2	–	6.4	–	3.9	4.1	–	21.5	–
fra_Latn	56.4	56.6	57.3	56.2	59.7	71.1	71.3	71.7	69.7	72.7
fur_Latn	18.5	19.8	–	39.6	–	40.6	42.5	–	56.8	–
fuv_Latn	1.2	0.4	–	6.0	–	8.5	5.8	–	23.9	–
gaz_Latn	0.6	0.4	–	12.6	14.6	8.0	7.3	–	37.5	40.3
gla_Latn	15.5	16.3	–	28.7	32.2	38.9	39.0	–	50.2	52.7

Continued on next page

Table 11 – continued from previous page

Language	spBLEU200					chrF2++				
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google
gle_Latn	25.8	26.3	32.8	41.4	44.1	47.1	47.5	53.0	58.0	60.1
glg_Latn	39.4	40.0	–	40.1	41.9	61.3	61.5	–	59.8	61.5
grn_Latn	0.7	0.6	–	16.4	15.3	6.3	5.7	–	36.6	36.4
guj_Gujr	18.9	19.4	–	37.2	39.2	37.4	37.1	–	53.3	55.2
hat_Latn	24.5	24.8	–	30.5	31.8	47.0	47.2	–	51.9	53.4
hau_Latn	6.2	6.3	15.7	31.4	30.6	22.2	22.4	40.3	53.5	53.2
heb_Hebr	35.3	35.4	–	46.8	48.8	51.2	50.7	–	59.8	61.2
hin_Deva	29.2	29.4	–	40.6	43.0	48.7	48.6	–	57.3	59.3
hne_Deva	14.1	15.5	–	33.7	–	34.0	36.1	–	54.3	–
hrv_Latn	37.8	38.2	–	38.9	42.5	57.0	57.2	–	57.2	60.2
hun_Latn	34.8	34.9	–	38.1	40.9	54.6	54.5	–	55.5	58.1
hye_Armn	14.3	14.8	–	40.2	42.7	33.2	33.5	–	53.2	56.3
ibo_Latn	3.2	4.0	9.8	20.6	22.2	14.7	16.3	27.7	41.4	43.5
ilo_Latn	11.4	12.6	–	29.0	31.0	33.6	35.6	–	53.3	56.0
ind_Latn	48.8	48.7	–	49.2	55.0	68.5	68.5	–	68.7	72.6
isl_Latn	26.0	26.0	–	33.9	40.8	44.8	45.0	–	50.0	55.8
ita_Latn	37.6	37.7	–	38.3	40.0	59.4	59.5	–	57.3	59.1
jav_Latn	16.9	18.9	–	30.3	30.3	41.2	42.7	–	54.8	55.1
jpn_Jpan	30.5	31.3	–	20.1	35.3	33.1	33.7	–	27.9	37.1
kab_Latn	1.3	1.5	–	16.9	–	11.9	12.9	–	35.6	–
kac_Latn	0.0	0.1	–	14.3	–	2.9	4.8	–	37.5	–
kam_Latn	1.3	1.1	–	6.1	–	8.9	9.0	–	25.9	–
kan_Knda	18.6	19.4	–	39.6	41.9	37.9	38.2	–	53.4	55.7
kas_Arab	4.0	3.3	–	18.2	–	15.5	14.3	–	34.2	–
kas_Deva	2.3	2.7	–	4.7	–	12.9	13.9	–	17.1	–
kat_Geor	15.2	15.7	23.2	34.6	37.5	32.5	33.5	41.4	48.1	51.4
kaz_Cyrl	12.9	13.4	–	34.0	38.7	33.9	33.4	–	51.8	56.0
kbp_Latn	0.4	1.4	–	11.3	–	4.0	9.4	–	28.3	–
kea_Latn	13.0	18.7	–	22.5	–	37.6	43.0	–	42.8	–
khk_Cyrl	8.0	8.5	–	27.1	33.1	26.1	26.6	–	43.9	49.8
khm_Khmr	5.7	6.0	–	23.0	27.4	21.5	21.1	–	36.4	40.3
kik_Latn	0.8	2.0	–	15.4	–	8.8	11.6	–	37.1	–
kin_Latn	3.4	3.1	–	27.2	34.3	18.7	18.0	–	49.7	56.1
kir_Cyrl	8.4	8.9	–	27.4	30.5	25.8	26.6	–	44.5	48.2
kmb_Latn	0.4	0.4	–	4.5	–	4.9	6.1	–	24.9	–
kmr_Latn	8.3	9.4	14.3	19.6	20.0	25.3	27.4	34.3	39.3	40.0
knc_Arab	0.5	1.1	–	6.5	–	5.3	4.7	–	9.8	–
knc_Latn	2.6	0.9	–	8.2	–	13.4	8.1	–	27.4	–
kon_Latn	0.9	1.3	–	18.9	–	8.5	10.5	–	45.3	–
kor_Hang	25.6	25.9	–	26.7	30.0	34.4	34.9	–	36.0	38.6
lao_Laoo	2.9	4.0	–	29.6	29.6	18.5	21.5	–	46.2	44.0
lij_Latn	7.6	10.3	–	37.2	–	32.8	35.2	–	53.8	–
lim_Latn	15.1	19.8	21.0	25.8	–	40.2	42.7	45.1	47.9	–
lin_Latn	2.6	2.5	–	21.9	21.4	14.8	14.7	–	48.0	48.4
lit_Latn	30.0	30.6	–	35.4	41.7	51.5	51.8	–	54.7	59.4
lmo_Latn	6.7	8.3	–	10.5	–	29.9	30.6	–	34.9	–
ltg_Latn	5.3	5.4	–	36.4	–	29.2	29.1	–	53.6	–
ltz_Latn	25.4	27.5	–	36.7	35.3	48.7	48.9	–	56.0	55.6
lua_Latn	1.0	1.1	–	9.8	–	8.1	9.3	–	35.2	–
lug_Latn	1.6	1.3	–	14.0	14.4	11.6	10.6	–	39.8	41.3
luo_Latn	0.8	0.1	–	15.2	–	7.0	5.0	–	38.5	–
lus_Latn	4.6	4.7	–	15.1	–	17.6	17.8	–	38.0	–
lvs_Latn	33.0	33.5	36.7	35.4	–	55.1	55.2	57.3	54.8	–
mag_Deva	18.6	19.4	24.8	39.4	–	39.1	39.9	44.6	58.5	–
mai_Deva	10.2	12.1	–	27.1	19.6	28.9	31.2	–	46.7	40.6
mal_Mlym	14.6	14.9	–	38.3	43.2	32.3	32.0	–	51.6	56.2
mar_Deva	14.5	14.7	–	30.3	33.4	34.3	34.6	–	48.0	51.0
min_Arab	1.6	3.5	–	–	–	10.1	14.9	–	–	–
min_Latn	11.5	13.1	–	28.7	–	37.0	38.4	–	52.4	–
mkd_Cyrl	36.0	36.5	–	42.6	46.5	57.0	57.3	–	60.6	63.7
mlt_Latn	29.9	30.3	–	50.3	59.7	49.4	49.8	–	66.0	71.6
mni_Beng	1.8	2.0	–	27.5	0.1	11.4	10.5	–	38.7	0.6
mos_Latn	0.2	0.2	–	6.8	–	3.9	4.3	–	24.3	–
mri_Latn	15.1	14.5	–	20.7	18.3	34.8	34.0	–	44.2	42.4
mya_Mymr	2.1	2.8	–	17.7	24.5	19.8	20.6	–	32.0	40.4
nld_Latn	36.3	36.5	–	35.6	38.0	56.5	56.7	–	54.9	57.3

Continued on next page

Table 11 – continued from previous page

Language	spBLEU200					chrF2++				
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google
nno_Latn	37.1	38.3	–	33.4	25.6	58.7	59.4	–	53.6	50.7
nob_Latn	40.2	39.8	–	38.4	–	60.5	60.2	–	58.6	–
npi_Deva	19.0	19.6	–	28.7	–	39.5	39.3	–	45.5	–
nso_Latn	3.7	4.6	–	26.5	29.8	16.7	19.0	–	50.8	54.0
nus_Latn	0.1	0.5	–	14.4	–	3.0	5.5	–	29.0	–
nya_Latn	4.9	5.5	–	17.7	21.1	20.6	22.6	–	44.0	48.0
oci_Latn	30.4	33.3	–	41.0	–	55.1	57.0	–	58.8	–
ory_Orya	11.6	12.6	–	30.2	38.9	27.5	29.8	–	45.7	53.4
pag_Latn	5.7	8.3	–	20.2	–	22.6	26.7	–	46.3	–
pan_Guru	21.0	21.5	–	36.4	39.7	37.4	37.6	–	49.0	51.9
pap_Latn	25.4	33.2	–	42.2	–	51.6	56.5	–	60.2	–
pbt_Arab	5.1	5.8	9.2	22.9	–	19.7	21.1	26.7	39.4	–
pes_Arab	29.4	30.4	–	36.1	39.8	48.6	48.8	–	51.3	54.3
plt_Latn	8.2	8.3	–	25.3	25.9	31.4	30.9	–	50.0	51.2
pol_Latn	32.1	32.6	–	32.5	36.3	49.7	50.0	–	48.9	52.1
por_Latn	56.4	56.9	–	52.9	58.6	71.4	71.7	–	67.9	72.3
prs_Arab	25.7	27.5	–	33.8	–	44.8	47.4	–	53.6	–
quy_Latn	0.7	0.6	–	5.8	8.2	9.3	9.5	–	26.9	34.0
ron_Latn	46.2	46.9	49.0	44.7	50.0	64.0	64.2	65.3	61.3	65.0
run_Latn	3.1	2.3	–	19.6	–	16.6	14.7	–	42.5	–
rus_Cyrl	38.9	38.9	–	41.0	43.9	56.6	56.5	–	56.3	58.7
sag_Latn	0.1	0.1	–	10.5	–	4.6	5.1	–	35.7	–
san_Deva	4.7	5.4	–	8.0	10.0	21.8	22.6	–	26.1	30.3
sat_Olck	0.0	1.9	–	18.5	–	0.2	14.4	–	26.3	–
scn_Latn	11.2	13.0	–	24.4	–	35.9	37.2	–	46.8	–
shn_Mymr	0.5	1.3	–	15.1	–	7.6	16.6	–	34.4	–
sin_Sinh	6.1	6.9	–	36.0	40.4	19.5	20.1	–	43.8	51.2
slk_Latn	38.6	38.4	–	42.9	48.4	56.8	57.0	–	59.0	63.1
slv_Latn	35.7	36.0	–	38.1	42.4	55.5	55.7	–	56.2	59.6
smo_Latn	6.3	8.0	–	26.9	–	22.8	26.3	–	50.0	–
sna_Latn	3.2	3.4	8.4	19.7	20.8	15.3	16.3	29.2	43.4	44.4
snd_Arab	9.1	10.5	–	31.9	32.6	22.5	24.9	–	48.1	48.7
som_Latn	8.1	8.1	–	18.4	18.9	29.4	29.7	–	43.0	43.7
sot_Latn	5.7	5.4	–	20.7	22.5	20.7	20.9	–	46.1	47.8
spa_Latn	33.8	33.9	–	33.1	35.0	56.5	56.7	–	53.8	55.5
srd_Latn	16.3	18.5	–	35.8	–	42.1	43.8	–	55.6	–
srp_Cyrl	37.5	37.9	–	43.4	48.1	56.5	57.2	–	59.7	63.4
ssw_Latn	1.9	0.5	5.8	19.9	–	10.6	6.7	24.1	43.3	–
sun_Latn	13.9	14.5	–	21.6	24.4	39.0	38.6	–	44.7	48.7
swe_Latn	52.5	52.2	–	50.1	54.2	68.5	68.4	–	65.9	69.4
swh_Latn	38.0	38.6	–	36.8	44.6	60.1	60.3	–	58.6	64.4
szl_Latn	12.8	15.1	–	38.4	–	35.5	36.7	–	53.7	–
tam_Taml	13.6	13.4	20.9	36.6	38.7	33.8	34.5	42.7	53.7	55.8
taq_Latn	0.8	0.3	–	4.9	–	8.2	6.2	–	23.1	–
taq_Tfng	0.6	1.4	–	5.6	–	5.2	8.3	–	16.7	–
tat_Cyrl	6.7	7.3	–	30.4	30.4	21.5	23.6	–	46.8	48.2
tel_Telu	17.4	18.0	–	41.6	44.7	34.4	35.6	–	55.9	58.2
tgk_Cyrl	10.8	11.7	–	35.3	35.6	29.3	30.4	–	51.2	51.8
tgl_Latn	35.0	35.0	–	38.3	39.8	60.8	60.6	–	60.5	61.8
tha_Thai	33.5	33.6	–	35.1	45.2	43.1	43.2	–	42.7	49.7
tir_Ethi	1.6	1.9	–	17.8	17.6	5.8	6.7	–	25.8	26.3
tpi_Latn	14.0	15.8	22.7	17.8	–	37.1	39.2	49.5	41.6	–
tsn_Latn	3.8	4.2	–	25.6	–	17.0	18.6	–	48.5	–
tso_Latn	2.8	3.0	–	26.7	26.1	15.0	16.0	–	50.0	50.9
tuk_Latn	6.2	7.7	–	22.6	35.8	25.2	25.9	–	42.1	52.7
tum_Latn	3.6	2.9	–	13.3	–	16.5	14.8	–	35.2	–
tur_Latn	38.5	38.5	–	41.5	46.4	57.9	57.8	–	58.3	62.4
twi_Latn	3.0	3.0	–	15.2	17.4	13.4	14.2	–	37.9	40.9
tzm_Tfng	1.1	2.2	–	21.0	–	8.3	11.7	–	32.3	–
uig_Arab	6.5	8.5	–	30.5	40.2	20.5	24.7	–	45.3	54.3
ukr_Cyrl	37.4	37.4	39.2	40.1	42.8	55.0	55.4	56.3	56.3	58.6
umb_Latn	0.4	0.1	–	4.1	–	5.3	4.9	–	26.6	–
urd_Arab	21.9	22.2	–	30.5	32.7	41.7	41.8	–	48.9	50.0
uzn_Latn	17.4	18.8	–	30.0	37.8	39.9	40.9	–	50.6	56.4
vec_Latn	15.7	17.5	–	28.2	–	41.0	42.8	–	51.6	–
vie_Latn	40.7	40.7	–	43.3	–	58.5	57.9	–	59.5	–

Continued on next page

Table 11 – continued from previous page

Language	spBLEU200					chrF2++				
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google
war_Latn	24.3	25.0	28.4	35.0	–	49.3	49.5	54.0	57.4	–
wol_Latn	2.1	1.2	–	9.6	–	10.6	8.3	–	29.7	–
xho_Latn	5.3	6.0	–	25.4	29.5	21.9	23.3	–	48.6	52.2
ydd_Hebr	10.6	18.7	–	18.4	16.8	31.0	38.1	–	38.6	37.7
yor_Latn	2.5	3.3	–	10.5	4.9	11.4	13.7	–	25.5	20.0
yue_Hant	26.4	33.8	–	16.6	–	22.3	27.2	–	17.9	–
zho_Hans	36.3	36.5	–	26.6	43.6	31.0	31.9	–	22.8	37.8
zho_Hant	29.3	30.4	–	12.4	–	24.8	26.0	–	14.0	–
zsm_Latn	41.4	41.3	–	45.5	47.5	64.5	64.3	–	66.5	68.0
zul_Latn	6.7	7.3	–	31.4	32.0	25.2	26.3	–	53.3	53.9

Figure 4: chrF scores across all MT systems and languages

Figure 5: BLEU scores across all MT systems and languages

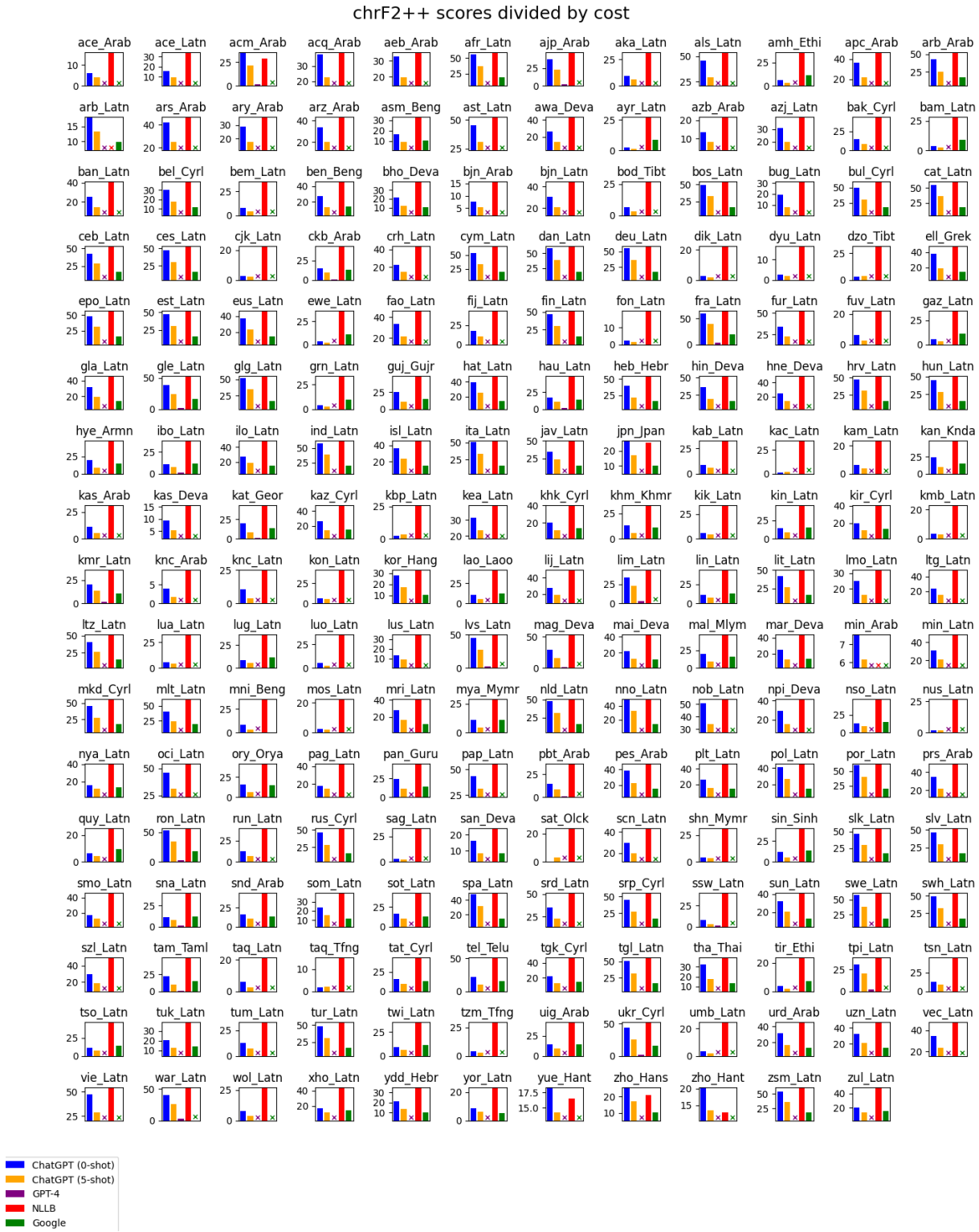

Figure 6: chrF scores divided by the estimated cost of each MT system, across all MT systems and languages

Figure 7: BLEU scores divided by the estimated cost of each MT system, across all MT systems and languages

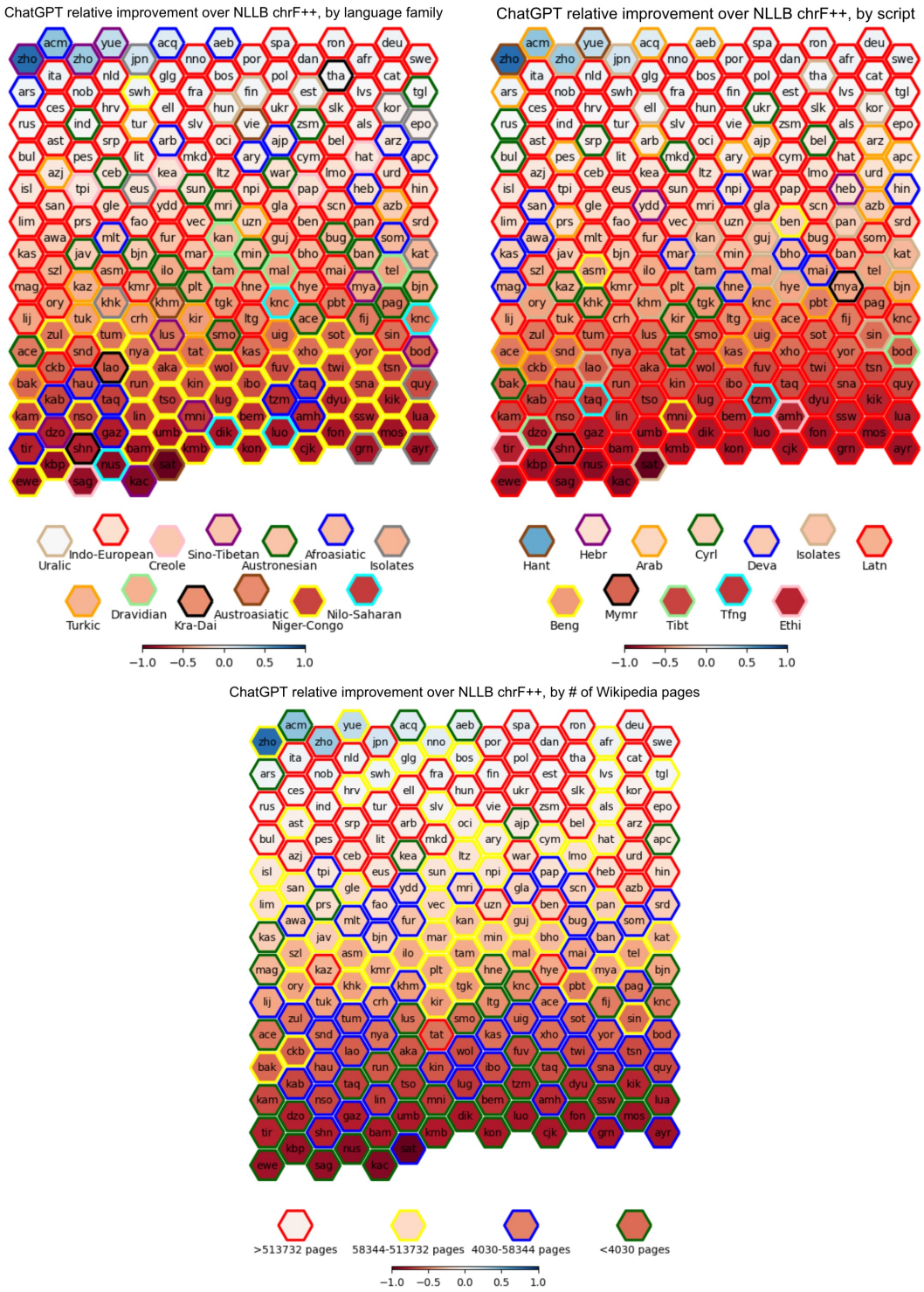

Figure 8: ChatGPT *relative improvement* over NLLB chrF (color scale), with languages organized by family, script, and number of Wikipedia pages (divided in quartiles). Hexagons (one per language) are displayed in descending order across rows, with the highest ChatGPT relative improvement over NLLB chrF2++ at the top left, and the lowest at the bottom right. Group hexagons at the bottom of each plot display the average color for each group and are organized in like manner.

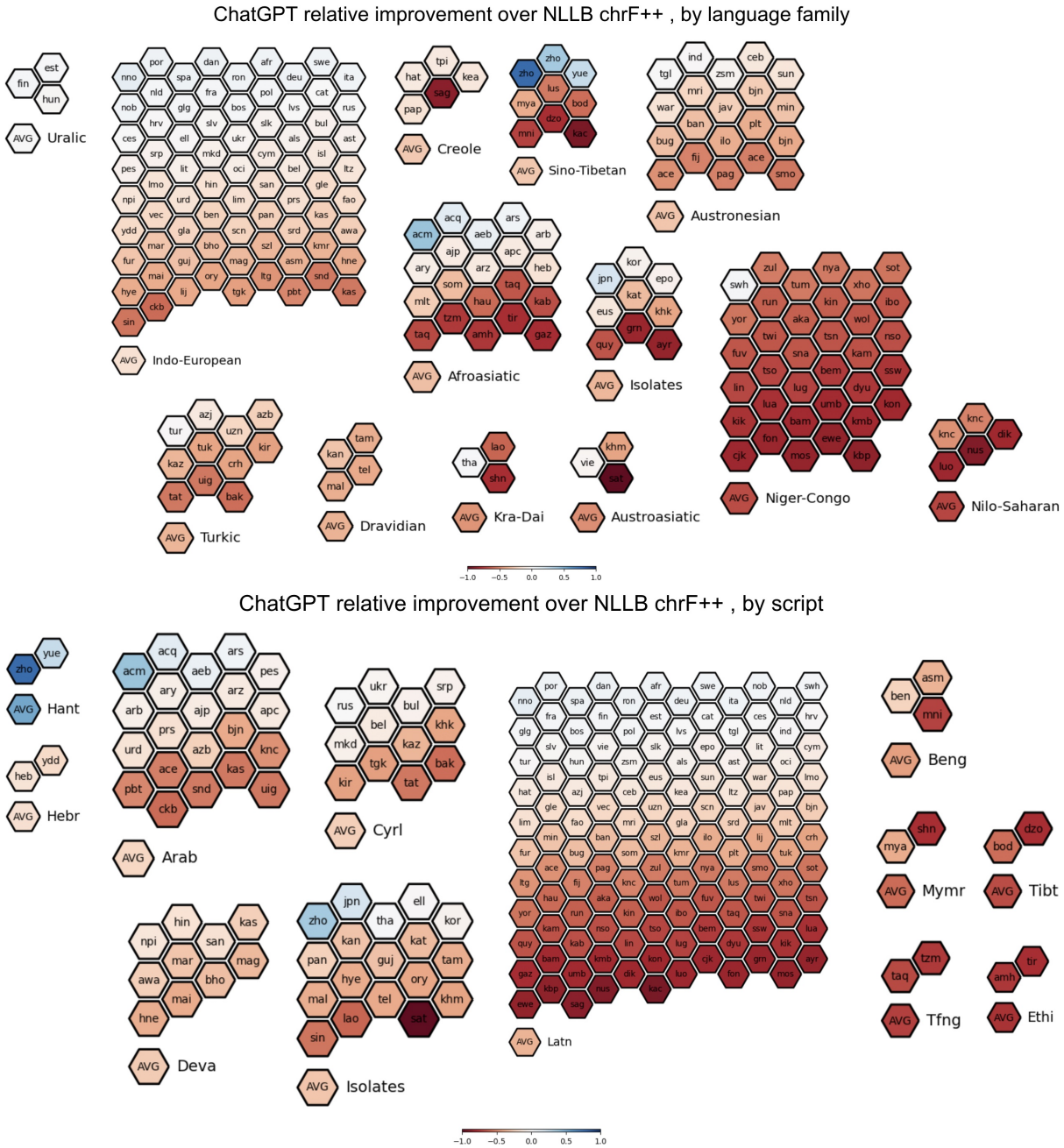

Figure 9: Alternative visualizations to those in Figure 8. Groups and languages are organized the same here: from top left to bottom right in descending order of the ChatGPT *relative improvement* over NLLB (using averages for the groups).

FLORES lang.	substitution for wiki_ct
arb	Used macrolanguage ‘Arabic’ (ara) because ‘Standard Arabic’ (arb) not present
bho	Used macrolanguage ‘Bihari’ (bih) because ‘Bhojpuri’ (bho) not present
dik	Used macrolanguage ‘Dinka’ (din) because ‘Southwestern Dinka’ (dik) not present
fuv	Used macrolanguage ‘Fula’ (ful) because "Nigerian Fulfulde" (fuv) not present
knc	Used macrolanguage ‘Kanuri’ (kau) because ‘Central Kanuri’ (knc) not present
lvs	Used macrolanguage ‘Latvian’ (lav) because ‘Standard Latvian’ (lvs) not present
plt	Used macrolanguage ‘Malagasy’ (mlg) because ‘Plateau Malagasy’ (plt) not present
khk	Used macrolanguage ‘Mongolian’ (mon) because ‘Halh Mongolian’ (khk) not present
gaz	Used macrolanguage ‘Oromo’ (orm) because ‘West Central Oromo’ (gaz) not present
pes	Used macrolanguage ‘Persian’ (fas) because ‘Western Persian’ (pes) not present
pbt	Used macrolanguage ‘Pashto’ (pus) because ‘Southern Pashto’ (pbt) not present
quy	Used macrolanguage ‘Quechua’ (que) because ‘Ayuacucho Quechua’ (quy) not present
als	Used macrolanguage ‘Albanian’ (sqi) because ‘Tosk Albanian’ (als) not present
uzn	Used macrolanguage ‘Uzbek’ (uzb) because ‘Northern Uzbek’ (uzn) not present
ydd	Used macrolanguage ‘Yiddish’ (yid) because ‘Eastern Yiddish’ (ydd) not present
zsm	Used macrolangauge ‘Malay’ (msa) because ‘Standard Malay’ (zsm) not present

Table 12: FLORES-200 languages for which we used the Wikipedia page count associated with a macrolanguage of another ISO 639-3 code

FLORES lang.	reason for assigning wiki_ct = 0
acm	Macrolanguage ‘Arabic’ (ara) appears to be in ‘Standard Arabic’ (arb), not ‘Mesopotamian Arabic’ (acm)
acq	Macrolanguage ‘Arabic’ (ara) appears to be in ‘Standard Arabic’ (arb), not ‘Tai’izzi Arabic’ (acq)
aeb	Macrolanguage ‘Arabic’ (ara) appears to be in ‘Standard Arabic’ (arb), not ‘Tunisian Arabic’ (aeb)
ajp	Macrolanguage ‘Arabic’ (ara) appears to be in ‘Standard Arabic’ (arb), not ‘South Levantine Arabic’ (ajp)
apc	Macrolanguage ‘Arabic’ (ara) appears to be in ‘Standard Arabic’ (arb), not ‘North Levantine Arabic’ (apc)
ars	Macrolanguage ‘Arabic’ (ara) appears to be in ‘Standard Arabic’ (arb), not ‘Najdi Arabic’ (ars)
mag	Macrolanguage ‘Bihari’ (bih) appears to be in ‘Bhojpuri’ (bho), not ‘Magahi’ (mag)
prs	Macrolanguage ‘Persian’ (fas) appears to be in ‘Western Persian’ (pes), not ‘Dari’ (prs)

Table 13: FLORES-200 languages for which we used assigned wiki_ct to be zero, despite the existence of Wikipedia pages in a corresponding macrolanguage

Table 14: Estimated costs in USD to translate the FLORES-200 *devtest* set ENG→X for each targe language and MT system, along with BLEU and chrF scores divided by the cost estimates, where applicable. The cost is roughly $0.09 for NLLB and $2.66 for Google Translate for all target languages.

Lang.	spBLEU200/cost					chrF2++/cost					cost estimate (USD$)		
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4
ace_Arab	0.9	0.9	–	5.1	–	6.3	4.0	–	16.0	–	0.3	1.5	29.0
ace_Latn	4.0	2.2	–	10.7	–	15.8	9.1	–	34.2	–	0.3	1.0	18.9
acm_Arab	22.4	13.9	1.2	10.8	–	35.6	21.6	1.8	29.3	–	0.3	1.1	24.3
acq_Arab	24.6	14.9	–	24.8	–	37.7	22.4	–	38.9	–	0.3	1.1	24.6
aeb_Arab	19.3	11.6	–	18.3	–	32.7	19.4	–	35.2	–	0.3	1.1	24.1
afr_Latn	39.5	25.8	–	41.0	13.3	56.2	36.8	–	59.4	18.5	0.2	0.8	17.1
ajp_Arab	25.0	15.3	1.3	33.4	–	37.4	22.5	2.0	47.3	–	0.3	1.1	23.7
aka_Latn	2.3	1.4	–	10.8	–	9.5	6.1	–	31.8	–	0.4	1.2	22.9
als_Latn	27.6	17.5	–	36.3	–	46.0	28.9	–	53.8	–	0.2	0.9	20.3
amh_Ethi	2.3	1.1	–	28.8	9.3	6.4	3.2	–	35.9	11.5	0.6	2.4	50.8
apc_Arab	24.2	14.7	–	33.8	–	36.2	21.8	–	46.6	–	0.3	1.1	23.6
arb_Arab	29.9	17.6	–	39.6	13.3	42.7	25.0	–	52.6	17.1	0.3	1.1	24.8
arb_Latn	4.2	4.2	–	–	2.1	18.1	13.4	–	–	9.7	0.3	1.0	21.5
ars_Arab	28.5	17.3	–	33.8	–	41.7	24.7	–	46.5	–	0.3	1.2	24.8
ary_Arab	15.3	9.2	–	21.5	–	28.8	17.2	–	35.8	–	0.3	1.1	24.4
arz_Arab	20.9	12.5	–	29.6	–	33.7	20.0	–	43.1	–	0.3	1.1	24.3
asm_Beng	5.7	3.6	–	20.6	6.4	16.1	8.8	–	32.9	10.2	0.4	2.0	42.6
ast_Latn	26.4	18.1	–	31.9	–	45.3	30.6	–	52.5	–	0.2	0.8	16.5
awa_Deva	11.5	6.4	–	25.3	–	26.1	14.0	–	43.2	–	0.4	1.6	34.6
ayr_Latn	0.1	0.0	–	7.0	2.0	2.8	1.5	–	27.4	8.6	0.7	1.5	19.9
azb_Arab	2.7	1.6	–	5.0	–	13.6	8.1	–	21.6	–	0.3	1.3	26.4
azj_Latn	13.4	8.7	–	22.7	–	31.1	19.7	–	39.6	–	0.2	1.0	22.5
bak_Cyrl	4.0	2.2	–	27.9	–	14.8	8.2	–	43.5	–	0.4	1.5	31.8
bam_Latn	0.3	0.3	–	8.6	2.6	3.7	2.8	–	28.1	8.9	0.6	1.4	22.1
ban_Latn	9.0	4.8	–	17.9	–	25.4	14.6	–	41.2	–	0.2	0.9	17.7
bel_Cyrl	15.3	9.1	–	25.1	8.2	30.0	17.3	–	38.7	12.1	0.3	1.3	27.4
bem_Latn	1.1	0.5	–	12.5	–	7.3	4.1	–	35.0	–	0.4	1.2	20.1
ben_Beng	15.4	7.7	–	33.0	10.3	27.2	13.6	–	45.8	14.0	0.4	1.9	40.6
bho_Deva	8.8	4.8	–	21.7	5.7	21.9	11.9	–	39.3	10.9	0.4	1.6	34.2
bjn_Arab	1.5	1.3	–	5.3	–	7.7	5.5	–	15.7	–	0.4	1.4	29.0
bjn_Latn	9.2	6.8	–	20.2	–	30.2	21.0	–	44.5	–	0.2	0.8	17.2
bod_Tibt	0.1	0.1	–	7.7	–	6.4	3.3	–	26.9	–	1.0	3.4	71.3
bos_Latn	33.3	21.9	–	37.6	12.0	49.9	32.4	–	54.3	16.9	0.2	0.9	18.2

Continued on next page

Table 14 – continued from previous page

Lang.	spBLEU200/cost					chrF2++/cost					cost estimate (USD$)		
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4
bug_Latn	4.3	1.4	–	8.4	–	19.2	8.4	–	31.1	–	0.2	1.0	18.7
bul_Cyrl	35.4	21.6	–	46.1	14.5	49.5	30.2	–	59.7	18.5	0.2	1.1	22.5
cat_Latn	40.0	26.4	–	45.2	14.0	54.8	36.0	–	60.0	18.3	0.2	0.8	17.2
ceb_Latn	23.1	15.5	–	31.8	11.0	42.1	28.2	–	52.9	17.0	0.2	0.9	18.5
ces_Latn	33.7	21.3	–	39.1	12.6	47.5	30.0	–	53.0	16.5	0.2	0.9	19.5
cjk_Latn	0.1	0.1	–	3.7	–	2.8	2.0	–	22.4	–	0.6	1.3	18.7
ckb_Arab	3.5	2.5	0.3	24.6	7.0	14.5	9.4	0.9	43.3	13.0	0.4	1.6	35.1
crh_Latn	4.9	3.6	–	25.3	–	23.0	15.2	–	43.4	–	0.2	0.9	19.5
cym_Latn	39.6	25.2	–	53.9	17.4	53.3	33.8	–	65.3	20.4	0.2	0.9	19.7
dan_Latn	44.0	29.3	–	46.2	15.1	58.7	38.9	–	61.3	19.2	0.2	0.8	16.8
deu_Latn	40.2	26.9	–	43.0	14.0	55.1	36.7	–	58.0	18.2	0.2	0.8	16.5
dik_Latn	0.1	0.1	–	5.6	–	2.8	1.8	–	22.3	–	0.6	1.5	20.3
dyu_Latn	0.3	0.1	–	2.5	–	2.7	1.8	–	16.3	–	0.7	1.4	19.2
dzo_Tibt	0.0	0.1	–	12.0	–	2.8	3.5	–	31.4	–	1.8	3.6	76.8
ell_Grek	26.1	13.3	–	35.5	10.9	37.7	19.2	–	47.7	14.6	0.4	1.7	36.8
epo_Latn	31.5	20.8	–	39.5	11.0	48.7	31.7	–	56.7	16.4	0.2	0.9	18.1
est_Latn	29.4	19.3	–	33.7	11.3	47.3	30.7	–	51.8	16.4	0.2	0.9	18.2
eus_Latn	15.9	10.5	–	26.8	9.3	36.8	23.6	–	46.2	14.9	0.2	0.9	18.2
ewe_Latn	0.4	0.3	–	15.8	4.6	3.7	2.4	–	35.9	10.9	0.6	1.6	23.1
fao_Latn	15.0	10.1	–	29.2	–	33.4	21.8	–	45.9	–	0.2	0.9	19.3
fij_Latn	4.3	2.4	–	21.8	–	17.8	10.5	–	43.1	–	0.3	1.0	19.8
fin_Latn	29.6	19.1	–	33.8	10.7	46.6	29.9	–	51.0	15.8	0.2	0.9	18.9
fon_Latn	0.1	0.1	–	5.9	–	2.3	1.4	–	19.8	–	0.7	1.9	28.4
fra_Latn	47.5	31.7	3.3	51.9	16.3	60.0	39.9	4.1	64.4	19.9	0.2	0.8	16.6
fur_Latn	15.4	10.7	–	36.6	–	33.8	23.0	–	52.4	–	0.2	0.9	18.0
fuv_Latn	0.9	0.2	–	5.5	–	6.2	2.6	–	22.0	–	0.4	1.2	18.0
gaz_Latn	0.4	0.2	–	11.6	4.0	5.5	3.3	–	34.6	11.0	0.5	1.2	20.7
gla_Latn	12.5	8.1	–	26.5	8.8	31.5	19.5	–	46.3	14.4	0.2	1.0	21.2
gle_Latn	21.1	13.3	1.5	38.2	12.0	38.5	24.1	2.4	53.5	16.4	0.2	1.0	20.7
glg_Latn	33.3	22.6	–	37.0	11.5	51.8	34.7	–	55.2	16.8	0.2	0.8	16.4
grn_Latn	0.5	0.2	–	15.1	4.2	4.1	2.4	–	33.8	9.9	0.5	1.3	19.8
guj_Gujr	12.5	5.8	–	33.9	10.7	24.7	11.1	–	48.6	15.1	0.5	2.4	51.2
hat_Latn	20.5	13.6	–	28.2	8.7	39.3	25.8	–	47.9	14.6	0.2	0.8	17.4
hau_Latn	4.9	3.2	0.8	29.0	8.4	17.6	11.5	2.0	49.4	14.5	0.3	0.9	18.9
heb_Hebr	27.2	15.3	–	43.0	13.3	39.4	21.9	–	55.0	16.7	0.3	1.3	28.3
hin_Deva	21.5	11.3	–	37.3	11.8	35.8	18.7	–	52.6	16.2	0.4	1.6	34.6
hne_Deva	10.4	6.0	–	30.9	–	25.0	14.0	–	49.8	–	0.4	1.6	34.1
hrv_Latn	31.5	20.7	–	35.9	11.6	47.5	30.9	–	52.8	16.5	0.2	0.8	18.0
hun_Latn	28.6	18.1	–	35.1	11.2	44.9	28.3	–	51.2	15.9	0.2	0.9	19.7
hye_Armn	8.6	3.7	–	36.5	11.7	19.9	8.5	–	48.4	15.4	0.7	2.9	63.9
ibo_Latn	2.4	1.9	0.4	19.0	6.1	11.0	7.8	1.2	38.2	11.9	0.3	1.1	21.9
ilo_Latn	9.3	6.5	–	26.8	8.5	27.3	18.5	–	49.2	15.3	0.2	0.9	19.2
ind_Latn	41.1	27.5	–	45.4	15.0	57.8	38.6	–	63.5	19.8	0.2	0.8	16.3
isl_Latn	21.3	13.5	–	31.3	11.2	36.8	23.4	–	46.1	15.3	0.2	0.9	19.7
ita_Latn	31.7	21.0	–	35.4	10.9	50.1	33.2	–	52.9	16.1	0.2	0.8	16.8
jav_Latn	14.1	10.3	–	28.0	8.3	34.4	23.4	–	50.6	15.0	0.2	0.8	17.3
jpn_Jpan	24.9	16.0	–	18.5	9.7	27.1	17.2	–	25.7	10.1	0.2	1.0	20.4
kab_Latn	1.0	0.7	–	15.6	–	9.1	6.2	–	32.8	–	0.3	1.1	21.5
kac_Latn	0.0	0.0	–	13.2	–	1.6	1.9	–	34.6	–	0.8	1.5	21.1
kam_Latn	0.9	0.5	–	5.6	–	6.4	4.1	–	23.9	–	0.4	1.2	19.9
kan_Knda	11.8	5.3	–	36.0	11.5	24.1	10.5	–	48.6	15.2	0.6	2.6	58.0
kas_Arab	2.9	1.2	–	16.7	–	11.3	5.4	–	31.4	–	0.4	1.7	33.8
kas_Deva	1.7	1.0	–	4.3	–	9.3	5.4	–	15.7	–	0.4	1.6	34.1
kat_Geor	9.3	4.0	0.4	31.5	10.2	19.8	8.6	0.6	43.8	14.1	0.6	2.9	62.7
kaz_Cyrl	9.8	5.7	–	31.3	10.6	25.9	14.1	–	47.6	15.3	0.3	1.4	29.1
kbp_Latn	0.2	0.5	–	10.4	–	2.3	3.3	–	26.0	–	0.7	1.9	34.3
kea_Latn	11.0	10.3	–	20.8	–	31.6	23.6	–	39.5	–	0.2	0.8	17.3
khk_Cyrl	6.1	3.6	–	24.9	9.1	19.9	11.3	–	40.4	13.6	0.3	1.4	28.9
khm_Khmr	3.5	1.7	–	21.0	7.5	13.3	5.8	–	33.2	11.0	0.6	2.6	57.2
kik_Latn	0.5	0.8	–	14.2	–	6.0	4.6	–	34.2	–	0.5	1.5	26.1
kin_Latn	2.6	1.5	–	25.1	9.4	14.6	9.0	–	45.8	15.3	0.3	1.0	19.6
kir_Cyrl	6.4	3.9	–	25.2	8.3	19.6	11.5	–	41.0	13.2	0.3	1.3	27.5
kmb_Latn	0.3	0.2	–	4.2	–	3.2	2.7	–	23.0	–	0.5	1.3	19.7
kmr_Latn	6.6	4.8	0.7	18.1	5.5	20.2	14.0	1.6	36.3	10.9	0.3	1.0	20.1
knc_Arab	0.4	0.4	–	6.0	–	3.9	1.8	–	9.0	–	0.3	1.6	28.2
knc_Latn	2.1	0.4	–	7.6	–	10.6	3.6	–	25.3	–	0.3	1.2	21.1
kon_Latn	0.7	0.6	–	17.4	–	5.9	5.0	–	41.8	–	0.4	1.1	18.8
kor_Hang	20.9	13.1	–	24.6	8.2	28.1	17.6	–	33.2	10.5	0.2	1.0	21.0
lao_Laoo	1.6	1.0	–	26.9	8.1	10.2	5.5	–	42.0	12.0	0.8	2.9	62.0
lij_Latn	6.3	5.5	–	34.3	–	27.4	18.8	–	49.6	–	0.2	0.9	18.8
lim_Latn	12.6	10.8	1.1	23.8	–	33.6	23.2	2.4	44.2	–	0.2	0.8	17.8
lin_Latn	2.0	1.3	–	20.2	5.9	11.2	7.4	–	44.3	13.2	0.3	1.0	18.2
lit_Latn	24.6	15.8	–	32.7	11.4	42.2	26.7	–	50.5	16.2	0.2	0.9	20.0
lmo_Latn	5.6	4.4	–	9.7	–	25.0	16.1	–	32.2	–	0.2	0.9	19.1
ltg_Latn	4.4	2.8	–	33.6	–	23.9	15.0	–	49.4	–	0.2	0.9	20.0
ltz_Latn	21.1	14.6	–	33.9	9.7	40.4	25.9	–	51.7	15.2	0.2	0.9	18.8
lua_Latn	0.7	0.5	–	9.0	–	5.7	4.4	–	32.5	–	0.4	1.1	18.6
lug_Latn	1.2	0.6	–	12.9	3.9	8.6	5.1	–	36.7	11.3	0.4	1.1	18.7
luo_Latn	0.5	0.1	–	14.0	–	5.0	2.2	–	35.5	–	0.4	1.2	17.9
lus_Latn	3.5	2.4	–	13.9	–	13.5	9.0	–	35.1	–	0.3	1.0	18.8
lvs_Latn	27.0	16.9	1.7	32.7	–	45.1	27.9	2.6	50.5	–	0.2	1.0	20.7
mag_Deva	13.7	7.5	0.7	36.2	–	28.8	15.4	1.3	53.7	–	0.4	1.6	34.3
mai_Deva	7.5	4.6	–	24.9	5.4	21.3	11.9	–	42.9	11.1	0.4	1.6	35.3

Continued on next page

Table 14 – continued from previous page

Lang.	spBLEU200/cost					chrF2++/cost					cost estimate (USD$)		
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4
mal_Mlym	9.1	4.0	–	34.9	11.8	20.1	8.7	–	47.0	15.4	0.6	2.7	58.5
mar_Deva	10.5	5.5	–	27.8	9.1	24.8	12.9	–	44.0	13.9	0.4	1.7	36.3
min_Arab	1.2	1.4	–	–	–	7.6	6.2	–	–	–	0.3	1.4	30.1
min_Latn	9.7	7.1	–	26.5	–	31.1	21.0	–	48.4	–	0.2	0.8	17.6
mkd_Cyrl	28.9	17.5	–	39.3	12.7	45.6	27.5	–	55.8	17.4	0.2	1.1	23.2
mlt_Latn	24.2	15.1	–	46.4	16.3	40.0	24.9	–	60.8	19.6	0.2	1.0	21.3
mni_Beng	1.3	0.6	–	25.1	0.0	8.0	3.3	–	35.4	0.2	0.4	2.2	45.8
mos_Latn	0.1	0.1	–	6.3	–	2.3	1.8	–	22.4	–	0.7	1.4	20.5
mri_Latn	12.0	7.2	–	19.1	5.0	27.8	16.9	–	40.8	11.6	0.3	1.0	21.0
mya_Mymr	1.2	0.6	–	16.1	6.7	10.9	4.7	–	29.1	11.0	0.8	3.4	73.8
nld_Latn	30.5	20.4	–	32.9	10.4	47.6	31.7	–	50.7	15.7	0.2	0.8	16.6
nno_Latn	31.3	21.4	–	30.8	7.0	49.4	33.1	–	49.5	13.8	0.2	0.8	16.8
nob_Latn	33.9	22.4	–	35.5	–	51.0	33.9	–	54.1	–	0.2	0.8	16.4
npi_Deva	13.9	7.5	–	26.4	–	29.0	15.0	–	41.8	–	0.4	1.6	34.7
nso_Latn	2.8	2.3	–	24.4	8.1	12.5	9.3	–	46.8	14.8	0.3	1.0	19.9
nus_Latn	0.1	0.2	–	13.2	–	1.8	1.9	–	26.6	–	0.7	1.9	30.3
nya_Latn	3.8	2.8	–	16.3	5.8	16.1	11.5	–	40.6	13.1	0.3	1.0	19.4
oci_Latn	25.4	18.1	–	37.9	–	46.1	31.0	–	54.3	–	0.2	0.8	17.9
ory_Orya	6.5	2.8	–	27.3	10.6	15.4	6.6	–	41.4	14.6	0.8	3.5	78.3
pag_Latn	4.6	4.6	–	18.6	–	18.2	14.7	–	42.7	–	0.2	0.8	16.3
pan_Guru	13.9	6.3	–	33.2	10.9	24.8	11.1	–	44.7	14.2	0.5	2.4	52.3
pap_Latn	21.4	18.2	–	39.0	–	43.5	31.0	–	55.6	–	0.2	0.8	17.4
pbt_Arab	3.8	2.4	0.3	21.1	–	14.8	8.9	0.9	36.2	–	0.3	1.4	29.1
pes_Arab	23.1	13.7	–	33.2	10.9	38.0	22.0	–	47.2	14.8	0.3	1.2	26.1
plt_Latn	6.6	4.2	–	23.3	7.1	25.4	15.6	–	46.1	14.0	0.2	1.0	20.3
pol_Latn	26.7	17.5	–	30.0	9.9	41.3	26.8	–	45.1	14.2	0.2	0.9	18.4
por_Latn	47.8	32.4	–	48.9	16.0	60.5	40.8	–	62.7	19.8	0.2	0.8	15.9
prs_Arab	20.2	12.6	–	31.1	–	35.1	21.6	–	49.3	–	0.3	1.2	25.4
quy_Latn	0.5	0.3	–	5.4	2.2	6.5	4.4	–	24.8	9.3	0.4	1.2	19.4
ron_Latn	38.6	25.3	2.6	41.3	13.7	53.3	34.6	3.4	56.6	17.8	0.2	0.9	18.1
run_Latn	2.4	1.2	–	18.1	–	12.8	7.2	–	39.2	–	0.3	1.0	19.7
rus_Cyrl	31.6	19.4	–	37.8	12.0	46.0	28.2	–	51.9	16.1	0.2	1.0	21.5
sag_Latn	0.1	0.0	–	9.7	–	2.8	2.1	–	32.9	–	0.6	1.4	19.3
san_Deva	3.5	2.0	–	7.3	2.7	16.1	8.5	–	24.0	8.3	0.4	1.7	35.7
sat_Olck	0.0	0.4	–	16.8	–	0.1	3.1	–	23.8	–	1.1	3.6	80.2
scn_Latn	9.3	6.9	–	22.5	–	29.9	19.7	–	43.2	–	0.2	0.9	18.9
shn_Mymr	0.3	0.3	–	13.6	–	4.1	3.2	–	31.0	–	0.8	4.1	93.0
sin_Sinh	3.7	1.9	–	32.8	11.0	11.9	5.5	–	39.9	14.0	0.6	2.7	57.0
slk_Latn	31.8	20.0	–	39.6	13.2	46.8	29.7	–	54.4	17.2	0.2	0.9	19.6
slv_Latn	29.7	19.3	–	35.2	11.6	46.2	30.0	–	51.9	16.3	0.2	0.9	18.2
smo_Latn	4.8	3.9	–	24.8	–	17.5	13.0	–	46.1	–	0.3	1.0	20.5
sna_Latn	2.4	1.7	0.4	18.2	5.7	11.5	8.0	1.4	40.0	12.1	0.3	1.0	20.3
snd_Arab	6.7	4.3	–	29.3	8.9	16.5	10.2	–	44.2	13.3	0.4	1.4	30.2
som_Latn	6.6	4.1	–	17.0	5.2	24.0	15.2	–	39.7	11.9	0.2	1.0	20.1
sot_Latn	4.5	2.7	–	19.1	6.1	16.1	10.4	–	42.5	13.1	0.3	1.0	20.1
spa_Latn	28.6	19.2	–	30.6	9.6	47.9	32.1	–	49.7	15.2	0.2	0.8	16.3
srd_Latn	13.6	9.8	–	33.0	–	35.0	23.2	–	51.3	–	0.2	0.9	18.8
srp_Cyrl	29.9	17.9	–	40.0	13.1	45.1	27.0	–	55.0	17.3	0.3	1.1	24.0
ssw_Latn	1.3	0.2	0.3	18.4	–	7.5	2.9	1.1	40.0	–	0.4	1.3	21.1
sun_Latn	11.6	7.9	–	19.9	6.7	32.5	20.9	–	41.3	13.3	0.2	0.8	17.8
swe_Latn	44.2	29.3	–	46.3	14.8	57.7	38.5	–	60.9	19.0	0.2	0.8	16.5
swh_Latn	31.5	20.5	–	34.0	12.2	49.8	32.1	–	54.1	17.6	0.2	0.9	18.6
szl_Latn	10.6	7.8	–	35.4	–	29.4	19.0	–	49.5	–	0.2	0.9	19.9
tam_Taml	8.8	4.0	0.4	33.4	10.6	22.0	10.3	0.8	49.0	15.3	0.5	2.4	51.2
taq_Latn	0.6	0.1	–	4.5	–	5.8	2.7	–	21.3	–	0.4	1.3	20.1
taq_Tfng	0.2	0.3	–	5.1	–	1.8	2.0	–	15.1	–	1.9	3.2	65.1
tat_Cyrl	5.0	3.1	–	28.0	8.3	16.1	9.9	–	43.1	13.2	0.3	1.4	28.8
tel_Telu	11.1	5.1	–	37.9	12.2	21.9	10.1	–	50.9	15.9	0.6	2.5	54.8
tgk_Cyrl	8.2	5.0	–	32.5	9.7	22.3	13.1	–	47.1	14.1	0.3	1.3	28.1
tgl_Latn	29.1	18.4	–	35.3	10.9	50.5	31.9	–	55.8	16.9	0.2	0.9	19.2
tha_Thai	25.4	13.5	–	32.3	12.4	32.7	17.4	–	39.2	13.6	0.3	1.5	32.3
tir_Ethi	1.0	0.6	–	16.2	4.8	3.6	2.0	–	23.5	7.2	0.6	2.4	51.9
tpi_Latn	11.5	8.3	1.1	16.4	–	30.5	20.6	2.5	38.4	–	0.2	0.9	18.9
tsn_Latn	2.9	2.0	–	23.6	–	12.8	9.0	–	44.7	–	0.3	1.1	20.4
tso_Latn	2.1	1.5	–	24.6	7.1	11.3	7.7	–	46.1	13.9	0.3	1.1	20.4
tuk_Latn	5.0	3.9	–	20.8	9.8	20.5	12.9	–	38.8	14.4	0.2	1.0	21.2
tum_Latn	2.8	1.3	–	12.3	–	12.8	6.9	–	32.5	–	0.3	1.1	22.2
tur_Latn	32.0	20.6	–	38.3	12.7	48.1	31.0	–	53.8	17.1	0.2	0.9	18.4
twi_Latn	2.1	1.4	–	14.0	4.7	9.5	6.5	–	34.9	11.2	0.4	1.2	21.8
tzm_Tfng	0.6	0.5	–	19.0	–	4.3	2.8	–	29.3	–	0.9	3.1	64.7
uig_Arab	4.7	3.1	–	28.0	11.0	14.8	9.1	–	41.5	14.8	0.4	1.7	37.0
ukr_Cyrl	29.6	17.5	1.5	36.9	11.7	43.6	25.9	2.2	51.9	16.0	0.3	1.1	24.4
umb_Latn	0.2	0.1	–	3.8	–	3.6	2.1	–	24.5	–	0.5	1.3	19.1
urd_Arab	16.5	8.9	–	28.0	8.9	31.3	16.8	–	44.9	13.7	0.3	1.5	32.2
uzn_Latn	14.3	9.7	–	27.7	10.3	32.7	21.1	–	46.7	15.4	0.2	0.9	19.8
vec_Latn	13.2	9.6	–	26.0	–	34.4	23.6	–	47.6	–	0.2	0.8	17.2
vie_Latn	33.1	20.4	–	39.9	–	47.6	28.9	–	54.9	–	0.2	1.0	21.4
war_Latn	20.0	13.3	1.4	32.3	–	40.7	26.3	2.7	53.0	–	0.2	0.9	18.7
wol_Latn	1.6	0.6	–	8.9	–	7.9	4.0	–	27.4	–	0.3	1.1	18.5
xho_Latn	4.2	3.1	–	23.4	8.1	17.4	12.0	–	44.8	14.3	0.3	0.9	19.2
ydd_Hebr	7.3	6.6	–	16.8	4.6	21.4	13.5	–	35.3	10.3	0.4	1.8	39.1
yor_Latn	1.9	1.5	–	9.7	1.3	8.7	6.1	–	23.5	5.5	0.3	1.2	24.3
yue_Hant	21.7	17.7	–	15.3	–	18.4	14.2	–	16.5	–	0.2	0.9	19.4

Continued on next page

Table 14 – continued from previous page

Lang.	spBLEU200/cost					chrF2++/cost					cost estimate (USD$)		
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4
zho_Hans	30.4	19.7	–	24.6	11.9	25.9	17.2	–	21.1	10.3	0.2	0.9	18.1
zho_Hant	24.1	15.8	–	11.5	–	20.5	13.5	–	12.9	–	0.2	0.9	19.7
zsm_Latn	34.8	23.1	–	42.0	13.0	54.2	35.9	–	61.4	18.6	0.2	0.8	16.7
zul_Latn	5.4	3.7	–	29.0	8.8	20.1	13.3	–	49.2	14.7	0.3	1.0	20.1