An Empirical Study of Many-Shot In-Context Learning for Machine Translation of Low-Resource Languages

Anonymous ACL submission

Abstract

In-context learning (ICL) allows large language models (LLMs) to adapt to new tasks from a few examples, making it promising for languages underrepresented in pre-training. Recent work on many-shot ICL suggests that modern LLMs can further benefit from larger ICL examples enabled by their long context windows. However, such gains depend on careful example selection, and the inference cost can be prohibitive for low-resource language communities. In this paper, we present an empirical study of many-shot ICL for machine translation from English into ten truly low-resource languages recently added to FLORES+. We analyze the effects of retrieving more informative examples, using out-of-domain data, and ordering examples by length. Our findings show that many-shot ICL becomes more effective as the number of examples increases. More importantly, we show that BM25-based retrieval substantially improves data efficiency: 50 retrieved examples roughly match 250 many-shot examples, while 250 retrieved examples perform similarly to 1,000 many-shot examples. We further show that ICL provides additional gains on top of fine-tuning.

1 Introduction

Large Language Models (LLMs) have shown strong performance on many natural language processing (NLP) tasks, including generation tasks such as summarization and machine translation (MT) in multilingual settings (Pu et al., 2023; Zhu et al., 2024; Gemini-Team et al., 2024). However, their performance remains limited for very low-resource and endangered languages.

In-context learning (ICL) (Brown et al., 2020; Dong et al., 2024), which leverages few-shot examples, is one of the most promising emergent abilities of LLMs and has shown strong potential for low-resource languages (LRLs) (Lin et al., 2022; Zhang et al., 2023). However, performance does not necessarily improve as the number of few-shot examples increases, leading many studies to limit the number of shots (e.g., to less than 20) (Zhang et al., 2024; Tanzer et al., 2024; Court and Elsner, 2024). This issue has led to various streams of work focusing on choosing the best few-shot examples, either through similarity between examples, retrieving the most similar to target examples, domain effects, and use of pseudo parallel data (Agrawal et al., 2023; Marashian et al., 2025; Pei et al., 2025). However, most of these studies do not cover LRLs where retrieval is often weaker.

Recently, a new paradigm has emerged that focuses on prompting with large numbers of examples, or "many-shot" prompting (e.g., 1,000 examples) (Agarwal et al., 2024). This approach seems beneficial for LRLs that had limited representation during pretraining, and can make better use of a few thousand examples than fine-tuning, which typically requires more data (Adelani et al., 2022; Vieira et al., 2024). However, despite its promise, the high inference cost makes it challenging for practical use. Exploring more effective many-shot example selection could improve its scalability for low-resource language communities.

In this paper, we investigate the effectiveness of many-shot machine translation (MT) from English into 10 truly LRLs with limited exposure during LLM pretraining. We select these languages carefully based on the recency of the benchmarks, focusing on newly added languages in FLORES+. We further examine how to reduce inference costs by retrieving a smaller set of more effective many-shot examples using simple methods such as BM25. Besides, we study the impact of out-of-domain examples, such as religious texts, on in-domain translation performance for Wikipedia, as well as the effect of example ordering from short to long, inspired by curriculum learning. We finally compare many-shot ICL against fine-tuning of an open-weight model (`Gemma 3 27B`) .

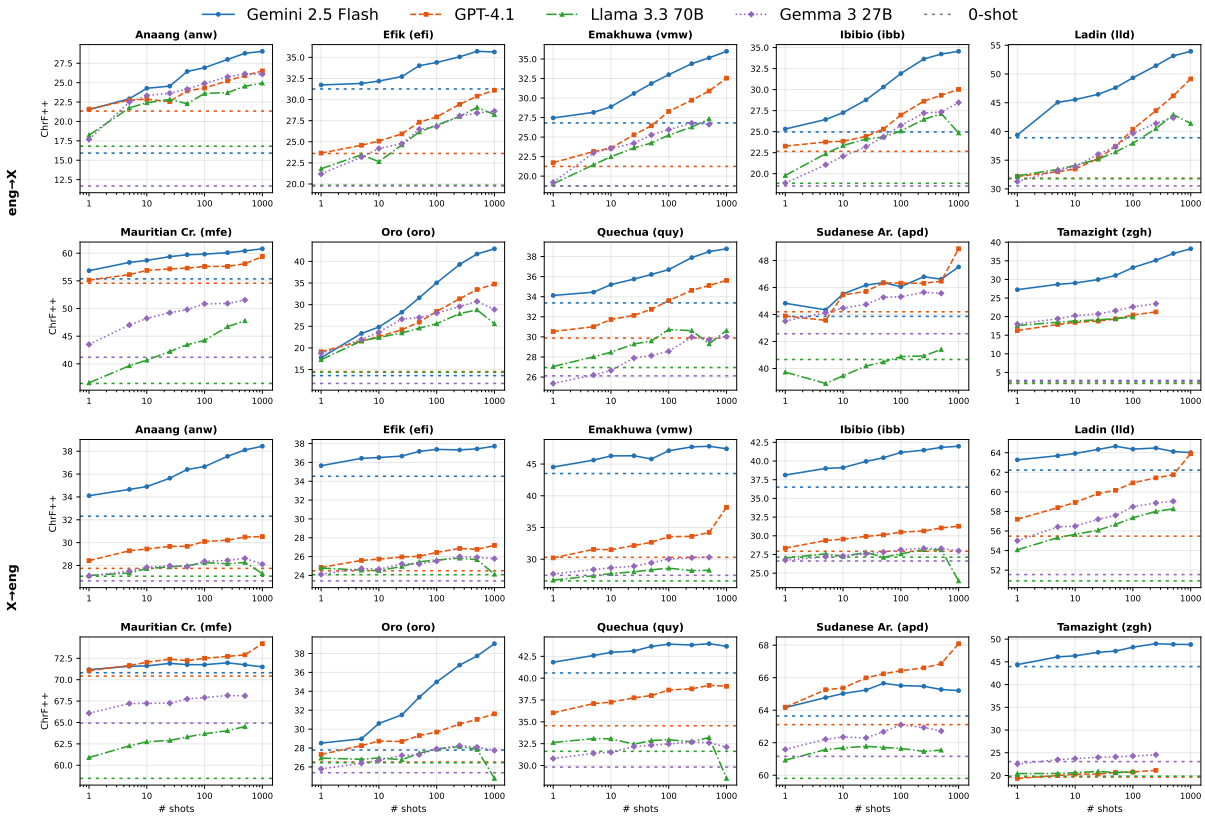

Figure 1: Per-language scaling curves (chrF++, random selection). Top two rows: eng→X; bottom two rows: X→eng.

Our evaluation on both open-weight and proprietary models, such as `Gemini 2.5 Flash` and `GPT-4.1`, shows the effectiveness of the many-shot approach on these languages, where performance more than doubles in some languages with more examples. Similarly, leveraging a simple BM25 retrieval from English examples drastically reduces cost, where BM25 with 50 examples roughly matches 250 many-shot examples, and BM25 with 250 examples has similar performance as 1,000 many-shot examples. We find that many shots from within the same domain are consistently better for most languages; however, some languages are not affected by this domain mismatch, which shows some promise in the case of absence of in-domain examples. Furthermore, the ordering of examples does not seem to have a strong effect on many-shot performance in our experiments. Finally, we show that ICL with only 10–25 retrieved examples on open-source models without fine-tuning matches fine-tuning on the full 1,012-sentence pool, and increasing the number of samples continues to improve performance even on top of fine-tuning.

2 Experimental Setup

Focus Languages We focus on ten extremely low-resource languages that were recently added to FLORES+ or extended from FLORES-200 (NLLB Team et al., 2024), including four Nigerian languages (Anaang, Efik, Ibibio, and Oro), Sudanese Arabic, Emakhuwa, Ladin, Mauritian Creole, Tamazight, and Quechua[1]. Eight languages use the Latin script; the exceptions are Tamazight (Tifinagh) and Sudanese Arabic (Arabic script). Appendix A provides more details.

Experimental Design For each test sentence, we construct a prompt containing a task instruction, k parallel example pairs (source and target), and the test sentence as a query; the model then generates the translation in one pass. We test $k \in \{0, 1, 5, 10, 25, 50, 100, 250, 500, 1000\}$; the full prompt template appears in Appendix B. Our evaluation targets three research questions.

(1) Scaling We evaluate with randomly sampled examples across several k, and ten languages in two directions ("eng→X" and "X→eng"), us-

[1]While Tamazight and Quechua were originally part of FLORES-200, Tamazight is improved by the community.

Shots	Method	Anaang	Sudanese	Efik	Ibibio	Ladin	Mauritian	Oro	Quechua	Emakhuwa	Tamazight	Avg
						Gemini 2.5 Flash, eng → X (chrF++)						
5	BM25	**25.8**	**45.7**	**33.5**	**28.7**	**47.6**	**58.6**	**28.6**	**35.9**	**30.9**	**32.0**	36.7
	Random	22.9	44.3	31.9	26.4	45.1	58.3	23.3	34.5	28.2	28.7	34.4
	Δ	+2.9	+1.4	+1.6	+2.3	+2.5	+0.3	+5.3	+1.4	+2.7	+3.3	+2.4
50	BM25	**28.2**	**46.9**	**35.2**	**32.8**	**51.8**	**60.3**	**38.8**	**38.0**	**34.2**	**36.4**	40.3
	Random	26.4	46.4	34.0	30.3	47.6	59.7	31.6	36.2	31.9	31.1	37.5
	Δ	+1.8	+0.5	+1.2	+2.5	+4.2	+0.6	+7.2	+1.8	+2.3	+5.3	+2.7
250	BM25	**28.9**	**46.8**	**35.4**	**34.2**	**53.2**	**60.5**	**41.8**	**38.6**	**35.3**	**37.7**	41.2
	Random	28.0	**46.8**	35.1	33.6	51.4	60.1	39.3	37.9	34.4	35.1	40.2
	Δ	+0.9	0.0	+0.3	+0.6	+1.8	+0.4	+2.5	+0.7	+0.9	+2.6	+1.1
1,000	BM25	**29.5**	47.0	**35.9**	34.4	**54.1**	**61.1**	**43.2**	**38.8**	35.7	**38.3**	41.8
	Random	29.0	**47.5**	35.6	**34.5**	53.9	60.8	42.9	**38.8**	**36.0**	38.2	41.7
	Δ	+0.5	–0.5	+0.3	–0.1	+0.2	+0.3	+0.3	0.0	–0.3	+0.1	+0.1

Table 1: BM25 vs. Rand (chrF++, eng→X, Gemini). **Bold**: winner. Δ: green = BM25 better, red = Random better.

ing `Gemini 2.5 Flash` (Comanici et al., 2025), `GPT-4.1`, `Llama 3.3 70B` (Grattafiori et al., 2024) and `Gemma 3 27B` (Team et al., 2025).

(2) Example Selection We compare selecting ICL examples randomly vs. BM25 (Robertson and Zaragoza, 2009) retrieval on the English source side and therefore restrict this experiment to eng→X. We report other retrieval models such as Qwen Embedding model in Appendix F.

(3) Domain Mismatch On the 7 languages where Bible translations exist[2] (Table 6), we compare Bible examples against in-domain BM25 and random selection, all in the "eng→X" direction. We retrieve Bible examples with BM25, the same setup as the in-domain experiments, so differences in performance reflect domain rather than retrieval. This tests whether out-of-domain data benefits from scaling the way in-domain data does.

3 Results

Here we mainly report results with chrF++; results with spBLEU, shown in Appendix E, lead to similar conclusions.

3.1 Effect of Scaling Many-Shot Examples

Figure 1 shows the results for different numbers of ICL examples from $k = 1, 5, .., 1000$. The results show consistent improvement across all LLMs, especially the proprietary models (`Gemini 2.5 Flash` and `GPT-4.1`). In the "eng→X" direction, `Gemini 2.5 Flash` achieves the best overall results across all shot settings, with performance gains ranging from 3.7 to 35.6 from 0-shot to 1,000-shot. Performance is also moderately higher when comparing a few-shot setting, say 10-shot to a many-shot setting, say 1,000-shot, with gains of 2.0 to 18.1. Tamazight benefited the most from few-shot (1-shot) while Oro benefited the most from many-shot; We discuss possible explanations and provide human evaluation in Appendix I. In contrast, the "X→eng" direction shows significantly less improvement than generation into a LRL. However, the trend remains consistently positive except for Sudanese Ar. and Mauritian Cr., which may reflect confusion with related dominant languages such as Arabic and French.

Open weight models experience in-consistencies in higher many-shot. `Gemma 3 27B` and `Llama 3.3 70B` sometimes fail at $k = 500$ or $k = 1,000$ due to limited context window, and weaker ability to process large numbers of ICL examples. Many-shot is still most effective for proprietary models, aligning with the base LLM's competence.

Larger many-shot does not always lead to the best performance. In some cases, $k = 250$ or $k = 500$ leads to better results than $k = 1,000$, especially in the "X→eng" direction, while for some other languages the performance keeps increasing.

3.2 Discussion

Example Selection: BM25 vs. Random Table 1 compares random sampling of ICL examples to BM25 retrieval on the English source side (eng→X).[3] BM25 outperforms random selection for nearly all languages, especially at low shot counts, although the difference narrows as more examples are added. Most importantly, it is more data efficient: 50-shot with BM25 achieves the same average performance as 250-shot of Random ICL (≈ 40.2), and 250-shot BM25 matches 1,000-

[2]We excluded Sudanese Arabic Bible following native speaker feedback that it is too similar to Standard Arabic.

[3]Applied to the English side because applying BM25 to LRLs may yield inconsistent retrieval across languages.

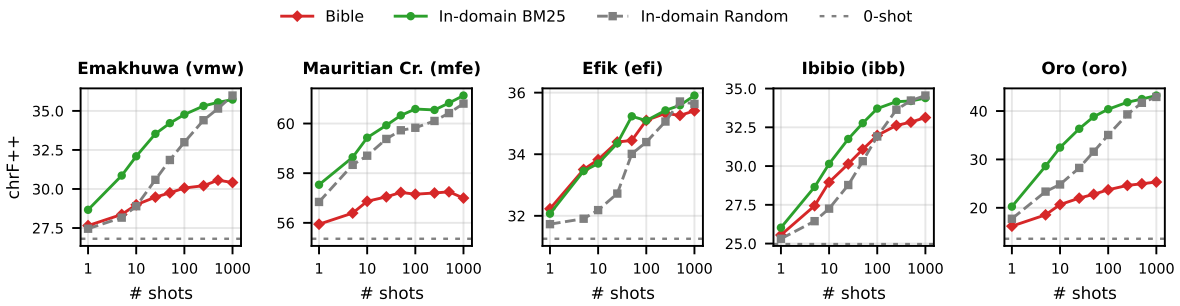

Figure 2: Bible vs. in-domain examples (chrF++, eng→X, Gemini 2.5 Flash). Bible examples plateau or degrade with more shots, while in-domain examples scale consistently.

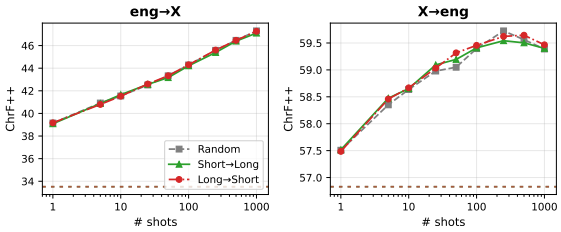

Figure 3: Effect of example ordering on MT (chrF++, Gemini 2.5 Flash, averaged across vmw, zgh, lld, mfe, and apd). **S2L** and **L2S** sort examples by source length.

Variant	0	5	10	25	50	250
Base `Gemma`	19.8	25.5	27.0	28.4	29.2	29.3
+ LoRA_{250}	24.1	26.3	27.0	28.2	28.9	29.4
+ LoRA_{1012}	27.3	28.5	28.9	29.3	29.7	30.1

Table 2: LoRA fine-tuning combined with many-shot ICL on eng→Efik (chrF++): columns are BM25-retrieved in-context shots.

shot random ICL. This provides more practical usability of many-shot for low-resource communities, drastically reducing inference cost.

Effect of Out-of-Domain Examples Figure 2 shows that many-shot ICL from within the same domain (Wikipedia) is more effective than that of out-of-domain (Bible) across all languages. However, some languages do benefit more from the religious domain than others, such as Efik and Ibibio. This finding suggests that all hope is not lost if in-domain examples are not available; out-of-domain examples may still be useful when they remain distributionally close to in-domain text—a closeness that depends on the language property.

Effect of Ordering ICL Examples Figure 3 shows the results of ordering examples from Short-to-Long (S2L) and Long-to-Short (L2S). We did not observe any significant impact of ordering by length, especially in eng→X. Semantic relevance via BM25 retrieval seems to be more important.

Comparison with Fine-tuning Fine-tuning is an alternative to many-shot ICL for open-weight users. We LoRA fine-tune `Gemma 3 27B` on the full 1,012-sentence pool (LoRA_{1012}) and on a 250-sentence BM25-selected subset (LoRA_{250}) for eng→{Efik, Ibibio}, evaluated at 0 to 250 BM25 shots. For eng→Efik, without ICL examples, LoRA_{1012} reaches 27.3 chrF++, which is slightly below base `Gemma 3 27B` with 25 shots (28.4), while LoRA_{250} reaches 24.1, slightly below the base model with 5 shots (25.5). LoRA and ICL are complementary, but the marginal benefit of LoRA quickly narrows. Adding ICL on top of LoRA still improves performance, even though the examples are drawn from the same pool used for fine-tuning. While fine-tuning is helpful before ICL, open-weight models still lag behind the closed-source baseline: `Gemini 2.5 Flash` with only 5 shots already exceeds LoRA_{1012} at 250 shots (33.5 vs. 30.1). Also, given the huge parameter sizes of proprietary LLMs, fine-tuning is typically not an option, but ICL can be always be leveraged. eng→Ibibio follows the same qualitative pattern. See full result of fine-tuning in Appendix G.

4 Conclusion

In this paper, we study many-shot in-context learning for MT across ten LRLs, scaling from 0 to 1,000 parallel examples and evaluating four LLMs. Performance improves roughly log-linearly as the number of ICL examples increases, with larger gains for translation into LRLs (eng→X) than for X→eng, which shows only moderate improvements. For example selection, retrieval based on similarity on the English side improves sample efficiency: 50 BM25-retrieved examples match the performance of 250 randomly selected ones, and 250 BM25 examples match 1,000 random ICL examples. Out-of-domain ICL examples yield smaller overall gains than in-domain examples, although they remain beneficial for some LRLs.

Limitations

All language pairs in this study involve English as either the source or target language; we did not evaluate non-English-centric translation, which may exhibit different scaling behavior.

Furthermore, we used a single fixed prompt template and did not explore prompt engineering or chain-of-thought strategies, which could affect scaling in ways we have not measured. Additionally, our evaluation is limited to automatic metrics (spBLEU, chrF++); human evaluation would provide a more complete picture of translation quality, particularly for languages where automatic metrics may be less reliable. We also did not explore embedding-based metrics such as COMET (Rei et al., 2020) and MetricX (Juraska et al., 2023) because the languages studied are extremely low-resource and not covered by these metrics.

More broadly, parallel data for extremely low-resource languages may contain annotation errors or inconsistencies, which can affect both the quality of in-context examples and the reliability of reference translations for these automatic evaluation. Hence, translation outputs for these languages should be verified by native speakers before deployment.

Finally, it is important to note that the full set of experiments in this paper required more than $30,000 in API credits, highlighting a significant barrier to reproducibility for low-resource language communities. To mitigate this, we release all results and analyses to reduce the need for others to replicate these costly runs.

References

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, and 26 others. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie C. Y. Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-Shot In-Context Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Felermino Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. Expanding FLORES+ benchmark for more low-resource settings: Portuguese-emakhuwa machine translation evaluation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 579–592, Miami, Florida, USA. 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, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Sara Court and Micha Elsner. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Andrew Drozdov, Honglei Zhuang, Zhuyun Dai, Zhen Qin, Razieh Rahimi, Xuanhui Wang, Dana Alon, Mohit Iyyer, Andrew McCallum, Donald Metzler, and Kai Hui. 2023. PaRaDe: Passage ranking using demonstrations with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14242–14252, Singapore. Association for Computational Linguistics.

Samuel Frontull, Thomas Ströhle, Carlo Zoli, Werner Pescosta, Ulrike Frenademez, Matteo Ruggeri, Daria

Valentin, Karin Comploj, Gabriel Perathoner, Silvia Liotto, and Paolo Anvidalfarei. 2025. Bringing Ladin to FLORES+. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1061–1071, Suzhou, China. Association for Computational Linguistics.

Gemini-Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Oluwadara Kalejaiye, Luel Hagos Beyene, David Ifeoluwa Adelani, Mmekut-mfon Gabriel Edet, Aniefon Daniel Akpan, Eno-Abasi Urua, and Anietie Andy. 2025. Ibom NLP: A step toward inclusive natural language processing for Nigeria's minority languages. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 372–382, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Xiaonan Li and Xipeng Qiu. 2023. Finding support examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6219–6235, Singapore. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024. In-context learning with retrieved demonstrations for language models: A survey. *Transactions on Machine Learning Research*. Survey Certification.

Ali Marashian, Enora Rice, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2025. From priest to doctor: Domain adaptation for low-resource neural machine translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7087–7098, Abu Dhabi, UAE. Association for Computational Linguistics.

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, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Alp Oktem, Mohamed Aymane Farhi, Brahim Essaidi, Naceur Jabouja, and Farida Boudichat. 2025. Correcting the tamazight portions of FLORES+ and OLDI seed datasets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1072–1080, Suzhou, China. Association for Computational Linguistics.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 2025. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics.

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

Yush Rajcoomar. 2025. KozKreolMRU WMT 2025 CreoleMT system description: Koz kreol: Multistage training for English–mauritian creole MT. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1183–1190, Suzhou, China. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Luis Frentzen Salim, Esteban Carlin, Alexandre Morinvil, Xi Ai, and Lun-Wei Ku. 2026. Beyond many-shot translation: Scaling in-context demonstrations for low-resource machine translation. *arXiv preprint arXiv:2602.04764*.

Hadia Mohmmedosman Ahmed Samil and David Ifeoluwa Adelani. 2026. Sudanese-flores: Extending FLORES+ to sudanese arabic dialect. In *7th Workshop on African Natural Language Processing*.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *The Twelfth International Conference on Learning Representations*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, and et al. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Inacio Vieira, Will Allred, Séamus Lankford, Sheila Castilho, and Andy Way. 2024. How much data is enough data? fine-tuning large language models for in-house translation: Performance evaluation across multiple dataset sizes. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 236–249, Chicago, USA. Association for Machine Translation in the Americas.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. In-context example selection via similarity search improves low-resource machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International conference on machine learning*, pages 41092–41110. PMLR.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024. Teaching Large Language Models an Unseen Language on the Fly. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8783–8800, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Languages

We provide the details of the studied languages in Table 6.

ICL examples are sampled from the devtest split (1,012 sentences), and evaluation is performed on the dev split (997 sentences).

For the Bible data, we use versions obtained from bible.com. These texts are used strictly for research purposes without redistribution, in accordance with their terms of use.

B Prompt Templates

All experiments use the following prompt format without any dictionary augmentation.

Many-shot prompt ($k \geq 1$ examples).

```
You are an expert translator. I am going to
give you one or more example pairs of text
snippets where the first is in {SRC} and the
second is a translation of the first snippet
into {TGT}. The sentences will be written
{SRC}: {example_src_1}
{TGT}: {example_tgt_1}
{SRC}: {example_src_2}
{TGT}: {example_tgt_2}
...
{SRC}: {example_src_k}
{TGT}: {example_tgt_k}
After the example pairs, I am going to provide
another sentence in {SRC} and I want you to
translate it into {TGT}. Give only the
translation, and no extra commentary,
formatting, or chattiness. Translate the text
from {SRC} to {TGT}.
{SRC}: {query}
{TGT}:
```

Zero-shot prompt.

```
You are an expert translator. Translate the
text from {SRC} to {TGT}. Give only the
translation, and no extra commentary,
formatting, or chattiness.
{SRC}: {query}
{TGT}:
```

Here `{SRC}` and `{TGT}` are replaced with the full language names (e.g., "English", "Ibibio"). In the many-shot setting, the k example pairs are drawn from the example pool, either by random sampling or BM25 retrieval (Appendix D).

C Related Work

While ICL (Brown et al., 2020) has been shown to be very effective on tasks such as machine translation (Dong et al., 2024), performance is highly sensitive to the choice of demonstrations (Luo et al., 2024), including factors such as example order (Lu et al., 2022), diversity (Li and Qiu, 2023), and difficulty (Drozdov et al., 2023). In MT, Agrawal et al. (2023) show that relevance-based retrieval improves few-shot ICL for high-resource languages. However, it remains unclear how example selection interacts with scale in the many-shot setting, particularly for extremely LRLs.

Agarwal et al. (2024) challenge this paradigm by scaling ICL to hundreds or thousands of examples with Gemini 1.5 Pro, showing substantial improvements across summarization, reasoning, and MT, including low-resource translation into Bemba and Kurdish. Salim et al. (2026) investigate scaling in-context token budget to 1M tokens, but the evaluation is limited to smaller models like `Qwen 2.5 7B` with worse performance for larger shots. Our work extends the many-shot setting to a systematic study of 10 extremely LRLs, with a particular focus on retrieval-based example selection and domain effects.

D Retrieval Details

BM25 example selection uses BM25 (Robertson and Zaragoza, 2009) with default parameters. Indexing and querying both use the English source side; English sentences are tokenized by lowercasing, replacing punctuation with whitespace, removing digits, and splitting on whitespace. For each test sentence, we score every pool sentence with BM25 and select the top-k by descending score; the corresponding parallel pairs are then placed in the prompt according to one of three orderings:

- **Default** (used in the main BM25 results, §3.2): randomly shuffle the top-k pairs with a fixed seed.
- **Similar first** (Table 19): place pairs in descending BM25 score, so the most-similar pair appears at the start of the prompt (farthest from the query).
- **Dissimilar first** (Table 19): reverse of *similar first*, so the most-similar pair appears immediately before the query.

E Full Model Results

eng→X. Each table reports Random (in-domain), BM25 retrieval, BM25−Random delta (Δ), and Bible (cross-domain) results across all shot counts. Gemini 2.5 Flash: Table 9 (chrF++) and Table 10 (spBLEU). GPT-4.1: Table 11 (chrF++) and Table 12 (spBLEU). Llama 3.3 70B: Table 13 (chrF++) and Table 14 (spBLEU). Gemma 3 27B: Table 15 (chrF++) and Table 16

(spBLEU). All eight tables share the following formatting: the six languages with a Bible translation appear on the left; the four without (✗) on the right. **Bold** marks the best score per column. Gray indicates the 0-shot baseline. Δ values are green when BM25 outperforms Random and red otherwise. Blue at the bottom shows the gain from Random at 1,000 shots over 0-shot.

X→eng (random selection only). Table 17 (chrF++) and Table 18 (spBLEU) report X→eng scaling results with random selection across all four models. Gray indicates the 0-shot baseline; **bold** marks the best per column. The *gain* row shows the ratio of the best score to 0-shot (blue; **green bold** for ≥2×).

Missing entries (–) indicate runs that exceeded the model's context window.

F Ordering Comparison

We test whether the order in which examples appear in the prompt affects translation quality. Table 19 varies the ordering of retrieved examples (BM25 and dense retrieval); Table 20 sorts randomly selected examples by sentence length. In both cases, differences are negligible and no strategy consistently wins.

G Fine-tuning Comparison

Table 7 reports per-shot, per-direction spBLEU and chrF++ for the case study in §3.2, in the same stacked format as Table 20. `Gemini 2.5 Flash` is included as a closed-API reference; its BM25 values are copied from Table 9 and Table 10. Aggregate averages reported in the main text are computed over the two directions.

Cost considerations. Translating the 997-sentence query set using 250 shots ICL costs approximately $2–$3 per direction with `Gemini 2.5 Flash` (Batch API rate) for typical Latin-script LRLs (e.g., Mauritian Creole, Ladin, Efik, Ibibio), and up to $12 for under-represented scripts such as Tifinagh; `GPT-4.1` costs approximately $46 per direction at standard pricing on eng→Efik, or $23 at the Azure Batch rate we used in our experiments.

The cost of fine-tuning an open-weight alternative depends on the choice of model and compute budget. Renting a single A100 80GB GPU typically costs more than $2 per hour by most providers such as Lambda (A100 cost $2.79) , but the total cost scales with model size and training-pass duration. Deployment cost for self-hosted open-weight models is harder to characterize, as it depends on the user's available hardware and throughput requirements; matching the translation quality of the strongest closed-API models in our experiments would typically require open-weight models substantially larger than those we fine-tuned, which are themselves nontrivial to self-host efficiently.

For practitioners with access to a closed-API model, it delivers both the strongest quality we measure (e.g., 35.4 vs. 30.1 chrF++ at 250 shots on eng→Efik; Table 7) and a cost on the order of a few dollars per test set, accessible to low-resource language communities without infrastructure investment. For open-weight-only deployments, based on our experimental results, you should only fine-tune without using ICL if you need to reuse the model extensively and only require a relatively mediocre translation performance. If you are only translating a small amount, then ICL alone is sufficient to achieve the same performance.(Based on our experiments, open-source models only need 10-25 shots to achieve the performance of model fine-tuned using full 1012 sentence example pools, which is much cheaper than the 250 shots we discussed.). If you want to have the best possible performance, you still need to add ICL, and in this case, fine-tuning may no longer be necessary, because in our experiments, under 250-shot ICL, the performance with and without loRA fine-tuning was very similar.

The total of more than $30,000 reported in our Limitations corresponds to our actual API billing across the full experimental program, including pilot and exploration runs beyond the configurations reported in the main results. Because input prompt length—and thus per-token API cost—scales with the number of shots, less than 25% of this total was spent on experiments at $k \leq 250$; the remaining ∼75% comes from high-shot scaling experiments at $k \in \{500, 750, 1000\}$ (only the $k = 500$ and $k = 1000$ endpoints are reported in the main scaling figures).

H Dictionary Augmentation

Bilingual dictionaries are a common augmentation for prompt-based MT in low-resource settings, with prior work reporting gains at low shot counts (Ghazvininejad et al., 2023; Zhang et al., 2024; Pei et al., 2025). These gains, however, de-

pend on the availability of comprehensive, high-quality dictionaries — a condition that is often not met for extremely low-resource languages, whose digital dictionaries (where they exist at all) tend to be small, incomplete, or prone to OCR and typographical errors. We test whether dictionary augmentation persists in the many-shot regime under such realistic LRL conditions on eng→Efik and eng→Anaang, the two directions for which we have a digital dictionary, comparing a 25-shot BM25 prompt augmented with the dictionary against both the matched 25-shot baseline and a 50-shot baseline without the dictionary (Table 8). At a matched 25-shot budget, the dictionary yields only marginal gains on `GPT-4.1` (+0.55 and +0.05 chrF++) and slight losses on `Gemini 2.5 Flash` (−0.41 and −0.19). At a matched prompt budget, adding the dictionary on top of 25 shots is consistently outperformed by simply scaling to 50 retrieved examples (all four Δ_{50} values negative). Two factors plausibly contribute: the limited coverage and quality of the digital dictionaries available for these languages, and the diminishing marginal value of a separate dictionary once retrieved examples already supply relevant target-language vocabulary. We therefore do not interpret these results as a refutation of dictionary augmentation in general, which remains useful at lower shot counts and with higher-coverage dictionaries.

I Qualitative Analysis

Per-language scaling patterns. We provide a post-hoc analysis of the language-specific scaling patterns observed in §3.1. Following Zebaze et al. (2025), we view ICL gains as reflecting two largely independent bottlenecks: *task and target-language anchoring*, which may be resolved with very few examples; and *target-language generation quality*, which may continue to benefit from additional lexical, orthographic, and syntactic evidence. Under this interpretation, Tamazight's large 0→1-shot gain is consistent with a script-anchoring effect: Tifinagh is severely under-represented in pretraining, so the model fails to produce target-script output at 0-shot (near-zero chrF++), and a single in-context example may suffice to anchor it to the correct target script. We verified this by inspecting actual outputs: at 0-shot, the model produces valid Tamazight content transliterated in Latin script rather than written in Tifinagh; a single Tifinagh in-context example switches the output to the correct

Language	Script	0→1	1→1000
Tamazight	Tifinagh	+24.6	+11.0
Oro	Latin + diacritics	+4.1	+25.2

Table 3: 0→1 and 1→1000 chrF++ gains for Tamazight and Oro (Random selection), computed from Table 9.

English (source)	The pilot was identified as Squadron Leader Dilokrit Pattavee.
Reference (Tifinagh)	ⵉⵣⵔⵉⵍ ⵏ ⵓⵎⵙⵙⵓⵏ ⵙⴳⴰ ⵓⵎⵣⵣⵓⵍ ⵉ ⵉⴷⵍⵓⵙⵓ ⴷⵉⵏⵉⴽⵓⵙⵜ ⵜⵓⵜⵓⵔⵉⵙ
`Gemini Flash` **0-shot**	Amhar ittwasn d Amghar n Tagrawt Dilokrit Pattavee.
`Gemini Flash` **1-shot**	ⵉⵍⵓⵣ ⵙⵜⵜⵍⵓⵙⵙⵉ ⴷ ⵉⵎⵙⵙⴳⴰⵔ ⵉ ⵜⵓⴳⵓⵎⵜ ⴷⵉⵏⵉⴽⵓⵙⵜ ⵜⵓⵜⵓⵔⵉⵙ

Table 4: Illustrative example of script anchoring on eng→Tamazight. At 0-shot, `Gemini 2.5 Flash` produces valid Tamazight content transliterated in Latin script; a single Tifinagh in-context example switches the output to Tifinagh. The proper name "Dilokrit Pattavee" is preserved across all three versions (left in Latin at 0-shot, transliterated to Tifinagh at 1-shot and in the reference), confirming the failure mode is target-script anchoring rather than translation quality.

script (Table 4). By contrast, Oro uses an extended Latin script (with diacritics) but remains extremely low-resource; its continued gains with more shots suggest that additional examples may provide useful target-language evidence beyond task identification. The remaining eight languages show intermediate scaling behavior (see Table 9 for the full scaling curves). We emphasize that these are hypotheses supported by observed scaling patterns and language metadata, not causal claims.

Native-speaker assessment for Oro. To complement the aggregate chrF++ trends in §3.1, we asked a native speaker of Oro to rate eng→Oro translations from `Gemini 2.5 Flash` using Direct Assessment (DA, 0–100; higher is better) for a single FLORES+ sentence across shot counts (Table 5). The example illustrates that low-shot chrF++ gains can be misleading: at $k \leq 10$ the model produces fluent-looking output that mimics the target-language orthography but is rated as not useful (DA=0); only at $k \geq 100$ does the translation become adequate to the native speaker. We observed a similar pattern for Ibibio. We report this as an illustrative example rather than a systematic human evaluation.

Setting	Translation output	DA
Source (eng): The valves opening allowed a pressure release for the system and oil flowed on a pad to a tank that can hold 55,000 barrels (2.3 million gallons).		
Reference (oro): Ku inuẹ onu ku ovim ẹtei osi ọwọńọ kini ara ẹwẹ aga ọdọńọ ku ata ogbono uga omu okeme idọńọ izim miniọn otu ara 55,000 (miniọn gallon 2.3).		
1-shot	Nmeghe mkpọkọ ahụ mere ka nrụgide tọhapụ maka usoro ahụ, ma mmanụ wee reree n'elu ebe gaa na tankị nke nwere ike ijide barel puku iri ise na ise (galọn nde abụọ na puku narị atọ).	0
10-shot	Ima valves egegin ọwọ ku ufene ifrẹ ọnọ imu udip ma ọkpọ-ọkpọ ọgọ ku pad ku tank omu ekeme imuọ 55,000 barrels (2.3 igbẹghẹ gallons).	0
100-shot	Kini ima valve egebuana, emu egeyak ulu ozighi ọnọ ubonokpọ onu, ma ara ọgọwọñ ku ufene edi ufene-mmọñ omu okeme imuọ mkpọ su barrel 55,000 (miniọn 2.3 galọn).	50
250-shot	Ima valve omu egedi ntuñọ ẹgeyak ntuñọ ulagha ọnọ ubonokpọ ma ara ọgọwẹ ku ogbogro aga udọñ omu okeme imuọ 55,000 barrel (miniọn 2.3 galọn).	60

Table 5: Illustrative native-speaker Direct Assessment (DA, 0–100; higher is better) of eng→Oro translations from `Gemini 2.5 Flash` for a single FLORES+ sentence, across shot counts k. Low-shot outputs mimic the target-language orthography but are inadequate (DA=0); adequacy emerges only at $k \geq 100$.

Language	Family	Morphology	Script	Dataset	Has a Bible?
Emakhuwa (vmw)	Atlantic–Congo / Bantu	Agglutinative	Latin	FLORES+[a]	✓
Moroccan Tamazight (zgh)	Afro-Asiatic / Berber	Fusional	Tifinagh	FLORES+[b]	✗
Ladin (lld)	Indo-European / Romance	Fusional	Latin	FLORES+[c]	✗
Mauritian Creole (mfe)	French Creole	Analytic	Latin	FLORES+[d]	✓
Ay. Quechua (quy)	Quechuan	Agglutinative	Latin	FLORES+[e]	✓
Sudanese Arabic (apd)	Afro-Asiatic / Semitic	Fusional	Arabic	Sudanese-Flores[f]	✓
Anaang (anw)	Atlantic–Congo / Cross River	Agglutinative	Latin	Ibom-NLP[g]	✗
Efik (efi)	Atlantic–Congo / Cross River	Agglutinative	Latin	Ibom-NLP[g]	✓
Ibibio (ibb)	Atlantic–Congo / Cross River	Agglutinative	Latin	Ibom-NLP[g]	✓
Oro (oro)	Atlantic–Congo / Cross River	Agglutinative	Latin	Ibom-NLP[g]	✓

[a] (Ali et al., 2024) [b] (Oktem et al., 2025) [c] (Frontull et al., 2025) [d] (Rajcoomar, 2025)
[e] (NLLB Team et al., 2024) [f] (Samil and Adelani, 2026) [g] (Kalejaiye et al., 2025)

Table 6: Overview of 10 target languages. The bottom four are from the Ibom region of Nigeria.

Shots	Variant	spBLEU / chrF++	
		efi	ibb
0	Base Gemma	3.6/19.8	3.8/18.5
	+ LoRA(250)	6.4/24.1	6.5/24.2
	+ LoRA(1012)	8.4/27.3	**8.0/26.4**
	Gemini 2.5 Flash	**9.1/31.3**	6.3/25.0
1	Base Gemma	4.8/22.1	4.9/20.2
	+ LoRA(250)	6.3/24.4	6.4/24.0
	+ LoRA(1012)	8.9/27.8	**8.7/26.9**
	Gemini 2.5 Flash	**10.1/32.1**	7.5/26.0
5	Base Gemma	6.8/25.5	6.7/23.2
	+ LoRA(250)	7.5/26.3	7.5/25.2
	+ LoRA(1012)	9.3/28.5	9.1/27.5
	Gemini 2.5 Flash	**11.5/33.5**	**10.1/28.7**
10	Base Gemma	7.7/27.0	7.7/24.7
	+ LoRA(250)	8.0/27.0	8.0/25.8
	+ LoRA(1012)	9.9/28.9	9.5/27.7
	Gemini 2.5 Flash	**11.9/33.7**	**11.5/30.1**
25	Base Gemma	8.9/28.4	9.1/26.5
	+ LoRA(250)	9.1/28.2	8.8/26.4
	+ LoRA(1012)	10.3/29.3	9.9/28.1
	Gemini 2.5 Flash	**12.6/34.4**	**13.1/31.7**
50	Base Gemma	9.7/29.2	9.7/27.1
	+ LoRA(250)	9.5/28.9	9.4/27.0
	+ LoRA(1012)	10.7/29.7	10.4/28.6
	Gemini 2.5 Flash	**13.5/35.2**	**13.9/32.8**
100	Base Gemma	10.3/29.5	9.7/27.6
	+ LoRA(250)	9.9/29.2	9.5/27.1
	+ LoRA(1012)	10.7/29.6	10.6/28.8
	Gemini 2.5 Flash	**13.4/35.1**	**14.7/33.7**
250	Base Gemma	10.3/29.3	9.8/27.3
	+ LoRA(250)	10.0/29.4	10.0/27.5
	+ LoRA(1012)	10.9/30.1	10.4/28.5
	Gemini 2.5 Flash	**13.8/35.4**	**15.2/34.2**

Table 7: Each cell reports *spBLEU / chrF++* on eng→{Efik, Ibibio} comparing Gemma 3 27B fine-tuning variants with Gemini 2.5 Flash, at 0 to 250 BM25-retrieved shots. **Bold** marks the variant with the highest score in that cell (spBLEU and chrF++ agree on the winner everywhere). Gemini 2.5 Flash cannot be LoRA-tuned through its API and is shown only as a closed-API reference at matched in-context counts.

Model	Direction	25+dict	25	Δ_{25}	50	Δ_{50}
GPT-4.1	eng→Efik	29.68	29.13	+0.55	29.90	−0.22
	eng→Anaang	25.63	25.58	+0.05	26.12	−0.49
Gemini 2.5 Flash	eng→Efik	33.94	34.35	−0.41	35.23	−1.29
	eng→Anaang	27.43	27.62	−0.19	28.21	−0.78

Table 8: Dictionary augmentation versus scaling in-context examples (chrF++). "25+dict" prepends a bilingual word list to a 25-shot BM25 prompt; "25" and "50" are no-dict BM25 baselines. Δ_{25} reports (25+dict) − (25): dictionary gain at matched shot count. Δ_{50} reports (25+dict) − (50): dictionary gain versus reallocating the same prompt budget to more examples. Δ_{50} is negative across all four model–language pairs: at a matched prompt budget, adding a dictionary on top of 25 shots is consistently worse than simply scaling to 50 shots.

Shots	Src	chrF++ (Bible langs)						chrF++ (no Bible)			
		efi	ibb	mfe	oro	quy	vmw	anw	lld	zgh	apd
	0-shot	31.3	25.0	55.4	13.6	33.4	26.8	15.9	38.9	2.6	43.9
1	Random	31.7	25.3	56.8	17.7	34.1	27.5	21.5	39.4	27.2	**44.8**
	BM25	**32.1**	**26.0**	**57.5**	**20.2**	**34.5**	**28.7**	**22.7**	**43.8**	**28.6**	44.6
	BM25−Random (△)	+0.3	+0.7	+0.7	+2.5	+0.3	+1.2	+1.2	+4.4	+1.4	−0.2
	Bible	**32.2**	25.6	55.9	16.2	33.6	27.7	✗	✗	✗	✗
5	Random	31.9	26.4	58.3	23.3	34.5	28.2	22.9	45.1	28.7	44.3
	BM25	**33.5**	**28.7**	**58.6**	**28.6**	**35.9**	**30.9**	**25.8**	**47.6**	**32.0**	**45.7**
	BM25−Random (△)	+1.5	+2.2	+0.3	+5.3	+1.5	+2.7	+2.9	+2.5	+3.4	+1.3
	Bible	**33.5**	27.4	56.4	18.5	33.2	28.4	✗	✗	✗	✗
10	Random	32.2	27.3	58.7	24.8	35.2	28.9	24.3	45.5	29.0	45.5
	BM25	**33.7**	**30.1**	**59.4**	**32.5**	**36.7**	**32.1**	**26.7**	**49.1**	**33.8**	**45.7**
	BM25−Random (△)	+1.5	+2.9	+0.7	+7.6	+1.5	+3.2	+2.5	+3.6	+4.7	+0.2
	Bible	**33.8**	29.0	56.9	20.7	32.9	29.0	✗	✗	✗	✗
25	Random	32.7	28.8	59.4	28.2	35.8	30.6	24.6	46.5	30.0	46.2
	BM25	**34.4**	**31.7**	**59.9**	**36.3**	**37.6**	**33.5**	**27.6**	**50.8**	**35.5**	**46.7**
	BM25−Random (△)	+1.6	+3.0	+0.6	+8.1	+1.8	+2.9	+3.1	+4.4	+5.5	+0.5
	Bible	**34.4**	30.1	57.0	22.0	32.7	29.5	✗	✗	✗	✗
50	Random	34.0	30.3	59.7	31.6	36.2	31.9	26.4	47.6	31.1	46.4
	BM25	**35.2**	**32.8**	**60.3**	**38.8**	**38.0**	**34.2**	**28.2**	**51.8**	**36.4**	**46.9**
	BM25−Random (△)	+1.2	+2.5	+0.6	+7.2	+1.8	+2.4	+1.8	+4.1	+5.3	+0.6
	Bible	34.4	31.1	57.2	22.8	32.5	29.8	✗	✗	✗	✗
100	Random	34.4	31.9	59.8	35.0	36.7	33.0	26.9	49.3	33.2	46.1
	BM25	**35.1**	**33.7**	**60.6**	**40.4**	**38.4**	**34.8**	**28.5**	**52.3**	**37.2**	**46.8**
	BM25−Random (△)	+0.7	+1.8	+0.8	+5.3	+1.7	+1.8	+1.6	+3.0	+4.0	+0.8
	Bible	**35.1**	32.0	57.2	23.7	32.6	30.1	✗	✗	✗	✗
250	Random	35.1	33.6	60.1	39.3	37.9	34.4	28.0	51.4	35.1	**46.8**
	BM25	**35.4**	**34.2**	**60.5**	**41.8**	**38.6**	**35.3**	**28.9**	**53.2**	**37.7**	**46.8**
	BM25−Random (△)	+0.4	+0.5	+0.4	+2.5	+0.7	+0.9	+0.9	+1.8	+2.6	0
	Bible	35.3	32.6	57.2	24.6	32.4	30.2	✗	✗	✗	✗
500	Random	**35.7**	**34.2**	60.4	41.7	38.5	35.2	28.8	53.2	37.0	46.6
	BM25	35.6	**34.2**	**60.8**	**42.5**	**38.8**	**35.5**	**29.2**	**53.7**	**38.0**	**47.1**
	BM25−Random (△)	−0.1	0	+0.4	+0.8	+0.3	+0.4	+0.4	+0.5	+1.1	+0.5
	Bible	35.3	32.8	57.2	25.0	32.2	30.6	✗	✗	✗	✗
1000	Random	35.6	**34.5**	60.8	42.9	**38.8**	**36.0**	29.0	53.9	**38.2**	**47.5**
	BM25	**35.9**	34.4	**61.1**	**43.2**	**38.8**	35.7	**29.5**	**54.1**	**38.3**	47.0
	BM25−Random (△)	+0.3	−0.2	+0.3	+0.3	0	−0.3	+0.4	+0.2	0	−0.5
	Bible	35.4	33.1	57.0	25.3	32.5	30.4	✗	✗	✗	✗
Gain: Random 1000 vs 0-shot		+4.4	+9.6	+5.4	+29.3	+5.4	+9.2	+13.1	+15.0	+35.6	+3.7

Table 9: eng→X results with Gemini 2.5 Flash (chrF++). Formatting conventions as described in §E.

Shots	Src	spBLEU (Bible langs)						spBLEU (no Bible)			
		efi	ibb	mfe	oro	quy	vmw	anw	lld	zgh	apd
	0-shot	9.1	6.3	32.2	2.3	9.9	6.4	3.2	17.5	0.8	25.5
1	Random	9.4	6.6	35.5	4.1	10.8	6.8	4.9	18.1	20.0	**26.6**
	BM25	**10.1**	**7.5**	**35.7**	**7.2**	**11.2**	**7.7**	**5.8**	**24.1**	**22.2**	25.9
	BM25−Random (△)	+0.7	+0.9	+0.2	+3.2	+0.4	+0.9	+1.0	+6.0	+2.2	−0.8
	Bible	9.9	7.3	33.7	3.9	10.4	7.0	✗	✗	✗	✗
5	Random	9.6	7.9	37.3	8.8	11.7	7.3	5.3	25.6	22.1	25.2
	BM25	**11.5**	**10.1**	**37.4**	**14.3**	**12.8**	**9.7**	**7.5**	**29.0**	**26.1**	**26.3**
	BM25−Random (△)	+2.0	+2.3	+0.1	+5.5	+1.1	+2.4	+2.2	+3.5	+4.0	+1.1
	Bible	11.4	9.5	34.2	4.8	10.2	7.4	✗	✗	✗	✗
10	Random	9.8	8.2	37.7	9.2	12.1	8.0	6.1	26.1	22.3	26.0
	BM25	**11.9**	**11.5**	**38.3**	**18.1**	**13.5**	**10.7**	**8.0**	**31.1**	**28.0**	**26.3**
	BM25−Random (△)	+2.1	+3.3	+0.6	+8.9	+1.4	+2.7	+2.0	+5.1	+5.7	+0.3
	Bible	11.6	11.0	34.8	5.7	9.7	7.8	✗	✗	✗	✗
25	Random	10.2	9.8	38.5	11.6	12.7	8.9	6.4	27.4	23.6	26.9
	BM25	**12.6**	**13.1**	**39.0**	**22.2**	**14.4**	**12.2**	**8.8**	**33.4**	**29.8**	**27.1**
	BM25−Random (△)	+2.3	+3.3	+0.5	+10.6	+1.7	+3.3	+2.4	+6.1	+6.2	+0.2
	Bible	12.4	12.3	35.1	6.1	9.5	8.2	✗	✗	✗	✗
50	Random	11.2	11.4	39.0	16.2	13.2	10.2	7.6	29.1	24.9	26.9
	BM25	**13.5**	**13.9**	**39.5**	**24.9**	**15.0**	**12.9**	**9.3**	**34.8**	**30.6**	**27.1**
	BM25−Random (△)	+2.3	+2.6	+0.4	+8.7	+1.8	+2.6	+1.7	+5.7	+5.7	+0.2
	Bible	12.4	12.9	35.4	6.5	9.2	8.5	✗	✗	✗	✗
100	Random	11.9	13.0	38.9	19.9	13.6	11.4	8.1	31.4	27.1	26.4
	BM25	**13.4**	**14.7**	**39.9**	**26.7**	**15.3**	**13.4**	**9.5**	**35.6**	**31.2**	**27.0**
	BM25−Random (△)	+1.5	+1.7	+0.9	+6.8	+1.7	+2.0	+1.4	+4.1	+4.2	+0.6
	Bible	13.0	13.9	35.3	6.8	9.4	8.7	✗	✗	✗	✗
250	Random	12.8	14.5	39.3	25.2	14.8	12.8	9.0	34.2	29.2	**27.3**
	BM25	**13.8**	**15.2**	**40.0**	**28.5**	**15.4**	**14.0**	**9.7**	**36.8**	**31.8**	27.1
	BM25−Random (△)	+1.0	+0.7	+0.7	+3.3	+0.6	+1.2	+0.6	+2.6	+2.7	−0.2
	Bible	13.3	14.3	35.4	7.4	9.2	8.9	✗	✗	✗	✗
500	Random	13.6	15.1	39.9	28.1	15.3	13.7	9.7	36.6	30.9	27.2
	BM25	**13.9**	**15.2**	**40.3**	**29.3**	**15.6**	**14.3**	**9.9**	**37.4**	**32.2**	**27.4**
	BM25−Random (△)	+0.3	+0.1	+0.4	+1.3	+0.3	+0.6	+0.2	+0.8	+1.3	+0.2
	Bible	13.1	14.7	35.4	7.7	9.0	9.1	✗	✗	✗	✗
1000	Random	13.5	**15.6**	40.7	30.0	15.6	**14.8**	9.9	37.7	32.2	**28.0**
	BM25	**14.1**	15.4	**40.9**	**30.3**	**15.8**	14.7	**10.2**	**38.0**	**32.6**	27.5
	BM25−Random (△)	+0.6	−0.3	+0.1	+0.4	+0.1	−0.1	+0.3	+0.3	+0.3	−0.5
	Bible	13.2	15.1	35.3	7.9	9.2	9.0	✗	✗	✗	✗
Gain: Random 1000 vs 0-shot		+4.4	+9.4	+8.5	+27.7	+5.7	+8.4	+6.6	+20.2	+31.4	+2.5

Table 10: eng→X results with Gemini 2.5 Flash (spBLEU). Formatting conventions as described in §E.

Shots	Src	chrF++ (Bible langs)						chrF++ (no Bible)			
		efi	ibb	mfe	oro	quy	vmw	anw	lld	zgh	apd
	0-shot	23.6	22.7	54.6	14.5	29.9	21.2	21.3	31.9	2.1	44.2
1	Random	23.7	23.3	55.1	19.1	30.5	21.7	21.6	32.2	16.3	**43.9**
	BM25	**24.6**	**24.0**	**55.4**	**21.7**	**30.9**	**23.5**	**22.3**	**33.8**	**18.1**	**43.9**
	BM25−Random (△)	+0.9	+0.7	+0.3	+2.6	+0.4	+1.7	+0.7	+1.6	+1.8	0
	Bible	24.5	23.5	53.8	17.1	30.1	22.1	✗	✗	✗	✗
5	Random	24.6	23.8	56.1	21.6	31.0	23.2	22.7	33.0	17.9	43.6
	BM25	**26.6**	**25.5**	**57.0**	**26.2**	**32.5**	**26.5**	**23.8**	**37.6**	**21.1**	**44.6**
	BM25−Random (△)	+2.0	+1.7	+0.9	+4.6	+1.5	+3.3	+1.0	+4.6	+3.2	+1.0
	Bible	26.5	24.8	53.8	19.5	30.6	23.7	✗	✗	✗	✗
10	Random	25.1	23.9	56.9	22.6	31.7	23.6	22.9	33.5	18.5	**45.5**
	BM25	**27.6**	**26.5**	**57.3**	**28.6**	**33.5**	**27.9**	**24.7**	**39.9**	**22.7**	44.9
	BM25−Random (△)	+2.5	+2.7	+0.4	+5.9	+1.8	+4.3	+1.8	+6.4	+4.3	−0.6
	Bible	27.4	25.5	54.3	20.5	30.8	24.4	✗	✗	✗	✗
25	Random	25.9	24.4	57.2	24.2	32.1	25.3	22.6	35.3	18.8	**45.7**
	BM25	**29.1**	**28.2**	**58.3**	**31.3**	**34.6**	**29.5**	**25.6**	**43.1**	**23.9**	45.4
	BM25−Random (△)	+3.2	+3.7	+1.1	+7.1	+2.4	+4.3	+3.0	+7.8	+5.1	−0.3
	Bible	28.3	26.8	54.0	21.4	30.9	25.7	✗	✗	✗	✗
50	Random	27.3	25.3	57.3	26.0	32.7	26.4	23.9	37.4	19.4	**46.3**
	BM25	**29.9**	**28.9**	**58.5**	**33.0**	**35.2**	**30.7**	**26.1**	**44.8**	**24.2**	45.8
	BM25−Random (△)	+2.6	+3.6	+1.2	+7.1	+2.4	+4.2	+2.2	+7.4	+4.8	−0.5
	Bible	29.2	27.5	54.3	22.0	31.1	26.6	✗	✗	✗	✗
100	Random	27.9	27.0	57.6	28.4	33.6	28.3	24.3	40.4	20.5	**46.3**
	BM25	**30.4**	**29.6**	**58.6**	**34.1**	**35.4**	**31.2**	**26.4**	**46.3**	**24.4**	46.3
	BM25−Random (△)	+2.5	+2.7	+1.0	+5.6	+1.8	+2.9	+2.1	+5.9	+3.9	−0.1
	Bible	29.8	28.3	54.6	22.5	31.2	27.1	✗	✗	✗	✗
250	Random	29.4	28.6	57.6	31.4	34.6	29.7	25.2	43.6	21.3	**46.3**
	BM25	**30.8**	**30.0**	**58.6**	**34.6**	**35.8**	**31.6**	**26.5**	**47.5**	**24.0**	46.2
	BM25−Random (△)	+1.4	+1.4	+0.9	+3.3	+1.1	+1.9	+1.3	+3.9	+2.7	−0.1
	Bible	30.6	29.2	54.7	22.7	31.2	27.8	✗	✗	✗	✗
500	Random	30.4	29.3	58.1	33.5	35.1	30.9	25.9	46.2	–	**46.5**
	BM25	**31.1**	**30.1**	**58.2**	**34.9**	**35.9**	**31.7**	**26.7**	**48.1**	–	46.4
	BM25−Random (△)	+0.7	+0.7	+0.1	+1.4	+0.8	+0.8	+0.8	+1.9	–	−0.1
	Bible	30.8	29.5	54.8	23.3	31.3	28.0	✗	✗	✗	✗
1000	Random	31.1	**30.0**	**59.4**	34.8	35.6	**32.6**	26.5	**49.1**	–	**48.9**
	BM25	**31.4**	**30.0**	58.5	**34.9**	**35.7**	31.9	**26.7**	48.4	–	46.3
	BM25−Random (△)	+0.3	0	−0.9	+0.1	+0.1	−0.7	+0.1	−0.8	–	−2.6
	Bible	30.9	29.7	54.6	23.1	31.4	28.1	✗	✗	✗	✗
Gain: Random 1000 vs 0-shot		+7.5	+7.4	+4.8	+20.2	+5.7	+11.3	+5.2	+17.3	–	+4.7

Table 11: eng→X results with GPT-4.1 (chrF++). Formatting conventions as described in §E.

Shots	Src	spBLEU (Bible langs)						spBLEU (no Bible)			
		efi	ibb	mfe	oro	quy	vmw	anw	lld	zgh	apd
	0-shot	6.9	5.8	32.9	2.7	8.0	3.8	5.9	9.8	0.3	25.2
1	Random	6.0	6.6	34.5	5.1	8.1	4.2	5.5	10.3	7.2	**24.8**
	BM25	**6.9**	**7.1**	**34.6**	**8.1**	**8.5**	**5.1**	**5.8**	**12.3**	**10.0**	24.3
	BM25−Random (△)	+0.9	+0.5	+0.1	+3.1	+0.4	+0.9	+0.3	+2.0	+2.8	−0.5
	Bible	**6.9**	6.8	32.3	3.9	7.9	4.4	✗	✗	✗	✗
5	Random	6.4	6.7	35.6	6.8	8.5	4.5	5.9	11.2	9.4	24.3
	BM25	**7.9**	**8.1**	**36.5**	**11.5**	**9.7**	**6.7**	**6.3**	**16.4**	**13.4**	**24.8**
	BM25−Random (△)	+1.5	+1.4	+0.9	+4.7	+1.2	+2.2	+0.4	+5.2	+4.0	+0.4
	Bible	7.7	7.7	32.4	5.0	8.0	4.8	✗	✗	✗	✗
10	Random	6.2	6.5	36.4	7.1	8.9	4.4	5.6	11.6	9.6	**26.1**
	BM25	**8.5**	**8.9**	**36.7**	**13.9**	**10.9**	**7.5**	**6.9**	**19.0**	**15.1**	25.0
	BM25−Random (△)	+2.3	+2.4	+0.4	+6.8	+1.9	+3.1	+1.3	+7.4	+5.5	−1.1
	Bible	8.0	8.2	32.7	5.4	8.2	5.0	✗	✗	✗	✗
25	Random	6.9	6.9	36.6	8.5	9.6	5.1	5.6	13.3	10.0	**26.1**
	BM25	**9.7**	**10.2**	**38.0**	**16.5**	**11.8**	**9.0**	**7.4**	**22.6**	**16.6**	25.5
	BM25−Random (△)	+2.7	+3.3	+1.4	+8.0	+2.2	+3.9	+1.8	+9.3	+6.6	−0.6
	Bible	8.4	9.2	32.1	5.5	8.0	5.6	✗	✗	✗	✗
50	Random	7.6	7.4	36.9	10.4	10.2	6.0	6.1	15.6	10.9	**26.9**
	BM25	**10.2**	**11.0**	**38.2**	**18.5**	**12.5**	**9.8**	**7.9**	**24.7**	**16.9**	26.0
	BM25−Random (△)	+2.6	+3.6	+1.3	+8.1	+2.3	+3.9	+1.8	+9.1	+6.0	−0.9
	Bible	8.8	10.0	32.5	5.8	8.1	6.0	✗	✗	✗	✗
100	Random	8.4	8.7	37.2	12.8	11.0	7.5	6.5	19.2	12.2	**26.6**
	BM25	**10.7**	**11.6**	**38.3**	**19.5**	**12.7**	**10.3**	**8.2**	**26.6**	**17.3**	26.5
	BM25−Random (△)	+2.2	+2.9	+1.1	+6.8	+1.7	+2.8	+1.7	+7.4	+5.1	−0.1
	Bible	9.1	10.5	32.9	5.9	8.1	6.4	✗	✗	✗	✗
250	Random	9.4	10.3	37.2	16.5	11.8	8.8	7.2	23.4	13.3	**26.6**
	BM25	**10.8**	**12.0**	**38.3**	**20.2**	**13.1**	**10.7**	**8.2**	**28.5**	**16.8**	26.4
	BM25−Random (△)	+1.4	+1.6	+1.1	+3.7	+1.3	+2.0	+1.0	+5.1	+3.5	−0.2
	Bible	9.4	11.3	32.8	6.0	7.9	6.7	✗	✗	✗	✗
500	Random	10.4	11.1	37.9	18.9	12.5	10.0	7.7	26.5	–	**26.7**
	BM25	**11.1**	**12.0**	**38.0**	**20.5**	**13.2**	**10.8**	**8.4**	**29.1**	–	**26.7**
	BM25−Random (△)	+0.7	+0.9	+0.1	+1.6	+0.7	+0.7	+0.7	+2.5	–	0
	Bible	9.7	11.4	33.0	6.4	8.0	7.0	✗	✗	✗	✗
1000	Random	11.2	**11.8**	**38.7**	**20.6**	**13.1**	**11.8**	8.3	**30.1**	–	**29.5**
	BM25	**11.4**	**11.8**	38.3	20.5	**13.0**	10.9	**8.4**	29.6	–	26.6
	BM25−Random (△)	+0.2	0	−0.4	−0.1	0	−0.9	+0.2	−0.5	–	−2.9
	Bible	9.8	**11.7**	32.6	6.3	8.0	7.0	✗	✗	✗	✗
Gain: Random 1000 vs 0-shot		+4.3	+6.0	+5.7	+17.9	+5.1	+8.0	+2.4	+20.3	–	+4.3

Table 12: eng→X results with GPT-4.1 (spBLEU). Formatting conventions as described in §E.

Shots	Src	chrF++ (Bible langs)						chrF++ (no Bible)			
		efi	ibb	mfe	oro	quy	vmw	anw	lld	zgh	apd
	0-shot	19.9	18.9	36.5	14.4	27.0	18.7	16.8	31.8	2.0	40.7
1	Random	21.8	19.8	36.6	17.3	27.1	19.0	18.2	32.3	17.6	39.7
	BM25	**22.8**	**21.6**	**39.3**	**19.8**	**28.1**	**21.2**	**19.8**	**33.7**	**19.6**	**40.2**
	BM25−Random (△)	+1.0	+1.8	+2.7	+2.5	+1.1	+2.2	+1.6	+1.4	+1.9	+0.5
	Bible	22.4	20.1	37.3	16.1	26.9	19.6	✗	✗	✗	✗
5	Random	23.5	22.4	39.7	21.6	28.0	21.5	21.8	33.4	18.5	38.9
	BM25	**25.5**	**24.7**	**43.6**	**26.2**	**30.2**	**24.8**	**23.5**	**37.5**	**21.4**	**40.3**
	BM25−Random (△)	+2.0	+2.3	+4.0	+4.6	+2.1	+3.4	+1.7	+4.1	+3.0	+1.4
	Bible	24.7	22.2	39.6	19.3	27.8	21.7	✗	✗	✗	✗
10	Random	22.6	23.4	40.7	22.4	28.5	22.5	22.4	34.1	18.8	39.5
	BM25	**24.4**	**25.8**	**45.2**	**27.8**	**30.7**	**25.6**	**24.2**	**37.5**	**22.4**	**40.2**
	BM25−Random (△)	+1.7	+2.5	+4.6	+5.3	+2.2	+3.1	+1.8	+3.4	+3.5	+0.7
	Bible	**25.5**	22.0	40.4	20.2	28.2	22.5	✗	✗	✗	✗
25	Random	24.6	24.1	42.2	23.5	29.3	23.6	22.8	35.2	19.1	40.2
	BM25	**26.0**	**26.7**	**47.7**	**29.8**	**30.8**	**27.2**	**24.6**	**39.6**	**22.6**	**40.3**
	BM25−Random (△)	+1.3	+2.6	+5.4	+6.3	+1.5	+3.6	+1.8	+4.4	+3.5	+0.1
	Bible	25.2	24.4	41.3	20.6	28.7	23.3	✗	✗	✗	✗
50	Random	**26.2**	24.4	43.5	24.6	29.6	24.2	22.3	36.4	19.5	40.5
	BM25	25.0	**27.6**	**49.3**	**30.0**	**31.3**	**26.5**	**24.1**	**41.0**	**22.8**	**40.8**
	BM25−Random (△)	−1.1	+3.2	+5.8	+5.5	+1.7	+2.3	+1.8	+4.6	+3.4	+0.3
	Bible	23.6	24.6	41.3	21.0	28.8	23.6	✗	✗	✗	✗
100	Random	**27.0**	25.1	44.2	25.6	30.7	25.2	23.6	37.9	19.9	40.9
	BM25	25.5	**27.3**	**49.8**	**31.1**	**32.0**	**27.6**	**25.1**	**41.5**	**20.2**	**41.3**
	BM25−Random (△)	−1.4	+2.3	+5.5	+5.5	+1.3	+2.4	+1.5	+3.5	+0.3	+0.5
	Bible	**27.7**	23.9	40.7	20.8	28.9	23.6	✗	✗	✗	✗
250	Random	28.0	26.5	46.7	27.9	30.6	26.3	23.7	40.5	–	40.9
	BM25	**28.4**	**27.7**	**50.3**	**30.6**	**32.4**	**27.7**	**24.9**	**42.8**	**10.4**	**41.3**
	BM25−Random (△)	+0.4	+1.3	+3.6	+2.7	+1.8	+1.4	+1.2	+2.3	–	+0.3
	Bible	26.6	24.0	43.7	20.6	28.8	23.3	✗	✗	✗	✗
500	Random	**29.1**	27.2	47.8	28.8	29.3	27.3	24.5	**42.9**	–	41.4
	BM25	27.5	**27.5**	**49.7**	**30.3**	**31.2**	**27.8**	**25.0**	42.7	–	**41.5**
	BM25−Random (△)	−1.6	+0.3	+2.0	+1.5	+1.8	+0.5	+0.4	−0.2	–	+0.1
	Bible	26.8	25.2	40.4	19.8	28.6	22.4	✗	✗	✗	✗
1000	Random	**28.2**	**24.8**	–	**25.6**	**30.6**	–	**24.9**	**41.4**	–	–
	BM25	23.8	20.4	**48.4**	24.9	29.2	**22.9**	23.3	38.4	–	**40.7**
	BM25−Random (△)	−4.4	−4.4	–	−0.7	−1.4	–	−1.6	−3.0	–	–
	Bible	19.2	17.4	38.4	–	–	–	✗	✗	✗	✗
Gain: Random 1000 vs 0-shot		+8.4	+6.0	–	+11.2	+3.7	–	+8.1	+9.6	–	–

Table 13: eng→X results with Llama 3.3 70B (chrF++). Formatting conventions as described in §E.

Shots	Src	spBLEU (Bible langs)						spBLEU (no Bible)			
		efi	ibb	mfe	oro	quy	vmw	anw	lld	zgh	apd
	0-shot	4.0	4.1	12.9	2.4	6.0	2.5	3.5	9.6	0.3	20.2
1	Random	4.4	3.8	13.0	4.4	6.1	2.5	3.5	10.2	7.2	19.3
	BM25	**5.0**	**5.0**	**15.7**	**6.9**	**6.5**	**3.5**	**4.4**	**12.2**	**10.2**	**19.6**
	BM25−Random (△)	+0.6	+1.2	+2.7	+2.4	+0.4	+1.0	+0.9	+2.0	+3.0	+0.3
	Bible	4.7	4.6	13.8	3.0	5.5	2.7	✗	✗	✗	✗
5	Random	5.1	5.0	15.6	6.3	6.4	3.0	4.3	11.9	8.8	18.1
	BM25	**6.3**	**5.8**	**19.8**	**10.5**	**7.6**	**4.3**	**5.7**	**16.8**	**12.8**	**19.3**
	BM25−Random (△)	+1.2	+0.8	+4.2	+4.2	+1.2	+1.3	+1.4	+4.9	+4.0	+1.2
	Bible	5.6	5.5	15.8	3.7	5.8	3.1	✗	✗	✗	✗
10	Random	**5.3**	5.2	16.4	7.3	6.7	3.4	4.4	12.5	9.9	18.4
	BM25	5.0	**6.3**	**21.6**	**12.1**	**7.9**	**4.9**	**6.2**	**18.0**	**14.3**	**18.7**
	BM25−Random (△)	−0.3	+1.1	+5.2	+4.8	+1.2	+1.4	+1.8	+5.4	+4.5	+0.3
	Bible	**6.1**	5.7	16.7	3.9	5.8	3.3	✗	✗	✗	✗
25	Random	5.6	5.7	18.0	7.4	7.1	3.6	4.9	14.1	10.5	**19.2**
	BM25	**6.3**	**6.9**	**24.0**	**14.1**	**7.6**	**5.8**	**6.6**	**18.2**	**15.1**	19.0
	BM25−Random (△)	+0.7	+1.2	+6.0	+6.7	+0.5	+2.2	+1.7	+4.1	+4.6	−0.3
	Bible	5.9	6.7	17.1	4.2	6.0	3.8	✗	✗	✗	✗
50	Random	**6.2**	5.6	19.5	8.9	7.5	4.3	5.2	15.5	11.0	19.2
	BM25	5.2	**7.8**	**26.5**	**15.0**	**8.7**	**6.1**	**6.5**	**21.6**	**15.9**	**19.5**
	BM25−Random (△)	−1.0	+2.2	+7.0	+6.1	+1.1	+1.7	+1.2	+6.1	+4.9	+0.3
	Bible	5.3	7.1	16.9	4.5	6.3	4.1	✗	✗	✗	✗
100	Random	**7.0**	6.9	20.3	10.4	8.3	5.0	5.8	17.7	11.9	19.5
	BM25	5.5	**8.0**	**26.9**	**16.6**	**9.2**	**6.7**	**7.0**	**21.1**	**14.0**	**20.2**
	BM25−Random (△)	−1.5	+1.1	+6.6	+6.2	+0.8	+1.7	+1.2	+3.4	+2.0	+0.7
	Bible	**7.3**	7.3	18.0	4.6	6.4	4.2	✗	✗	✗	✗
250	Random	**7.9**	8.1	23.2	13.4	9.0	5.7	6.1	20.6	–	19.9
	BM25	7.8	**8.7**	**27.5**	**16.3**	**9.3**	**6.9**	**7.1**	**22.7**	**7.5**	**20.0**
	BM25−Random (△)	−0.1	+0.7	+4.4	+2.9	+0.3	+1.3	+1.1	+2.1	–	+0.1
	Bible	6.8	8.0	20.1	4.8	6.9	4.2	✗	✗	✗	✗
500	Random	**8.7**	8.5	24.6	14.8	**9.2**	6.8	6.5	**23.3**	–	20.2
	BM25	7.1	**8.7**	**27.1**	**16.5**	8.5	**7.3**	**7.1**	22.8	–	**20.4**
	BM25−Random (△)	−1.5	+0.1	+2.5	+1.7	−0.7	+0.5	+0.6	−0.5	–	+0.3
	Bible	7.1	8.0	16.6	4.6	6.7	3.9	✗	✗	✗	✗
1000	Random	**8.7**	**7.6**	–	**13.1**	**9.9**	–	**6.8**	**23.6**	–	–
	BM25	5.9	4.0	**25.4**	12.2	9.3	**5.3**	6.6	21.1	–	**19.8**
	BM25−Random (△)	−2.8	−3.6	–	−1.0	−0.6	–	−0.3	−2.5	–	–
	Bible	4.6	4.4	16.0	–	–	–	✗	✗	✗	✗
Gain: Random 1000 vs 0-shot		+4.7	+3.5	–	+10.7	+3.9	–	+3.3	+14.0	–	–

Table 14: eng→X results with Llama 3.3 70B (spBLEU). Formatting conventions as described in §E.

Shots	Src	chrF++ (Bible langs)						chrF++ (no Bible)			
		efi	ibb	mfe	oro	quy	vmw	anw	lld	zgh	apd
	0-shot	19.8	18.5	41.2	11.8	26.1	18.7	11.7	30.5	2.9	42.6
1	Random	21.2	18.9	43.5	18.7	25.4	19.2	17.7	31.3	18.0	43.5
	BM25	**22.1**	**20.2**	**45.7**	**19.7**	**26.5**	**21.4**	**19.8**	**33.1**	**19.5**	**43.7**
	BM25−Random (△)	+0.9	+1.3	+2.2	+1.0	+1.1	+2.2	+2.1	+1.8	+1.5	+0.2
	Bible	21.4	18.6	43.0	15.4	25.8	19.7	✗	✗	✗	✗
5	Random	23.2	21.1	47.0	22.0	26.2	23.0	22.5	33.2	19.4	44.1
	BM25	**25.5**	**23.2**	**49.8**	**25.6**	**28.4**	**25.4**	**23.3**	**37.8**	**22.9**	**44.6**
	BM25−Random (△)	+2.3	+2.1	+2.7	+3.7	+2.2	+2.4	+0.8	+4.6	+3.5	+0.4
	Bible	24.1	20.3	44.6	17.9	26.3	22.3	✗	✗	✗	✗
10	Random	24.2	22.1	48.2	23.6	26.7	23.6	23.3	33.9	20.3	44.5
	BM25	**27.0**	**24.7**	**51.8**	**28.2**	**29.2**	**27.1**	**24.1**	**40.3**	**24.8**	**44.7**
	BM25−Random (△)	+2.8	+2.6	+3.6	+4.6	+2.6	+3.6	+0.7	+6.4	+4.5	+0.2
	Bible	25.0	21.7	45.2	18.7	26.6	23.4	✗	✗	✗	✗
25	Random	24.8	23.2	49.2	26.7	27.9	24.2	23.6	36.0	20.7	44.8
	BM25	**28.4**	**26.5**	**53.3**	**31.3**	**30.3**	**28.4**	**25.0**	**43.0**	**26.6**	**45.2**
	BM25−Random (△)	+3.7	+3.3	+4.1	+4.7	+2.4	+4.2	+1.3	+7.0	+5.9	+0.4
	Bible	26.4	22.9	46.0	19.8	26.9	24.2	✗	✗	✗	✗
50	Random	26.5	24.4	49.8	27.0	28.1	25.3	24.2	37.4	21.6	45.3
	BM25	**29.2**	**27.1**	**53.8**	**32.1**	**30.7**	**28.9**	**25.6**	**44.2**	**27.0**	**45.4**
	BM25−Random (△)	+2.7	+2.7	+4.0	+5.1	+2.6	+3.6	+1.5	+6.8	+5.4	+0.2
	Bible	27.0	23.5	46.1	20.1	27.1	24.5	✗	✗	✗	✗
100	Random	26.8	25.7	50.9	28.1	28.6	26.0	24.9	39.7	22.6	45.3
	BM25	**29.5**	**27.6**	**53.7**	**33.1**	**30.7**	**29.0**	**25.9**	**45.2**	**26.9**	**45.8**
	BM25−Random (△)	+2.7	+1.8	+2.8	+5.0	+2.1	+3.0	+1.0	+5.5	+4.3	+0.4
	Bible	27.1	23.9	46.2	19.7	27.4	24.7	✗	✗	✗	✗
250	Random	28.1	27.2	50.9	29.6	30.0	26.8	25.7	41.4	23.5	**45.6**
	BM25	**29.3**	**27.3**	**52.9**	**32.5**	**30.7**	**28.4**	**26.3**	**44.6**	**26.0**	**45.6**
	BM25−Random (△)	+1.3	+0.1	+2.0	+3.0	+0.7	+1.6	+0.5	+3.2	+2.5	0
	Bible	27.4	24.4	46.2	20.0	27.9	24.9	✗	✗	✗	✗
500	Random	28.4	27.3	51.5	30.8	29.7	26.6	**26.1**	42.4	–	45.6
	BM25	**28.9**	**27.6**	**52.0**	**31.4**	**30.5**	**27.5**	26.1	**43.7**	–	**45.7**
	BM25−Random (△)	+0.5	+0.3	+0.5	+0.6	+0.8	+0.9	−0.1	+1.3	–	+0.1
	Bible	26.8	24.1	47.3	20.4	27.7	24.4	✗	✗	✗	✗
1000	Random	28.6	**28.5**	–	**28.9**	**30.0**	–	**26.1**	–	–	–
	BM25	**28.7**	–	–	–	–	–	26.0	–	–	–
	BM25−Random (△)	+0.1	–	–	–	–	–	−0.1	–	–	–
	Bible	26.7	24.2	**45.2**	22.9	27.0	**23.5**	✗	✗	✗	✗
Gain: Random 1000 vs 0-shot		+8.9	+9.9	–	+17.1	+3.9	–	+14.4	–	–	–

Table 15: eng→X results with Gemma 3 27B (chrF++). Formatting conventions as described in §E.

Shots	Src	spBLEU (Bible langs)						spBLEU (no Bible)			
		efi	ibb	mfe	oro	quy	vmw	anw	lld	zgh	apd
	0-shot	3.6	3.8	17.1	2.5	5.0	3.0	2.0	8.6	0.8	23.5
1	Random	3.8	3.7	19.9	5.1	4.9	3.0	3.7	9.1	9.0	**24.1**
	BM25	**4.8**	**4.9**	**22.2**	**7.4**	**5.5**	**4.1**	**4.8**	**11.8**	**11.7**	**24.1**
	BM25−Random (△)	+0.9	+1.2	+2.4	+2.3	+0.7	+1.0	+1.1	+2.7	+2.8	0
	Bible	4.2	4.1	19.5	3.3	5.2	3.1	✗	✗	✗	✗
5	Random	4.9	5.1	23.1	7.3	5.5	4.4	5.8	11.4	11.3	24.4
	BM25	**6.8**	**6.7**	**26.8**	**11.4**	**7.1**	**5.9**	**6.1**	**17.1**	**15.9**	**24.6**
	BM25−Random (△)	+1.9	+1.7	+3.6	+4.1	+1.6	+1.5	+0.3	+5.7	+4.5	+0.2
	Bible	5.6	5.4	21.2	4.3	5.4	3.5	✗	✗	✗	✗
10	Random	5.3	5.7	24.6	8.5	6.0	4.5	6.1	12.2	12.7	24.6
	BM25	**7.7**	**7.7**	**29.3**	**14.0**	**8.1**	**7.1**	**6.6**	**20.2**	**18.2**	**24.9**
	BM25−Random (△)	+2.4	+2.0	+4.7	+5.5	+2.1	+2.6	+0.6	+8.0	+5.5	+0.2
	Bible	6.0	6.3	21.7	4.6	5.7	3.6	✗	✗	✗	✗
25	Random	6.3	6.4	26.1	10.4	7.1	4.6	6.2	14.5	13.5	24.8
	BM25	**8.9**	**9.1**	**31.5**	**16.6**	**9.1**	**8.5**	**7.0**	**23.6**	**20.5**	**25.0**
	BM25−Random (△)	+2.7	+2.7	+5.4	+6.1	+2.0	+3.9	+0.8	+9.2	+7.0	+0.3
	Bible	7.1	7.2	22.1	5.2	5.7	4.1	✗	✗	✗	✗
50	Random	7.1	7.4	26.9	10.7	7.4	5.9	6.3	16.2	14.8	**25.1**
	BM25	**9.7**	**9.7**	**32.2**	**17.3**	**9.6**	**9.0**	**7.5**	**24.9**	**20.9**	**25.1**
	BM25−Random (△)	+2.6	+2.3	+5.3	+6.6	+2.2	+3.1	+1.1	+8.7	+6.1	0
	Bible	7.6	7.8	22.0	5.4	6.3	4.5	✗	✗	✗	✗
100	Random	7.9	8.3	28.2	11.8	8.0	6.9	6.9	19.2	16.0	25.0
	BM25	**10.3**	**9.7**	**32.0**	**17.7**	**9.7**	**9.3**	**7.9**	**26.1**	**20.9**	**25.4**
	BM25−Random (△)	+2.4	+1.5	+3.9	+5.9	+1.7	+2.4	+1.0	+6.9	+4.8	+0.3
	Bible	7.8	8.2	22.2	5.1	6.3	4.8	✗	✗	✗	✗
250	Random	9.1	9.2	28.6	13.2	9.0	7.6	7.5	21.2	17.0	**25.4**
	BM25	**10.3**	**9.8**	**31.1**	**16.3**	**9.7**	**9.1**	**8.2**	**25.6**	**19.7**	**25.5**
	BM25−Random (△)	+1.3	+0.5	+2.5	+3.1	+0.7	+1.4	+0.7	+4.4	+2.7	0
	Bible	8.2	8.6	22.8	5.8	6.8	5.0	✗	✗	✗	✗
500	Random	9.4	9.5	29.4	13.6	8.7	7.4	8.0	22.5	–	**25.6**
	BM25	**10.1**	**10.1**	**30.0**	**14.6**	**9.5**	**8.4**	**8.1**	**24.4**	–	25.6
	BM25−Random (△)	+0.7	+0.6	+0.6	+1.0	+0.8	+0.9	+0.1	+1.9	–	−0.1
	Bible	7.9	8.2	23.8	6.2	6.5	4.9	✗	✗	✗	✗
1000	Random	**10.3**	**10.4**	–	**12.1**	**9.2**	–	**8.0**	–	–	–
	BM25	9.4	–	–	–	–	–	7.9	–	–	–
	BM25−Random (△)	−0.9	–	–	–	–	–	−0.1	–	–	–
	Bible	7.7	8.4	**22.2**	7.7	6.3	**4.6**	✗	✗	✗	✗
Gain: Random 1000 vs 0-shot		+6.7	+6.6	–	+9.7	+4.2	–	+6.0	–	–	–

Table 16: eng→X results with Gemma 3 27B (spBLEU). Formatting conventions as described in §E.

Shots	vmw	zgh	lld	mfe	apd	anw	efi	ibb	oro	quy
Gemini 2.5 Flash										
0	43.5	44.0	62.2	70.8	63.6	32.3	34.5	36.5	27.8	40.6
1	44.5	44.4	63.3	71.2	64.2	34.1	35.6	38.1	28.5	41.8
5	45.6	46.1	63.7	71.6	64.8	34.7	36.4	39.0	29.0	42.6
10	46.3	46.3	63.9	71.6	65.0	34.9	36.5	39.1	30.6	43.0
25	46.3	47.1	64.3	71.9	65.2	35.6	36.7	40.0	31.5	43.1
50	45.8	47.4	**64.7**	71.8	**65.7**	36.4	37.2	40.4	33.4	43.6
100	47.1	48.2	64.4	71.8	65.5	36.7	37.4	41.1	35.0	43.9
250	47.7	**49.0**	64.5	**72.0**	65.5	37.6	37.3	41.5	36.8	43.8
500	**47.8**	48.9	64.1	71.7	65.3	38.1	37.4	41.8	37.7	**43.9**
1,000	47.4	48.8	64.0	71.5	65.2	**38.4**	**37.7**	**42.0**	**39.0**	43.7
gain	×1.1	×1.1	×1.0	×1.0	×1.0	×1.2	×1.1	×1.1	×1.4	×1.1
GPT-4.1										
0	30.3	19.6	55.5	70.4	63.1	27.7	24.5	27.9	26.5	34.6
1	30.2	19.4	57.2	71.1	64.2	28.4	24.8	28.3	27.3	36.0
5	31.6	20.0	58.4	71.7	65.3	29.3	25.6	29.4	28.3	37.1
10	31.5	20.3	58.9	72.0	65.4	29.4	25.7	29.6	28.7	37.3
25	32.2	20.3	59.8	72.4	66.0	29.7	26.0	29.9	28.7	37.8
50	32.7	20.7	60.2	72.2	66.2	29.7	26.0	30.1	29.3	38.0
100	33.5	20.7	60.9	72.5	66.4	30.1	26.4	30.5	29.7	38.6
250	33.6	**21.1**	61.4	72.7	66.6	30.2	26.9	30.6	30.5	38.8
500	34.2	–	61.8	72.9	66.9	**30.5**	26.8	31.1	31.0	**39.2**
1,000	**38.2**	–	**63.9**	**74.2**	**68.1**	**30.5**	**27.2**	**31.3**	**31.6**	39.1
gain	×1.3	×1.1	×1.2	×1.1	×1.1	×1.1	×1.1	×1.1	×1.2	×1.1
Llama 3.3 70B										
0	26.6	19.8	50.9	58.5	59.8	27.1	24.1	27.1	26.4	31.6
1	26.7	20.4	54.1	60.9	60.9	27.1	24.8	27.0	26.9	32.6
5	27.4	20.4	55.3	62.3	61.6	27.3	24.6	27.6	26.8	33.1
10	27.7	20.6	55.7	62.7	61.7	27.7	24.5	27.2	26.9	33.1
25	28.0	**20.9**	56.1	62.9	**61.8**	27.9	25.0	27.8	26.8	32.4
50	28.3	20.6	56.7	63.3	61.7	28.0	25.4	27.1	27.4	32.9
100	**28.6**	**20.8**	57.4	63.7	61.6	**28.2**	25.6	27.6	27.9	33.0
250	28.2	–	58.0	64.0	61.5	28.2	**25.8**	**28.1**	**28.2**	32.7
500	28.3	–	**58.3**	**64.5**	61.5	**28.3**	25.7	**28.1**	27.9	**33.2**
1,000	–	–	–	–	–	27.3	24.1	24.0	24.8	28.5
gain	×1.1	×1.1	×1.1	×1.1	×1.0	×1.0	×1.1	×1.0	×1.1	×1.1
Gemma 3 27B										
0	27.5	23.1	51.5	64.9	61.2	26.6	23.4	26.6	25.4	29.8
1	27.7	22.5	55.0	66.1	61.6	27.1	24.1	26.7	25.8	30.8
5	28.4	23.5	56.4	67.2	62.2	27.5	24.7	27.2	26.4	31.4
10	28.6	23.7	56.5	67.2	62.4	27.8	24.7	27.3	26.7	31.5
25	28.9	24.0	57.2	67.3	62.3	28.0	25.2	27.6	27.2	32.2
50	29.4	24.1	57.6	67.8	62.7	27.9	25.2	27.8	27.3	32.3
100	30.0	24.3	58.5	67.9	**63.1**	28.4	25.5	28.1	27.9	32.5
250	30.2	**24.6**	58.9	**68.2**	62.9	28.4	**26.0**	**28.3**	**28.3**	**32.7**
500	**30.3**	–	**59.0**	**68.1**	62.7	**28.6**	25.9	**28.3**	28.1	32.6
1,000	–	–	–	–	–	28.1	25.8	28.0	27.7	32.1
gain	×1.1	×1.1	×1.1	×1.0	×1.0	×1.1	×1.1	×1.1	×1.1	×1.1

Table 17: X→eng scaling with random selection (chrF++). Formatting conventions as described in §E.

Shots	vmw	zgh	lld	mfe	apd	anw	efi	ibb	oro	quy
Gemini 2.5 Flash										
0	23.6	24.3	41.4	54.4	42.4	14.7	17.4	17.4	10.0	20.6
1	24.8	25.0	42.9	54.9	43.1	16.1	18.2	19.5	12.3	22.0
5	26.4	26.8	43.6	55.5	44.8	17.7	18.9	21.4	13.7	23.5
10	27.1	27.0	44.1	55.5	44.7	17.7	19.1	21.4	14.6	23.9
25	27.5	27.8	44.8	**56.0**	45.2	18.6	19.1	22.2	15.3	24.4
50	27.3	28.0	**45.3**	55.8	**45.9**	18.9	19.7	22.6	16.4	25.0
100	28.6	29.1	45.1	55.8	45.7	19.2	19.7	23.1	17.5	25.2
250	28.8	**29.9**	44.9	**56.1**	45.7	19.9	19.6	23.4	19.1	25.0
500	**28.9**	29.3	44.5	55.8	45.7	20.5	19.6	23.9	19.6	**25.4**
1,000	28.5	29.1	44.2	55.5	45.6	**20.5**	**20.1**	**24.0**	**20.7**	24.9
gain	×1.2	×1.2	×1.1	×1.0	×1.1	×1.4	×1.2	×1.4	×**2.1**	×1.2
GPT-4.1										
0	11.7	2.8	33.7	53.3	39.9	11.5	10.4	11.2	8.9	13.8
1	13.2	2.8	36.7	54.7	42.5	12.9	11.0	13.0	11.7	16.8
5	14.6	3.4	38.1	55.4	44.3	13.8	11.3	13.6	12.1	18.4
10	14.8	3.7	39.2	55.9	44.5	14.0	11.6	14.1	12.6	19.1
25	15.4	3.5	40.4	56.6	45.6	14.3	11.7	14.6	12.7	19.9
50	15.7	**3.9**	40.7	56.2	46.0	14.4	11.9	14.6	13.0	20.2
100	16.7	3.6	41.5	56.8	46.4	14.4	12.0	15.1	13.5	21.0
250	16.7	3.7	42.3	57.0	46.6	**14.7**	12.3	15.4	14.1	21.3
500	17.3	–	42.5	57.3	47.0	**14.8**	12.3	**15.7**	**14.3**	**21.6**
1,000	**21.4**	–	**45.6**	**59.3**	**49.4**	14.3	**12.7**	15.4	14.2	21.4
gain	×1.8	×1.4	×1.4	×1.1	×1.2	×1.3	×1.2	×1.4	×1.6	×1.6
Llama 3.3 70B										
0	8.3	1.4	29.1	39.1	36.4	8.9	9.1	9.2	7.6	13.5
1	8.8	2.0	33.3	42.2	38.3	10.3	10.1	10.5	8.9	14.3
5	10.7	2.2	34.7	44.1	39.5	12.0	10.3	11.5	10.4	**15.4**
10	**10.9**	**2.3**	35.3	44.7	40.0	**12.3**	10.5	11.5	10.6	15.4
25	10.4	2.0	35.7	44.6	40.2	11.3	10.5	11.6	9.8	15.3
50	10.5	2.0	36.4	45.2	40.3	11.6	10.9	10.6	10.5	15.2
100	10.8	**2.3**	37.5	45.3	40.2	11.1	10.9	11.8	10.5	14.6
250	10.5	–	38.3	45.4	40.0	11.3	**11.3**	**12.2**	**11.6**	14.0
500	10.5	–	**38.6**	**46.7**	**40.6**	10.6	11.1	11.7	10.3	14.7
1,000	–	–	–	–	–	10.7	10.3	10.5	9.6	13.3
gain	×1.3	×1.6	×1.3	×1.2	×1.1	×1.4	×1.2	×1.3	×1.5	×1.1
Gemma 3 27B										
0	9.2	3.9	30.1	46.6	38.3	9.4	9.6	10.0	6.9	11.0
1	10.5	4.9	34.7	48.1	39.8	10.6	9.9	11.4	8.8	12.6
5	11.5	5.5	36.3	49.7	41.2	11.7	10.6	12.2	10.1	13.9
10	12.0	5.7	36.8	49.7	41.5	**12.1**	10.7	12.1	10.3	14.4
25	12.1	6.0	37.6	49.7	41.9	11.5	10.9	12.3	10.9	15.0
50	12.7	6.0	38.0	50.3	42.5	11.8	11.0	12.6	11.0	15.2
100	13.3	6.4	39.5	50.6	**43.1**	11.8	11.1	12.7	**11.8**	15.6
250	13.3	**6.5**	**39.9**	**50.9**	43.0	11.7	**11.6**	12.7	**11.8**	**15.7**
500	**13.5**	–	39.8	50.8	42.6	11.8	11.4	**12.8**	11.7	**15.6**
1,000	–	–	–	–	–	11.9	11.2	12.7	**11.8**	15.1
gain	×1.5	×1.7	×1.3	×1.1	×1.1	×1.3	×1.2	×1.3	×1.7	×1.4

Table 18: X→eng scaling with random selection (spBLEU). Formatting conventions as described in §E.

Shots	Order	spBLEU					chrF++				
		vmw	zgh	lld	mfe	apd	vmw	zgh	lld	mfe	apd
Gemini 2.5 Flash											
BM25 retrieval											
1	Default	7.7	22.2	24.1	35.7	25.9	28.7	28.6	43.8	57.5	44.6
	Similar first	7.9	22.2	23.8	35.4	25.7	28.8	28.6	43.7	57.3	44.6
	Dissimilar first	7.9	21.8	24.1	35.6	25.8	28.8	28.6	43.8	57.4	44.7
5	Default	9.7	26.1	29.0	37.4	26.3	30.9	32.0	47.6	58.6	45.7
	Similar first	9.7	25.8	29.2	37.6	26.1	30.9	32.1	47.6	58.9	45.5
	Dissimilar first	9.7	25.6	28.9	36.9	26.3	30.9	31.7	47.5	58.5	45.6
10	Default	10.7	28.0	31.1	38.3	26.3	32.1	33.8	49.1	59.4	45.7
	Similar first	10.7	27.5	30.9	38.4	26.5	32.0	33.5	49.0	59.5	46.0
	Dissimilar first	10.8	27.5	30.9	38.0	26.6	32.0	33.4	49.0	59.2	46.1
25	Default	12.2	29.8	33.4	39.0	27.1	33.5	35.5	50.8	59.9	46.7
	Similar first	12.2	29.5	33.3	38.9	27.0	33.4	35.3	50.7	59.8	46.6
	Dissimilar first	12.3	29.5	33.3	38.9	26.9	33.4	35.6	50.7	60.0	46.6
Dense retrieval (Qwen3-Embedding-8B)											
1	Random	7.8	22.0	23.9	35.0	26.0	28.6	28.7	43.8	57.1	44.7
	Similar first	7.7	22.4	24.1	35.1	26.0	28.5	29.0	44.0	57.2	44.6
	Dissimilar first	7.6	21.8	24.1	35.0	26.0	28.6	28.4	43.9	57.0	44.7
5	Random	9.5	24.7	28.0	36.8	26.2	30.8	31.1	46.9	58.5	45.4
	Similar first	9.5	25.0	28.1	37.1	26.2	30.6	31.3	47.0	58.5	45.4
	Dissimilar first	9.5	25.0	27.9	36.9	26.3	30.7	31.1	46.8	58.4	45.4
10	Random	10.5	26.8	29.7	37.6	26.1	31.9	33.0	48.2	58.9	45.6
	Similar first	10.6	26.7	29.9	37.9	26.1	31.8	32.6	48.2	59.2	45.7
	Dissimilar first	10.5	27.1	29.9	37.9	26.4	31.9	32.9	48.2	59.1	45.9
25	Random	11.8	29.0	32.2	38.8	27.2	33.3	34.7	49.8	59.7	46.9
	Similar first	11.9	29.2	32.4	38.0	27.0	33.3	34.9	50.0	59.3	46.5
	Dissimilar first	11.7	28.8	32.7	38.5	27.2	33.1	34.7	50.3	59.7	46.7
GPT-4.1											
BM25 retrieval											
1	Default	5.1	10.0	12.3	34.6	24.3	23.5	18.1	33.8	55.4	43.9
	Similar first	5.3	10.5	12.4	34.6	24.4	23.3	18.2	33.9	55.4	43.9
	Dissimilar first	5.0	10.4	12.4	34.9	24.5	23.4	18.4	34.0	55.7	43.9
5	Default	6.7	13.4	16.4	36.5	24.8	26.5	21.1	37.6	57.0	44.6
	Similar first	6.4	13.6	16.4	36.1	24.7	26.2	21.3	37.6	56.9	44.5
	Dissimilar first	6.5	13.7	16.3	36.0	25.0	26.2	21.3	37.7	56.8	44.8
10	Default	7.5	15.1	19.0	36.7	25.0	27.9	22.7	39.9	57.3	44.9
	Similar first	7.4	15.1	18.9	36.6	25.3	27.6	22.4	39.8	57.3	45.1
	Dissimilar first	7.6	14.9	18.8	36.7	25.3	27.9	22.4	39.8	57.3	45.1
25	Default	9.0	16.6	22.6	38.0	25.5	29.5	23.9	43.1	58.3	45.4
	Similar first	8.7	16.2	22.6	37.6	25.7	29.4	23.6	43.0	58.0	45.6
	Dissimilar first	8.8	16.3	22.7	37.7	25.7	29.5	23.8	43.1	58.1	45.5
Dense retrieval (Qwen3-Embedding-8B)											
1	Random	5.1	9.9	11.6	34.4	24.7	23.4	18.2	33.5	55.4	44.0
	Similar first	5.1	9.6	11.5	34.5	24.7	23.3	18.1	33.4	55.4	44.0
	Dissimilar first	5.0	9.7	11.6	34.5	24.8	23.2	18.1	33.4	55.5	44.1
5	Random	6.3	12.1	14.6	36.1	24.6	25.9	20.4	36.1	56.8	44.5
	Similar first	6.3	12.1	14.6	36.1	24.6	25.8	20.3	36.2	56.9	44.5
	Dissimilar first	6.2	12.2	14.8	35.9	24.8	25.6	20.5	36.3	56.7	44.6
10	Random	7.3	13.2	16.4	36.3	24.9	27.3	21.5	38.0	57.1	44.9
	Similar first	7.1	13.7	16.7	36.5	25.0	27.2	21.8	38.1	57.2	44.9
	Dissimilar first	7.2	13.5	16.6	36.3	24.9	27.3	21.7	38.2	57.1	44.8
25	Random	8.1	14.8	20.4	37.0	25.5	28.9	22.7	41.2	57.7	45.5
	Similar first	8.0	14.7	20.1	36.9	25.5	28.6	22.8	41.0	57.6	45.4
	Dissimilar first	8.2	15.0	20.5	37.0	25.4	28.9	22.7	41.4	57.6	45.4

Table 19: Effect of example ordering within BM25 and dense retrieval (eng→X, FLORES). Left: spBLEU; right: chrF++. For BM25, "Default" randomly shuffles the BM25-retrieved top-k examples (the setting used in the main BM25 results, §3.2); for dense retrieval, "Random" shuffles the retrieved examples. "Similar first" places the most-similar example at the start of the prompt (farthest from the query); "Dissimilar first" places it immediately before the query. Differences across orderings are negligible (≤ 0.5 chrF++).

Shots	Order	eng→X										X→eng									
		spBLEU					chrF++					spBLEU					chrF++				
		vmw	zgh	lld	mfe	apd	vmw	zgh	lld	mfe	apd	vmw	zgh	lld	mfe	apd	vmw	zgh	lld	mfe	apd
Gemini 2.5 Flash																					
1	Random	6.8	20.0	18.1	35.5	26.6	27.5	27.2	39.4	56.8	44.8	24.8	25.0	42.9	54.9	43.1	44.5	44.4	63.3	71.2	64.2
	L2S	6.9	20.2	18.3	35.3	26.5	27.5	27.4	39.6	56.7	44.7	24.9	24.8	42.9	54.6	43.2	44.6	44.3	63.4	71.0	64.2
	S2L	6.8	19.9	18.3	35.4	26.6	27.3	26.9	39.6	56.7	44.8	24.7	25.1	43.0	54.9	43.1	44.5	44.5	63.4	71.2	64.1
	Pair-L2S	6.8	20.7	18.2	35.5	26.3	27.3	27.7	39.5	56.8	44.6	25.1	25.0	42.7	54.8	43.0	44.7	44.3	63.2	71.1	64.1
	Pair-S2L	6.9	20.2	18.2	35.2	26.2	27.4	27.2	39.5	56.6	44.5	25.0	25.1	42.9	54.7	43.0	44.6	44.4	63.4	71.1	64.0
5	Random	7.3	22.1	25.6	37.3	25.2	28.2	28.7	45.1	58.3	44.3	26.4	26.8	43.6	55.5	44.8	45.6	46.1	63.7	71.6	64.8
	L2S	7.3	21.8	25.3	37.2	25.5	28.0	28.6	44.9	58.1	44.4	26.6	26.9	43.4	55.4	44.6	45.9	46.3	63.7	71.5	64.9
	S2L	7.5	22.3	25.1	36.7	25.6	28.3	28.8	44.7	58.1	44.4	26.8	27.2	43.4	55.6	44.4	45.9	46.3	63.7	71.6	64.9
	Pair-L2S	7.3	21.9	25.5	37.1	26.7	28.0	28.6	45.0	58.3	45.4	26.4	26.9	43.8	55.5	44.6	45.7	46.1	63.8	71.5	64.9
	Pair-S2L	7.5	22.4	25.3	36.9	25.7	28.3	29.0	44.8	58.1	44.9	26.9	27.4	43.9	55.2	44.5	45.8	46.3	64.0	71.4	64.9
10	Random	8.0	22.3	26.1	37.7	26.0	28.9	29.0	45.5	58.7	45.5	27.1	27.0	44.1	55.5	44.7	46.3	46.3	63.9	71.6	65.0
	L2S	8.0	22.5	26.1	37.4	26.6	29.0	29.0	45.4	58.6	45.6	26.8	27.2	44.4	55.8	44.8	45.8	46.6	64.1	71.8	65.0
	S2L	8.3	22.6	26.0	37.8	26.5	29.1	28.9	45.4	59.0	45.8	26.4	27.5	44.1	56.1	44.7	45.9	46.5	63.8	71.9	65.1
	Pair-L2S	7.7	22.3	25.7	37.3	26.9	28.6	28.9	45.1	58.5	46.1	26.8	27.5	44.5	55.9	44.4	46.0	46.7	64.1	71.8	64.8
	Pair-S2L	8.1	22.5	26.2	37.7	26.8	29.2	29.3	45.5	58.7	46.0	26.0	27.2	44.3	55.6	45.1	45.5	46.5	63.9	71.7	65.3
25	Random	8.9	23.6	27.4	38.5	26.9	30.6	30.0	46.5	59.4	46.2	27.5	27.8	44.8	56.0	45.2	46.3	47.1	64.3	71.9	65.2
	L2S	9.1	23.7	27.2	37.9	27.8	30.8	29.9	46.3	59.1	46.9	27.6	27.5	44.5	55.9	45.5	46.8	47.0	64.1	71.8	65.6
	S2L	9.0	23.9	27.2	38.6	27.0	30.7	30.1	46.2	59.4	46.3	27.8	27.8	45.1	55.7	45.8	46.7	47.1	64.3	71.8	65.6
	Pair-L2S	9.3	23.8	27.3	37.8	27.4	30.7	29.8	46.3	59.1	46.4	27.9	27.9	44.8	56.5	45.7	46.9	47.2	64.4	72.2	65.5
	Pair-S2L	8.8	23.9	27.5	38.5	27.5	30.4	30.0	46.4	59.3	46.9	27.7	27.5	44.4	56.1	45.7	46.7	46.8	64.1	71.9	65.4
GPT-4.1																					
1	Random	4.2	7.2	10.3	34.5	24.8	21.7	16.3	32.2	55.1	43.9	13.2	2.8	36.7	54.7	42.5	30.2	19.4	57.2	71.1	64.2
	L2S	4.2	7.8	10.2	34.3	24.9	21.6	16.4	32.1	55.0	44.0	13.1	2.8	36.6	54.6	42.5	30.3	19.4	57.2	71.0	64.2
	S2L	4.4	7.6	10.2	34.2	24.7	21.7	16.9	32.0	54.9	43.8	13.1	2.6	36.4	54.7	42.6	30.3	19.2	56.9	71.1	64.3
	Pair-L2S	4.2	7.8	10.2	34.4	24.8	21.7	16.6	32.0	55.0	43.8	13.0	2.6	36.7	54.8	42.5	30.1	19.3	57.1	71.2	64.3
	Pair-S2L	4.0	8.0	10.1	34.4	24.9	21.6	16.7	32.0	55.1	43.9	13.1	2.5	36.7	54.7	42.5	30.2	19.3	57.2	71.1	64.3
5	Random	4.5	9.4	11.2	35.6	24.3	23.2	17.9	33.0	56.1	43.6	14.6	3.4	38.1	55.4	44.3	31.6	20.0	58.4	71.7	65.3
	L2S	4.3	8.8	11.1	35.4	24.6	22.9	17.7	32.9	56.0	43.9	14.2	3.3	38.1	55.2	43.9	31.5	20.2	58.4	71.5	65.1
	S2L	4.4	9.1	11.0	35.5	24.3	23.1	17.7	32.8	56.0	43.6	14.6	3.2	38.5	55.4	44.1	31.6	19.9	58.7	71.6	65.1
	Pair-L2S	4.5	9.1	10.9	35.6	24.6	23.1	17.7	32.9	56.2	44.0	14.6	3.4	38.1	55.4	44.1	31.7	20.2	58.4	71.6	65.2
	Pair-S2L	4.5	9.3	11.0	35.7	24.5	23.2	17.7	32.9	56.2	43.6	14.5	3.3	38.2	55.3	44.1	31.6	19.9	58.5	71.6	65.1
10	Random	4.4	9.6	11.6	36.4	26.1	23.6	18.5	33.5	56.9	45.5	14.8	3.7	39.2	55.9	44.5	31.5	20.3	58.9	72.0	65.4
	L2S	4.2	9.5	11.6	36.0	26.1	23.3	18.4	33.7	56.8	45.5	14.5	3.7	39.1	55.7	44.8	31.6	20.4	59.0	71.9	65.5
	S2L	4.5	9.8	11.6	35.9	25.8	23.7	18.3	33.6	56.6	45.3	14.7	3.6	39.1	55.7	44.2	31.6	20.2	59.0	71.9	65.2
	Pair-L2S	4.3	9.7	11.8	36.3	25.8	23.6	18.4	33.8	56.9	45.2	14.4	3.6	39.0	55.7	44.6	31.7	20.4	58.9	71.9	65.4
	Pair-S2L	4.6	9.6	11.5	36.5	25.5	23.6	18.3	33.4	56.9	44.8	14.6	3.5	38.8	56.0	44.6	31.5	20.3	58.6	72.1	65.4
25	Random	5.1	10.0	13.3	36.6	26.1	25.3	18.8	35.3	57.2	45.7	15.4	3.5	40.4	56.6	45.6	32.2	20.3	59.8	72.4	66.0
	L2S	4.9	9.9	13.5	36.4	26.1	25.0	18.7	35.7	57.2	46.0	15.0	3.5	39.7	56.2	45.8	32.1	20.4	59.4	72.2	66.0
	S2L	5.1	10.0	13.4	36.3	26.3	25.0	18.7	35.5	57.1	46.0	15.0	3.8	40.2	56.2	45.8	31.8	20.5	59.6	72.3	66.1
	Pair-L2S	5.1	9.9	13.3	36.3	26.3	25.0	18.7	35.6	57.1	46.1	15.2	3.8	39.7	56.1	45.5	32.3	20.6	59.5	72.2	65.9
	Pair-S2L	5.1	9.8	13.3	36.9	26.0	25.1	18.5	35.2	57.5	45.9	15.4	3.8	39.7	55.6	45.7	32.2	20.5	59.3	71.8	66.0

Table 20: Effect of length-based example ordering. Random: no ordering; L2S: Long-to-Short; S2L: Short-to-Long. Pair-L2S/Pair-S2L: ordered by the combined source+target length.