SMOL: Professionally translated parallel data for 115 under-represented languages

Isaac Caswell[ʕ*] Elizabeth Nielsen[ʕ*] Jiaming Luo[ʕ] Colin Cherry[ʕ]
Geza Kovacs[ʕ] Hadar Shemtov[ʕ] Partha Talukdar[ʕ] Dinesh Tewari[ʕ]
Baba Mamadi Diane[ɔ][nqo] Djibrila Diane[ɔ][nqo] Solo Farabado Cissé[ɔ][nqo]
Koulako Moussa Doumbouya[ð][nqo] Edoardo Ferrante[ʒ][lij] Alessandro Guasoni [ʒ][lij]
Christopher Homan[π][dje] Mamadou K. Keita[π][dje;mos] Sudhamoy DebBarma[α][trp] Ali Kuzhuget[θ][tyv;ru]
David Anugraha[ð][id] Muhammad Ravi Shulthan Habibi[æ][id] Sina Ahmadi [ü][ku++]
Anthony Munthali[ŵ][tum] Jonathan Mingfei Lau[ʕ][粵] Jonathan Eng[ʕ][粵]

[ʕ]Google {Research, Deepmind} [ð]Stanford University [ɔ]NKo USA INC
[ʒ]Conseggio pe-o patrimònio linguistico ligure [æ]Universitas Indonesia [ü]University of Zurich
[π]Rochester Institute of Technology [θ]tyvan.ru

[nqo;lij;dje;mos;trp;tyv;id;ku++;tum;粵] Led volunteer contributions for NKo, Ligurian, Zarma, Mooré, Kokborok, Tuvan, Russian, Indonesian, Kurdish languages, Tumbuka, and Cantonese, respectively

[*]{icaswell, eknielsen}@google.com

Abstract

We open-source SMOL (*Set of Maximal Overall Leverage*),[1] a suite of training data to unlock machine translation for low-resource languages. SMOL has been translated into 123 under-resourced languages (125 language pairs),[2] including many for which there exist no previous public resources, for a total of 6.1M translated tokens. SMOL comprises two sub-datasets, each carefully chosen for maximum impact given its size: SMOLSENT, a set of sentences chosen for broad unique token coverage, and SMOLDOC, a document-level resource focusing on a broad topic coverage. They join the already released GATITOS for a trifecta of paragraph, sentence, and token-level content. We demonstrate that using SMOL to prompt or fine-tune Large Language Models yields robust CHRF improvements. In addition to translation, we provide factuality ratings and rationales for all documents in SMOLDOC, yielding the first factuality datasets for most of these languages.

1 Introduction

There exist no professionally-translated data for most of the world's 7000 or so languages, rendering tasks like Machine Translation near impossible. High-quality data is needed. However, it is not clear how best to use a limited budget for an expensive task like professional translation. As shown by the GATITOS dataset (Jones et al., 2023), word-level translations provide large benefits to translation quality for low-resource languages at the lowest cost. However, gains quickly saturate, as single tokens are not very expressive. Sentence-level data is better for a model once token-level data saturates, but it has much more inherent redundancy; and document-level data is even more effective...and more redundant.

In this work, we release the SMOL dataset, which provides professionally translated sentence- and document-level data for 123 LRLs (125 language pairs). SMOL contains two sub-datasets:

- **SMOLSENT**: 863 English sentences covering 5.5k of the most common English tokens,[3] professionally translated into 90 languages.
- **SMOLDOC** 584 English documents covering a wide range of topics, domains, and tokens, generated by an LLM and professionally translated into 103 languages.

We demonstrate the utility of these data for fine-tuning and prompting LLMs for translation, and provide factuality annotations for all documents.

[1] google/smol

[2] Experiments are mostly on a subset of 115 languages, before volunteer translations of additional languages finished. The paper title reflects this.

[3] In this paper, 'token' refers to typographic units as an approximation to words, not subword tokens from a model's vocabulary.

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 1103–1123
November 8–9, 2025. ©2025 Association for Computational Linguistics

2 Related work

There are not many training datasets with human-translated data for Low-Resource Languages (LRLs), where we operationally define LRL as any language beyond the first 100 supported by most traditional crawls and MT providers (enumerated in Appendix section A).

Tatoeba (Tiedemann, 2020) is probably the most multilingual, but it is made of volunteer contributions and of unclear quality. The GATITOS dataset (Jones et al., 2023) consists of a 4000-entry lexicon translated into 170 LRLs, but is only token-level. Most similar to the present work, NLLB-SEED is a high-quality, sentence-level training set of 6k sentences selected from English Wikipedia and professionally translated into 44 LRLs (Team et al., 2022). There are also several professionally-translated evaluation sets, namely FLORES-101 and FLORES-200 (Goyal et al., 2022; Team et al., 2022), and NTREX (Federmann et al., 2022a).

While highly multilingual, professionally translated training data is rare, there is a growing number of bottom-up community data sources organized through research collectives like Masakhane (∀ et al., 2020), Turkish Interlingua (Mirzakhalov et al., 2021a,b), and GhanaNLP (Azunre et al., 2021a); and conferences and workshops like AfricaNLP, AmericasNLP (Mager et al., 2021) and ArabNLP. These datasets are usually generated by researchers fluent in the languages, and are therefore especially high quality. In addition to providing datasets, such efforts frequently also provide models and baselines, or even public interfaces, like the Khaya Translator Web App[4] by GhanaNLP for West African languages, and the lesan.ai[5] translation website for Ethiopian languages.

Participation is especially strong from the African continent, including corpora and models for pan-East-African languages (Babirye et al., 2022), languages from the Horn of Africa (Hadgu et al., 2022), Ethiopian languages (Teferra Abate et al., 2018; Gezmu et al., 2021), Ugandan languages (Akera et al., 2022), Emakhuwa (Ali et al., 2021), South-African languages (Eiselen and Puttkammer, 2014), Setswana and Sepedi (Marivate et al., 2020), Yorùbá (Adelani et al., 2021b,a), Oshiwambo (Nekoto et al., 2022), Igbo (Ezeani et al., 2020), Zulu (Mabuya et al., 2021), Twi (Azunre et al., 2021b), Gbe (Hacheme, 2021), Bambara (Tapo et al., 2021), and Fon (Emezue and Dossou, 2020). Outside of Africa, corpora have been created for languages of the Americas, including for four indigenous languages of Peru in Bustamante et al. (2020), the numerous results on the largely South- and Central American languages from the first AmericasNLP conference (Mager et al., 2021), and the Inuktitut language of Canada (Joanis et al., 2020). Datasets for lower-resourced languages of India have also sprung up, including the 13-language PMIndia (Haddow and Kirefu, 2020), and datasets focused on languages of the Northeast like Mizo (Thihlum et al., 2020), Khasi (Laskar et al., 2021) and Assamese (Laskar et al., 2020). Further West, PARME (Ahmadi et al., 2025) has provided some of the first human-translated content for Kurdish and Iranian languages. Finally, a variety of such datasets and models are available for public use on HuggingFace[6] or Zenodo.[7]

In addition to professionally translated data, there are also several web-crawled datasets for LRLs, including MADLAD (Kudugunta et al., 2023), OSCAR (Ortiz Suárez et al., 2019), Glot500-C (Imani et al., 2023), NLLB (Team et al., 2022), and the Bloom library (Leong et al., 2022).

3 Text Selection

Translation requires significant investment and can't be easily re-done, so great care needs be put into carefully choosing sentences to translate. For both sub-datasets SMOLDOC and SMOLSENT, selection or generation of source text is done in English. Selecting only English has clear biases, but also has advantages—most notably, for N languages, it requires N times less work to quality control. Future work should consider focusing on non-English sources.

3.1 SMOLSENT: Token Set Cover

Our basic motivation for creating SMOLSENT was to help models overcome vocabulary issues, which are common for the lowest-resource languages (Nielsen et al., 2025; Bapna et al., 2022). Therefore, we frame this as a set-cover problem, and pick the smallest set of sentences (from CommonCrawl[8]) that covers the largest set of target tokens. The tokens we chose to cover (the *target set*) were

[4] https://ghananlp.org/project/translator-webapp/

[5] https://lesan.ai/translate

[6] https://huggingface.co/datasets?multilinguality=multilinguality:translation&task_categories=task_categories:translation

[7] https://zenodo.org/communities/africanlp/

[8] https://commoncrawl.org/ we use all available snapshots as of August 20, 2022

Method	ChrF
Random	30.5
Token set-cover	**31.7**
N-gram DWD	30.0
Embedding DWD	27.5

Table 1: Held-out CHRF for data selection approaches

the English side of GATITOS, as well as the most common 2,500 tokens from an English web crawl. Set cover is NP-hard, so we approximate it with a greedy algorithm that iteratively picks the sentence with the highest *coverage percent*, defined as the percentage of its tokens that are in the target set.

Preliminary work on Token Set-Cover To evaluate the token set-cover approach, we started by selecting data from existing web-scraped parallel data. We pretrain a multilingual Neural Machine Translation (NMT) model on parallel data from 294 language pairs from MADLAD-400 (Kudugunta et al., 2023), with nine languages held out to simulate LRLs. We fine-tune this model on sets of existing parallel data in each of the held-out languages, and evaluate on FLORES-200. Details on the experimental set-up in Appendix B.1.

In addition to Greedy Token Set-Cover, we explore two methods that balance data diversity and data quality. First, we implement Ambati et al. (2011)'s 'density-weighted diversity' (DWD) metric, which is an n-gram based metric for diversity and quality. Second, we implement an embedding-based version of DWD, which takes the weighted harmonic mean of perplexity under the Palm 2 model (Anil et al., 2023) (proxy for quality), and embedding distance on mBERT sentence embeddings (proxy for diversity). We apply both methods to the English side of the parallel data only, to simulate the case where we don't yet have LRL translations. As a baseline, we randomly select sentences.

Table 1 shows results after finetuning. Greedy token set-cover performs the best, with diversity-based metrics actively hurting performance.

Researcher in the Loop (RITL) Despite its success in the ablation, Greedy Token Set Cover had several problems when we scaled it to select from among all the English sentences of CommonCrawl. Firstly, it is maximized by honeypots, or nonsense strings dense in content words (Appendix Table B.1); and secondly, it biases towards short sentences, causing length distribution artifacts.

These problems are not easy to solve with heuristics—for example, if you disqualify lists with commas you'll get ones with spaces, if you require sentences to have some function words or token-length diversity, you'll get other sorts of garbled sentences, and so on. However, a dataset like SMOL is small enough to manually inspect. Therefore we develop *Researcher in the Loop Greedy Set-Cover* (Algorithm 1), where the domain expert (the researcher) can look at and edit each individual sentence.[9] The result of this process is SMOLSENT, a set which uses 863 sentences to cover 5519 unique tokens. Qualitatively, SMOLSENT consists of complex sentences with wide vocabulary coverage; quantitative metrics are explored in Appendix B.3.

Algorithm 1 Researcher in Loop Greedy Set Cover

```
Res ← ...            ▷ Sentence reservoir, e.g. CommonCrawl
Toks ← ...                  ▷ Tokens to Cover, e.g. GATITOS
Cov ← { }         ▷ Set-cover, aka output of this algorithm
while not ToCover.empty() do
    batch ← TopScoringSentences(Res, Toks)
    chosen ← ResearchersChoice(batch)
    chosen ← LetResearcherEdit(chosen)
    Cov.add(chosen)
    RemoveCoveredToks(Toks , chosen)
    Res ← LetResearcherDiscardSentences(Res)
    Res.remove(chosen)
end while
return Cov
```

3.2 SMOLDOC: LLMs with prompt mesh

SMOLDOC follows a different and complementary approach. Whereas SMOLSENT consists of a small set of *sentences* that are *selected* from natural text, are *complex*, and cover many *tokens*; SMOLDOC instead consists of *documents* that are *generated* and are *simpler*, but cover many *topics*. It should be noted that the token-coverage approach described above failed resoundingly for longer documents, as the prevalence of the honeypots was magnified.

To generate SMOLDOC, we used a collection of templates to create a few thousand diverse prompts with a wide range of topics, domains, words, tenses, grammatical cases, and registers (e.g. formal, informal). Appendix C.2 gives details and examples.

Corpus Diversity Ranking for SMOLDOC Document-by-document evaluation as described above does not help one understand *corpus diversity*—for example, if an almost identical document appears twice, only one of them should be included.

[9]This work was conducted before the advent of LLMs. Today, this could be simplified using LLMs as autoraters.

Therefore, we rank all candidates by how much new information they add to the corpus, by iteratively finding the document contributing the least new information and removing it, thus ranking all documents. Our criterion for "new information" was the average character 9-gram Inverse Document Frequency (IDF) score of a document—in other words, how rare its substrings were across all of the documents in the pool so far. To down-weight internally repetitive documents, we substracted the fourth moment BREAD score (Caswell et al., 2023).

Language Tiers for SMOLDOC We wanted to translate more data for languages with more speakers. We break the languages into the five different groups, each with a larger subset of the generated documents. Each tier contains translations of the top N documents as ranked by corpus diversity. These can be seen in Appendix Table C.1.

Non-English-centric translations For SMOLDOC, we additionally collected data for four non-English-centric language pairs, from each of the East African languages of Amharic (`am`) and Swahili (`sw`) to each of regionally relevant languages Standard Arabic (`ar`) and Mandarin Chinese (`zh`). Including the reversed versions of these, this yields 8 total language pairs. Because of the difficulty of generating good source material in these languages, we used the existing SMOLDOC translations to Swahili/Amharic as the source text. However, due to the lack of appropriate evaluation sets, it is hard to know the value-add of this data over datasets pivoted through English.

4 Data Collection and Verification

Several languages are contributed by volunteers; they are listed as co-authors.[10] For the other languages, the translation provider we contracted has worked with us for many years, and has a pre-existing relationship with professional translators for all languages in the SMOL datasets. The translators are paid a fair wage, and their identities are contractually kept anonymous to us. We checked the delivery for duplicate translations, anomalous source/target length ratios, and similarity with Google Translate outputs. Very few languages were flagged this way. Following this, we ran FUNLANGID (Caswell, 2024) on all segments and found no issues. Manual inspection turned up several issues with nonunicode fonts (e.g. ô for ɔ) for West African languages, and nonstandard orthography for Santali; these issues were then fixed. The choice of script, orthography and translation variety was challenging for many communities, including Kurdish, Zaza-Gorani and Gilaki languages, all of which have more than one orthography and lack a standard variety.

The largest missing check is for fluency, which is hard to measure without trusted native speakers *outside of the translation agency*, or trusted LLMs; neither of which exist for all SMOL languages.

5 Finetuning and In-Context Learning

We use fine-tuning and ICL as tools to demonstrate the value of the SMOL dataset. As this is a data paper, these experiments are motivated by the maxim *"what could any researcher simply train with public APIs?"* More involved techniques, e.g. Reinforcement-Learning-based approaches, will likely lead to stronger results.

5.1 Evaluation

Since so many language pairs are covered, we evaluate on a combination of all available evaluation sets, namely FLORES-200 (Team et al., 2022), NTREX (Federmann et al., 2022b; Barrault et al., 2019), and an in-house eval set. Since no reliable embedding models exist for these languages, trained metrics are not an option, so we use CHRF (Popović, 2015) as implemented in SacreBleu (Post, 2018)[11] with NFKC unicode normalization as our metric. For ten-shot decoding, exemplars were selected from both sub-datasets of SMOL using CHRF-counterweighted RAG (Appendix D).

5.2 Finetuning Setup and Results

We finetuned Gemini 2.0 Flash for 40 epochs on SMOLDOC, SMOLSENT, a combination of the two (BOTH), and their combination plus GATITOS (BOTH+G). To simplify finetuning, we split SMOLDOC into sentence pairs (SMOLDOCSPLIT).

Results can be seen in Table 3. Finetuning on SMOLSENT gives an average gain of +2.7 CHRF points, and SMOLDOCSPLIT gives +2.6 CHRF points on its languages. Concatenating the two training datasets leads to a gain of +3.3 to +3.6 CHRF points, and adding in GATITOS bumps it

[10] Community contributions of translations or corrections are welcome; please reach out to the authors or join the TUSL Discord.

[11] signature: `case.mixed+numchars.6+numrefs.1 +space.False+tok.none+version.1.3.0`

Set	# languages	Total Dataset Examples	Tokens	Characters	Per Language Pair (LP) Examples	Tokens	Characters
GATITOS	176	693k	784k	4.6M	3.9k	4.5k	26k
SMOLSENT	81	70k	994k	6.1M	863	12k	75k
SMOLDOC	100	27k	5.1M	28M	263	50k	278k
BOTH	115	97k	6.1M	34M	827	52k	294k

Table 2: Statistics for the whole data set (left bloc) and per language-pair (LP) (right bloc) on the two SMOL datasets and their predecessor GATITOS in number of examples, tokens, and characters. The `# languages` column counts translated languages only, not the source languages of English, Swahili, and Amharic.

LP subset →	SMOL-SENT (80 LP)		SMOL-DOC (73 LP)		Intersect (38 LP)		HARD (32 LP)	
Model ↓	0-shot	10-shot	0-shot	10-shot	0-shot	10-shot	0-shot	10-shot
G. TRANSLATE	-	-	-	-	**43.2**	-	-	-
NLLB-54B	-	-	-	-	40.0	-	-	-
CLAUDE 3.5 SON.	37.5	**39.7**	38.3	**40.9**	41.0	42.8	30.0	**33.5**
GPT-4O	29.9	34.1	31.8	36.3	35.4	38.5	15.9	23.7
GEMINI 2.0 PRO	**38.9**	38.9	**39.9**	40.3	42.6	42.2	**31.4**	31.7
GEMINI-2.0 FLASH	35.6	38.4	36.9	39.7	40.2	41.4	26.3	30.4
+ SMOLSENT	38.3	38.3	38.8	38.8	40.6	40.6	32.5	32.6
+ SMOLDOC	35.3	35.4	39.5	39.5	41.2	41.2	31.8	31.8
+ BOTH	38.9	38.9	40.5	40.5	41.8	41.8	33.4	33.4
+ BOTH+G	**39.4**	**39.3**	**41.0**	**40.9**	**42.1**	**42.2**	**33.9**	**33.9**
Δ_{FT}	+3.8	+0.9	+4.1	+1.2	+1.9	+0.8	+7.6	+3.5

Table 3: Finetuning Gemini 2.0 Flash on SMOL for four subsets of language pairs. The first two columns show LPs in SMOLSENT and those in SMOLDOC, to show the different effects of each split. The third shows those in both SMOL datasets AND the closed domain NMT models, for an even comparison to NMT models. Finally, the HARD column shows LPs in both SMOL splits but NOT in Google Translate, or not closely related to a language in Google Translate, to approximate the especially hard languages to learn.

up to +3.8 to +4.1 CHRF points, passing all baselines except Google Translate. The 10-shot RAG results on the un-tuned model are very close to the finetuned 0-shot results, and the finetuned models show no benefit from multi-shot decoding, suggesting that these are two different ways of giving the same information—inference-time versus training time. The 10-shot random results (not included in table) were much lower.

Gains were highest on languages that are not related to mid- or high-resource languages, and lowest on dialects close to major languages. As a heuristic to measure this, we exclude languages that are on Google Translate or closely related to languages on it (Appendix G). The average gain on these languages jumps to +7.6 CHRF.

Figure 1 shows the learning curve on a development subset of 37 languages. Although it may be surprising that so many epochs are needed before convergence, we found that further increasing learning rate led to overfitting. The sharp drop near the beginning suggests a domain mismatch between pretraining and finetuning, and suggests that

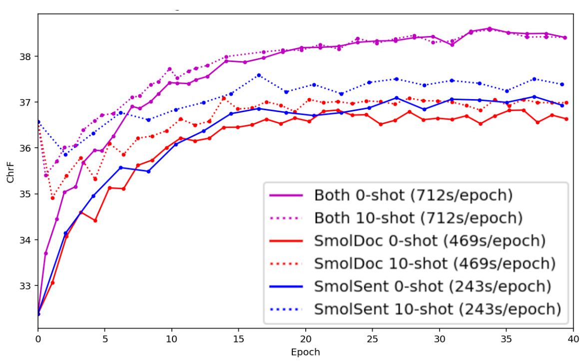

Figure 1: Training curves (CHRF) for finetuned models on a subset of 37 en→xx language pairs, trained on SMOLDOC, SMOLSENT, and their combination BOTH.

the same data could be used much more effectively with a better training set-up than explored here.

5.3 The Problem with xx→en training

Our initial experiments used all data for both en→xx and xx→en . However, the models lost performance on all tasks. The root cause turned out to be the multiway-parallel data with English tar-

Rating	Definition of Rating
N/A	True/False does not apply here. Most stories, dialogues, or fictional works would be considered N/A, unless they are promoting a falsehood about the real world.
Not Sure	Claims are made that may not be true, but you aren't sure. Choose this if it would take over 10 minutes to verify the factuality of the claim.
No Issues	All claims are factual and accurate. (Out-of-date is fine, e.g. "Barack Obama is the US President")
Minor Issue(s)	There are small inaccuracies. E.g., it may be broadly correct but frame something in a misleading way.
Clear Issue(s)	There are clear mistakes in factuality.

Table 4: Factuality Rubric

gets. LLMs are especially susceptible to repetition in data (Lee et al., 2022), and with 115 language pairs, for every one epoch over the data, the model saw about 115 epochs for each individual target sentence. Therefore, it wildly overfit and lost performance on all language pairs. Mitigating such overfitting is an important research direction to pursue, since many promising datasets are multiway parallel, e.g. FLORES-101 (Goyal et al., 2022), FLORES-200 (Team et al., 2022), NTREX (Federmann et al., 2022b; Barrault et al., 2019), and others. However, this is out of scope for the present paper, so we restrict our experiments to en→xx .

Seeing the same *source* many times likely also has deleterious effects and should also be studied; but these effects, if they exist, are small enough that we were still able to see net gains.

6 Factuality Review

Since SMOLDOC contains LLM-generated sources, they contain some factual inaccuracies. We therefore do a full human audit and assign factuality codes to each document. Each of the 584 documents was rated by three raters. Each rating is accompanied by a detailed explanation, including sources cited. Inter-annotator agreement was high, with Cohen's κ between each pair of raters between 0.82-0.87. The error code distribution can be seen in Figure 2. The rubric is presented in Table 4.

All ratings and rationales are made available. In addition, each datum in SMOLDOC is given with a simple `factuality` annotation, which has the value `has_errors` if any one of the ratings was any of `Minor Issues` or `Clear Issues`, and `ok` otherwise. For some use-cases, like question-answering, practitioners may want to filter out nonfactual data; for others, like translation, one may not be troubled by factual errors. In addition to filtering, this also provides the first factuality dataset for most of these languages.

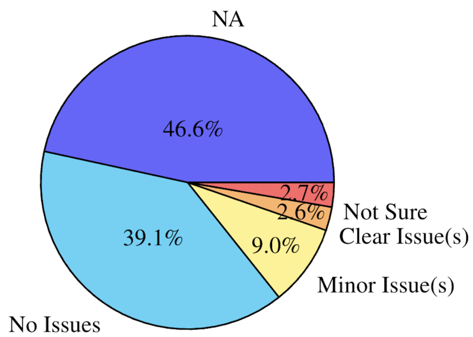

Figure 2: SMOLDOC factuality ratings.

7 Conclusion

We have open-sourced the SMOL dataset, a professionally-translated dataset covering 123 low resource languages and targeting the tasks of translation and factuality. It comprises SMOLDOC and SMOLSENT, two training datasets with the complementary strengths of sentence selection (complex, and high token coverage) and document generation (contextual, varied domains, simpler sentences) respectively. We demonstrate that finetuning Gemini 2.0 Flash on these yields to substantial improvements in translation quality. SMOL joins a growing body of resources to support underserved languages in the age of AI.

8 Limitations

The SMOL data would benefit from a more thorough review, audit, and correction from community members outside of the translators who created it. Future work on SMOL-like datasets should also focus on non-English source text that is not only maximally authentic in the given language, but also covers the topics and concepts most relevant to those languages. This approach is more difficult and would require significant work and review to do correctly. Finally, more research is needed to understand and prevent the overfitting that comes with multi-way parallel data.

References

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Awokoya, and Cristina España-Bonet. 2021a. MENYO-20k: A multi-domain English-Yorùbá corpus for machine translation and domain adaptation. *CoRR, arXiv:2103.08647v1*.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021b. "The Effect of Domain and Diacritics in Yoruba–English Neural Machine Translation". In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

Sina Ahmadi, Rico Sennrich, Erfan Karami, Ako Marani, Parviz Fekrazad, Gholamreza Akbarzadeh Baghban, Hanah Hadi, Semko Heidari, Mahîr Dogan, Pedram Asadi, Dashne Bashir, Mohammad Amin Ghodrati, Kourosh Amini, Zeynab Ashourinezhad, Mana Baladi, Farshid Ezzati, Alireza Ghasemifar, Daryoush Hosseinpour, Behrooz Abbaszadeh, Amin Hassanpour, Bahaddin Jalal Hamaamin, Saya Kamal Hama, Ardeshir Mousavi, Sarko Nazir Hussein, Isar Nejadgholi, Mehmet Ölmez, Horam Osmanpour, Rashid Roshan Ramezani, Aryan Sediq Aziz, Ali Salehi Sheikhalikelayeh, Mohammadreza Yadegari, Kewyar Yadegari, and Sedighe Zamani Roodsari. 2025. "PARME: Parallel Corpora for Low-Resourced Middle Eastern Languages". In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30032–30053, Vienna, Austria. Association for Computational Linguistics.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. 2022. Machine Translation For African Languages: Community Creation Of Datasets And Models In Uganda. In *3rd Workshop on African Natural Language Processing*.

Felermino D. M. A. Ali, Andrew Caines, and Jaimito L. A. Malavi. 2021. Towards a parallel corpus of Portuguese and the Bantu language Emakhuwa of Mozambique. *CoRR*, abs/2104.05753.

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2011. "Multi-Strategy Approaches to Active Learning for Statistical Machine Translation". In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. 2021a. NLP for Ghanaian Languages. *CoRR*, abs/2103.15475.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. 2021b. English-Twi Parallel Corpus for Machine Translation. *CoRR*, abs/2103.15625.

Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, Ronald Ogwang, Jeremy Tusubira Francis, Jonathan Mukiibi, Medadi Ssentanda, Lilian D Wanzare, and Davis David. 2022. Building Text and Speech Datasets for Low Resourced Languages: A Case of Languages in East Africa. In *3rd Workshop on African Natural Language Processing*.

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.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. "Findings of the 2019 Conference on Machine Translation (WMT19)". In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. "No Data to Crawl? Monolingual Corpus Creation from PDF Files of Truly low-Resource Languages in Peru". In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Isaac Caswell. 2024. FunLangID. `https://github.com/google-research/url-nlp/tree/main/fun_langid`.

Isaac Caswell, Lisa Wang, and Isabel Papadimitriou. 2023. Separating the wheat from the chaff with BREAD: An open-source benchmark and metrics to detect redundancy in text. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 324–338, Singapore. Association for Computational Linguistics.

Moussa Koulako Bala Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory 2. Condé, Kalo Mory Diané, Chris Piech, and Christopher Manning. 2023. Machine Translation for Nko: Tools, Corpora and Baseline Results.

Roald Eiselen and Martin Puttkammer. 2014. Developing text resources for ten South African languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3698–3703, Reykjavik, Iceland. European Language Resources Association (ELRA).

Chris Chinenye Emezue and Femi Pancrace Bonaventure Dossou. 2020. FFR v1.1: Fon-French neural machine translation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 83–87, Seattle, USA. Association for Computational Linguistics.

Ignatius Ezeani, Paul Rayson, Ikechukwu E. Onyenwe, Chinedu Uchechukwu, and Mark Hepple. 2020. Igbo-english machine translation: An evaluation benchmark. *CoRR*, abs/2004.00648.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022a. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022b. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online.

Andargachew Mekonnen Gezmu, Andreas Nürnberger, and Tesfaye Bayu Bati. 2021. Extended parallel corpus for amharic-english machine translation. *CoRR*, abs/2104.03543.

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.

Gilles Hacheme. 2021. English2gbe: A multilingual machine translation model for {Fon/Ewe} gbe. *arXiv preprint arXiv:2112.11482*.

Barry Haddow and Faheem Kirefu. 2020. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Asmelash Teka Hadgu, Gebrekirstos G. Gebremeskel, and Abel Aregawi. 2022. HornMT. `https://github.com/asmelashteka/HornMT`.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. "GATITOS: Using a New Multilingual Lexicon for Low-resource Machine Translation". In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji Darsh Kaushik, Partha Pakray, and Sivaji Bandyopadhyay. 2021. EnKhCorp1.0: An English–Khasi corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 89–95, Virtual. Association for Machine Translation in the Americas.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. EnAsCorp1.0: English-Assamese corpus. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 62–68, Suzhou, China. Association for Computational Linguistics.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. "Deduplicating Training Data Makes Language Models Better". In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rooweither Mabuya, Jade Abbott, and Vukosi Marivate. 2021. Umsuka English - isiZulu Parallel Corpus. Thank you to Facebook Research for funding the creation of this dataset.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. "Investigating an Approach for Low Resource Language Dataset Creation, Curation and Classification: Setswana and Sepedi". In *Proceedings of the first workshop on Resources for African Indigenous Languages*, pages 15–20, Marseille, France. European Language Resources Association (ELRA).

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr, et al. 2021a. A large-scale study of machine translation in the Turkic languages. *arXiv preprint arXiv:2109.04593*.

Jamshidbek Mirzakhalov, Anoop Babu, Aigiz Kunafin, Ahsan Wahab, Behzod Moydinboyev, Sardana Ivanova, Mokhiyakhon Uzokova, Shaxnoza Pulatova, Duygu Ataman, Julia Kreutzer, et al. 2021b. Evaluating multiway multilingual NMT in the Turkic languages. *arXiv preprint arXiv:2109.06262*.

Wilhelmina Nekoto, Julia Kreutzer, Jenalea Rajab, Millicent Ochieng, and Jade Abbott. 2022. Participatory Translations of Oshiwambo: Towards Sustainable Culture Preservation with Language Technology. In *3rd Workshop on African Natural Language Processing*.

Elizabeth Nielsen, Isaac Rayburn Caswell, Jiaming Luo, and Colin Cherry. 2025. Alligators all around: Mitigating lexical confusion in low-resource machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 206–221, Albuquerque, New Mexico. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores.

Allahsera Auguste Tapo, Michael Leventhal, Sarah Luger, Christopher M. Homan, and Marcos Zampieri. 2021. Domain-specific MT for Low-resource Languages: The case of Bambara-French. *CoRR*, abs/2104.00041.

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.

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. 2018. Parallel corpora for bi-directional statistical machine translation for seven Ethiopian language pairs. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 83–90, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zaitinkhuma Thihlum, Vanlalmuansangi Khenglawt, and Somen Debnath. 2020. Machine Translation of English Language to Mizo Language. In *2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, pages 92–97.

Jörg Tiedemann. 2020. The tatoeba translation challenge - realistic data sets for low resource and multilingual MT. *CoRR*, abs/2010.06354.

A Operational Definition of LRL

In this paper, we operationally define LRL as any language beyond the first 100 supported by most traditional crawls and MT providers. Since there is some variation in which languages exactly this is, we concretize it as the set of 104 languages supported in Google Translate prior to 2020. These are the languages for which launchable quality was possible before LLM-type models like M4 (Arivazhagan et al., 2019; Bapna et al., 2022) and PaLM (Anil et al., 2023) came on these scene. It is also worth noting that since these languages were on the product for much longer, they have much more machine-translated content online from services that used Google Translate for internationalization. These 104 languages are: `af am ar az be bg bn bs ca ceb co cs cy da de el en eo es et eu fa fi fil fr fy ga gd gl gu ha haw hi hmn hr ht hu hy id ig is it he ja jv ka kk km kn ko ku ky la lb lo lt lv mg mi mk ml mn mr ms mt my ne nl no ny pa pl ps pt ro ru sd si sk sl sm sn so sq sr st su sv sw ta te tg th tr uk ur uz vi xh yi yo zh zh-Hant zu`.

Rightly speaking, the languages outside of this set might better be termed "Very Low-Resource" instead of just "Low Resource", since the 104 languages above do include languages like Hawaiian, Javanese, Yiddish, and Hmong, which are by no stretch of the imagination high-resource. We will leave more rigorous definitions to future work.

B SMOLSENT details

B.1 Evaluating the SMOLSENT selection process

In Section 3.1, we describe experiments used to validate the selection process for SMOLSENT. We train a backbone MT system is pretrained on the MADLAD-400 dataset (Kudugunta et al., 2023). The following languages are held out of the training data to be used for fine-tuning experiments: Catalan, Icelandic, Marathi, Turkish, Maltese, Xhosa, Tamil, Basque, and Tajik. The model itself is a 1B parameter encoder-decoder Transformer and is trained from scratch on the MADLAD-400 data. Each of the candidate data selection methods is used to select data from the held-out languages, and then each candidate set is used in turn to finetune the backbone model. The results of this finetune step are reported in Table 1, where the set-cover approach is shown to be most effective.

B.2 Notes on Researcher in the Loop

Researcher In the Loop extends the greedy set cover approach thusly: rather than always picking the highest-scoring sentence, we iteratively show the researcher a batch of the 20 highest scoring sentences according to several scores, and let the researcher pick and optionally edit each sentence at each iteration. At each iteration, the researcher may also remove any number of this batch's sentences from the reservoir. Allowing the researcher to see and edit the sentences allows ensures that the sentences are of high-quality. To deal with the length bias issue, we showed not only sentences that maximize coverage percent, but also that maximize heuristics that weighted the coverage with the number of new tokens hit, like `log(coverage_percent)*n_hits` .

As described in the paper text, this approach is designed to combat issues such as honeypot sentences. Example Honeypot sentences can be seen in Table B.1

Sentence
Individual determine can get prolonged, reduce along with attractive.
Sell hand mood situation connect proper decision today spread true.
Demand indeed off forget act special well treat sometimes notice.
Agree board book oh trust by attractive supply deal together.
Picture exactly could ability impact advance then same admire across.
One physically courage both information language issue laugh common.

Table B.1: Honeypot Sentences for Greedy Selection: CommonCrawl has many sentences packed with content words but with no clear semantics or grammar.

B.3 Corpus statistics on SMOLSENT

To measure “Bang for our Buck” we define the *excess-token ratio* ξ as the number distinct tokens in the set cover divided by the number of target tokens, and use it along with the coverage percent to understand the SMOLSENT dataset. Table B.2 compares corpus statistics of SMOLSENT to four other corpora. `sametoks` picks a random set of sentences from CommonCrawl until it has the same number of tokens as SMOLSENT; this only covers 50% of the target tokens and has an excess-token-ratio ξ ratio of 3.3, much worse than SMOLSENT's value of 2.3. The `samecov` baseline randomly picks common-crawl sentences until it has the same token coverage as SMOLSENT, which necessitates

set	N sent	toks	types	$\xi(\downarrow)$	cov%($\uparrow$)
SMOLSENT	863	12k	5.5k	2.3	99.6%
sametoks	863	12k	3.8k	3.3	50.4%
samecov	57578	877k	38k	23.1	99.6%
SMOLDOC-T1	6979	108k	8.8k	12.3	80.5%
SMOLDOC-T5	820	12k	2.8k	4.3	40.2%

Table B.2: Corpus statistics of SMOLSENT, random selections of sentences from CommonCrawl, and tiers 1 and 5 of SMOLDOC.

set	N langs	ex	tok	char	ex/LP	tok/LP	char/LP
SMOLDOC-t1	5	2.9k	537k	3.0M	584	107k	604k
SMOLDOC-t2	31	14k	2.7M	15M	450	88k	495k
SMOLDOC-t3	24	6.7k	1.2M	6.8M	280	50k	281k
SMOLDOC-t4	8	1.0k	184k	1.0M	126	23k	128k
SMOLDOC-t5	34	2.2k	401k	2.2M	66	12k	65k
all-SMOL	GTr	97k	6.1M	34M	827	52k	294k

Table C.1: Statistics on the languages in the individual tiers of SMOLDOC.

a 67x larger set of sentences, and a correspondingly bloated excess token coverage ratio of 23.1. As a further reference we compare tiers 1 (largest) and 5 (smallest; comparative size to SMOLSENT) of SMOLDOC. As expected of machine-generated text, they have a worse ξ value, corresponding to a narrower spectrum of vocabulary used.

C SMOLDOCDetails

C.1 SMOLDOC data tier details

Per-tier statistics on the SMOLDOC dataset can be seen in Table C.1.

C.2 Details on SMOLDOC prompt creation

To avoid biases from overly tempted prompts, we put in significant effort to make sure the prompts were all very different. Each prompt drew at random from the following elements:

- random selection of English words to use in the response
- one of 600 manually created topics, e.g. "volcanic eruptions" or "A special tree"
- one of 50 tone/tense categories, e.g. "Please use the subjunctive mood.", "Use an effusive tone."
- A style prompt, e.g. "You are the author R.K. Narayan." or "You are a mother talking to her son."
- A text modality, e.g. story/dialogue/essay

In addition to this, we added a few more sources of prompts:

- Prompts based on urls, meant to simulate different web domains, like Wikipedia and reddit.
- Prompts based on continuing the sentences from SMOLSENT
- Prompts based on current events, history, and daily life in different countries
- Special effort was made to include dialogues (to get more spoken register) and recipes (unique domain that may also be important to translate).

For each prompt, we generated 8 responses (T=0.7). These were ranked by their simple token density (unique tokens over total tokens), and the top two were chosen for consideration. Using the Researcher-in-the-loop mentality ("measure twice cut once!"), we went over 1000+ responses by hand and scored/edited them. This was mainly to filter out questionable or boring responses. A typical paragraph scored as 0 would be LLM-speak like *"X is a complex and multifaceted problem with no easy solution. Here are some suggestions. Keep in mind that there is no one-size-fits all solution, and ultimately, the choice is up to you. [...]"*.

Example prompts can be seen in Table C.2.

C.3 SMOLDOC Errata

Orthographies Several languages use irregular orthographies. Most notable is Mooré/Mossi (mos), where different translators have used a variety of different conventions. After soliciting community feedback, we plan to release standardized versions to the data.

Example Prompts
You are Ernest Hemingway. Write a dialogue about road rage. Use a didactic tone.
Write a 1-paragraph story concerning an Irish wake.
You are a teenager talking to his friend. Please carefully craft a 1-paragraph bit about an engineer who subsists off coffee. Try to include the words "confirmed", "move" and "above".
Give a typical, yet interesting, example of something you would find on reddit.
Please write a few paragraphs about challenges facing Ethiopia.
Please write a long passage starting with 'Mum and Dad pause their debate when we hear this creepy clacking that sounds like hail falling.'
Write a recipe for baking an almond cake.

Table C.2: Representative sample of prompts use to generate the documents for SMOLDOC

document selection When collecting the data for SMOLDOC for the Indian languages, we mistakenly included a variety of documents that fell below the corpus diversity threshold described in Section 3.2.

D N-shot: CHRF-counterweighted RAG

To have a strong baseline for N-shot results, we adopt a RAG-based approach that resembles the greedy set-cover algorithm. For each sentence in the eval set, we want the best coverage of the source sentence n-grams as possible, with the least redundancy among exemplars. Therefore, we iteratively choose the exemplar whose source side has the minimum CHRF to the eval source. However, when counting the true positives in the CHRF calculation, we weight the count of each ngram n_i by $(1+c_i)^{-\alpha}$, where $c_i \in [0, \infty]$ is the number of times n_i has been seen among the exemplars chosen so far, and α is a parameter to control how close this algorithm is to ngram set-cover. We use $\alpha = 2$.The set of exemplars we choose from is the concatenation of SMOLSENT and SMOLDOCSPLIT.

E Prompts for Decoding

For 0-shot prompting, we used the following, fairly wordy prompt, the SL and TL standing for the source and target language name, respectively:

```
You are an expert translator. I am
going to give you some example
pairs of text snippets where the
first is in ${SL} and
the second is a translation of
the first snippet into
${TL}. The sentences
will be written
${SL}: <first sentence>
${TL}: <translated first sentence>
After the example pairs, I am
going to provide another sentence
in ${SL} and I want
you to translate it into
${TL}. Give only the
translation, and no extra
commentary, formatting, or
chattiness. Translate the text
from ${SL} to ${TL}.
```

For finetuned models, there is no need for such a wordy prompt, and indeed it only risks overfitting. Therefore, we used the following minimalist prompt:

```
Translate from ${SL} to ${TL}:
```

F Volunteer contributions

A few languages have extra details that need to be called out here.

F.1 Translations for Cantonese

A volunteer team of Cantonese speakers at Google pulled together to translate the maximal set of SMOL text. Mingfei Lau and Jonathan Eng were the main leaders of this effort, and the contributors to translation and post-editing were (alphabetically): Tsz Yan Au, Emily Awesome, Jason Chan, Siu Man Chan, Vicky Chan, Yiwang Chen, Kinton Cheung, Mingo Choi, Andy Chow, Ashley Chow, Olivia Chow, Daniel (Ying Wai) Fan, Thomas Fung, Vikki Ha, Joshua Kwong, Liam Lee Pong Lam, Jonas Lau, Ying Tung (Grace) Law, Crystal Lee, Aki Leung, Derek Leung, Jackie Leung, Thomas Leung, Mu Li, Alicia Liu, Malena Loosli, Chui McConnell, Ken Ng, Nicholas Ng, Tonia Shen, Helen Shum, Franky Sze, Eric Tang, Tommy Tse, Daniel Wong, Danny Wong, Maggie Wong, Pinki Wong, Jeffrey Yu, Shanelle Yu, Shing Fung Yue, Miranda Zhang, and Willis Zhang.

F.2 Translations for NKo

The initial delivery for the NKo language (`nqo`) had a wide variety of errors. We reached out to the authors from Doumbouya et al. (2023), who did a complete re-translation of the text.

F.3 Translations for Zazaki, Hawrami, and Gilaki

Sina Ahmadi gratefully acknowledges support from the UZH Postdoc Grant (reference number 269093).

F.4 Translations for Zarma

Annotation Pipeline The Zarma translation process of SMOL—all the subsets—was done through a combination of automatic and human in the loop methods. We leveraged some existing tools that our team developed to speed up the annotation process. We first used a baseline bidirectional model that we developed to produce initial translation of the samples. These machine translated samples were then passed through our Zarma grammatical error correction model. This model was built by pre-training gemma-2-9b on Zarma data and fine tuning the checkpoint on grammar error correction data set using Direct Preference Optimization (DPO) settings. The outputs from this stage—both languages side by side—were then given to our team of annotators for review.

The annotators were given some guidelines—in addition to the general guidelines from SMOL—for the annotations. These guidelines include:

- **Word adaptation:** rules for handling technical terms, proper nouns, and domain-specific vocabulary that might not have direct equivalents in Zarma. E.g: all the scientific/technical words remain unchanged; and words that have known french-ized equivalent in Zarma must be used in their french-ized forms (for better understandability).

- **Prioritize understandability:** guidelines to prioritize understandability and fidelity over word-for-word translation. We instructed annotators to focus on creating translations that sound natural and widely understandable by Zarma speakers.

- **Language specific constraints:** language specific guidelines that cannot be generalized.

The pipeline speeds up the process while maintaining the quality, since some of the outputs from the automatic stages were already correct.

Zarma Community Attitude Towards Tech : The Zarma community—and the whole Niger in general—are very open minded regarding technology. When we started our very first resource creation for the Zarma language, we received positive feedback and even help from the community, as long as we developed an openly accessible solution for the community. **For the SMOL annotation, that trust helps us to receive valuable help.** For instance, a government based institution verbally promised to accompany any language preservation—machine learning focus in our case—if the outcome will be open-sourced for community usage.

F.5 Post-Edits for Mooré

Annotation Pipeline The annotation process for Mooré did not involve any automated components; everything was annotated by humans. The annotation focuses more on the guidelines provided by SMOL, in addition to some more as in the Zarma case.

Mooré Community Perspective The Mooré community, similarly to the zarma community, are very open minded towards technology; especially if it touches cultural/language preservation. One main feedback we received from some elders (parents of one of the annotators) was a warning to ONLY USE standard Mooré orthography, not any equivalent. They want the language to be well documented according to the language standards.

F.6 Post-Edits for Indonesian

A volunteer team post-edited translations of `smoldoc` and `smolsent` datasets that had done by Gemini 2.5 Pro. The contributors to translation and post-editing were Muhammad Ravi Shulthan Habibi, David Anugraha, and Genta Indra Winata. The post-edits resulted in about 70% of the machine translations being changed.

Translators agreed that the system output was often too “formal", “stiff", or “awkward". The “formal" translations were furthermore not formal in an acceptible sense, but “too awkward and stiff, even for a more formal situations", as an annotator said. Each word choice might be correct and standard in Indonesian, but when combined in a sentence, the result sounded unnatural. Therefore, the majority of the post-edits focused on making the translations sound more natural.

Nonetheless, overall the system output was already quite reasonable in terms of register. In some cases, though, it leaned toward being too rigid. The post-edits tried to loosen that into a consistent "medium" range, but with some flexibility depending on the style of each sentence (sometimes slightly more formal, sometimes slightly less) so the overall text still feels natural and coherent.

F.7 Translations for Languages of the Russian Federation

Traditionally, speakers of hundreds of Cyrillic-based languages in the Russian Federation translate datasets via Russian. For the success of this project, I (Ali Kuzhuget) first funded a professional translation into Russian, engaging Andrey Anisimov as the main translator. The proofreading was conducted by Farhad Fatkullin, Vice-President of the National League of Translators, together with machine translation specialist David Dalé. I also oversaw formatting correctness and coordinated the overall translation workflow.

In parallel, I supervise the translation of the dataset from Russian into Tuvan, using a dedicated Telegram chatbot for large-scale dataset translation. This tool enables multiple rounds of validation and systematic assessment of translation quality. Currently, representatives of about a dozen Cyrillic languages are in the process of translating the SMOL dataset into their own languages through Russian and/or English (for example, Tuvan, Bashkir, Chuvash, and others), ensuring both linguistic accuracy and cultural relevance.

G Full results

Full per-language results can be seen in Table G.1. Results are sorted by the Δ_{FT}, which is the CHRF of the BOTH model minus the CHRF of the finetuned BOTH model—in other words, how much the finetuning on SMOL improved the baseline model.

Google Translate Languages and their cousins As mentioned in the results section, some languages see only very small improvements from finetuning on SMOL, and others even see losses. These are mainly either high-resource languages, or close relatives to higher-resource languages. In the full table G.1 below, The superscript $^{\text{GTr}}$ indicates a language supported by Google Translate at the time of these experiments; a superscript like $^{\sim\text{xx}}$ means that this language is closely related to the Google-Translate-supported language xx. We only consider the 108 languages that were present on Google Translate at the time of this work.

lang	cat	Δ_{FT}	G2F	+sS	+sD	+sB	+sG	Cld	+RAG	G2P	GPT4o	GTr	NLLB
ee	BOTH	+36.1	3.0	37.7	37.6	39.1	39.2	37.8	40.0	39.5	7.5	**42.7**	40.7
kr	BOTH	+10.8	17.3	25.6	25.9	28.1	28.8	22.7	26.3	20.3	22.2	**32.6**	31.0
kg	BOTH	+9.2	34.9	46.9	36.8	44.1	43.2	43.2	47.0	37.8	29.1	**50.2**	3.4
bem	BOTH	+7.3	40.0	44.8	44.7	47.3	49.2	43.3	47.7	42.3	33.3	**49.7**	41.8
dyu	BOTH	+5.3	17.9	22.5	23.3	23.2	23.7	23.9	**24.4**	21.0	4.5	22.4	12.5
din	BOTH	+4.6	20.3	23.8	22.9	24.9	25.1	23.3	25.9	21.4	1.6	25.1	**26.5**
luo	BOTH	+4.1	37.4	39.1	41.1	41.5	42.0	39.1	**42.0**	39.6	36.1	41.3	39.5
fon	BOTH	+3.6	21.3	24.3	23.9	24.9	25.3	20.4	23.7	23.8	1.9	**25.9**	24.2
bm	BOTH	+3.4	30.8	28.6	35.2	34.2	34.1	34.0	**36.2**	33.9	9.0	35.7	32.2
ak	BOTH	+2.6	35.5	36.1	37.8	38.1	**38.2**	34.4	38.1	37.3	32.2	34.5	33.3
ln	BOTH	+2.5	46.8	48.1	48.4	49.3	**49.3**	44.6	48.3	46.5	45.2	46.4	45.7
wo	BOTH	+1.1	30.3	30.0	30.4	31.4	31.6	31.4	32.2	30.7	29.8	**36.2**	30.9
ff	BOTH	+0.9	25.0	24.4	26.4	25.9	26.5	25.7	26.1	25.2	2.5	25.9	**27.1**
om	BOTH	-0.8	40.1	38.0	39.0	39.3	39.4	39.0	40.2	41.3	38.4	**41.4**	39.1
lg	BOTH	-1.1	42.5	39.9	41.4	41.4	41.7	42.0	43.1	43.5	41.0	**43.6**	41.1
ber	BOTH	-3.2	25.3	20.6	22.9	22.1	21.9	28.5	25.2	31.1	2.8	21.0	**32.4**
trp	SMOLDOC	+29.7	8.4	6.5	37.8	38.1	**39.1**	24.7	35.8	27.2	20.3	35.9	-
mni-M.	SMOLSENT	+26.4	2.9	30.0	1.2	29.3	29.3	29.6	31.8	33.6	1.3	**45.6**	0.8
gaa	BOTH	+23.1	22.7	44.5	44.4	45.8	47.4	34.7	44.0	40.9	6.6	**48.3**	-
dov	BOTH	+21.1	19.1	39.2	38.3	40.2	40.6	19.2	39.5	18.2	8.7	**41.7**	-
ahr$^{\sim\text{hi}}$	neither	+17.8	24.2	31.8	41.9	42.0	**42.8**	32.8	39.0	30.0	36.9	-	-
sus	BOTH	+17.8	11.3	28.3	26.8	29.1	30.3	26.1	29.4	20.7	5.6	**34.6**	-
nqo	BOTH	+17.5	0.2	17.9	17.1	17.7	17.5	17.1	17.9	17.2	1.1	**19.1**	-
alz	BOTH	+15.5	16.9	31.5	30.5	32.4	33.4	25.3	30.6	26.9	8.9	**36.6**	-
lu	BOTH	+13.8	27.6	37.5	39.3	41.4	**42.2**	27.2	37.0	34.8	21.9	-	-
cgg	BOTH	+12.2	32.6	40.5	40.5	44.8	**44.8**	37.3	42.2	37.7	28.3	42.8	-
ks-D.$^{\sim\text{hi}}$	neither	+11.7	14.9	26.6	18.8	26.6	**27.7**	23.8	27.1	21.5	19.2	-	21.1
brx	SMOLSENT	+11.6	24.3	36.0	0.4	35.9	**37.0**	30.8	35.9	36.2	5.2	-	-
mag$^{\sim\text{hi}}$	neither	+8.1	47.0	45.3	55.7	55.1	54.7	47.3	51.8	47.4	48.7	-	**59.4**
ki	BOTH	+7.7	32.6	38.2	38.0	40.3	40.6	35.9	**42.0**	39.1	10.5	-	38.4
aa	BOTH	+7.4	14.2	20.1	20.3	21.6	21.8	18.9	20.6	18.9	5.6	**23.1**	-
ks$^{\sim\text{ur}}$	neither	+7.3	22.1	29.4	0.4	29.4	29.7	28.0	30.4	30.5	26.3	-	**36.7**
nr$^{\sim\text{zu}}$	neither	+7.0	48.9	54.0	51.1	55.9	57.5	48.0	53.6	51.2	45.5	**59.5**	-
doi$^{\sim\text{hi}}$	neither	+6.6	34.3	28.1	**41.4**	40.9	41.3	35.9	39.5	38.2	27.7	40.4	-
sat-L.	SMOLSENT	+6.4	12.8	19.5	15.5	19.2	20.8	**25.3**	22.5	22.7	21.3	22.7	-
mfe$^{\sim\text{fr}}$	neither	+5.3	59.5	65.4	62.6	64.8	66.9	59.6	65.0	59.8	59.5	**67.5**	-
ach	BOTH	+5.2	33.2	43.1	32.5	38.4	39.2	32.4	37.3	35.1	23.8	**43.2**	-
ayl$^{\sim\text{ar}}$	neither	+4.5	47.3	51.7	51.9	51.8	**53.9**	45.3	48.9	46.3	48.6	-	-
st$^{\text{GTr}}$	neither	+4.5	49.9	54.0	55.2	54.4	55.0	49.4	**57.0**	53.1	49.2	49.0	47.2
ber-L.	BOTH	+4.2	26.1	27.9	30.3	30.3	30.9	27.6	32.7	32.1	21.2	**34.7**	-
apd-S.$^{\sim\text{ar}}$	neither	+3.6	42.3	**50.2**	43.3	45.9	47.1	43.2	45.0	42.5	45.6	-	-
ve$^{\sim\text{sn}}$	neither	+3.5	47.9	50.0	48.7	51.4	52.2	50.2	53.1	52.7	43.9	**56.8**	-
kri$^{\sim\text{en}}$	neither	+2.8	31.5	34.2	31.8	34.3	34.7	34.5	33.5	30.7	34.9	**34.9**	-
tiv	BOTH	+2.5	23.8	25.7	25.8	26.3	**26.5**	22.3	24.2	24.5	1.5	25.2	-
gn	SMOLSENT	+2.3	37.4	37.8	30.4	**39.7**	38.4	36.0	38.0	36.4	35.6	38.4	38.5
mos	BOTH	+2.3	18.2	20.9	18.9	20.5	21.1	24.3	**25.0**	20.9	1.3	-	23.8
tum$^{\sim\text{ny}}$	neither	+1.1	40.8	39.5	42.4	41.9	42.8	40.0	42.7	43.8	37.7	**45.4**	36.2
ti$^{\sim\text{am}}$	neither	+0.8	24.2	24.3	24.9	25.0	25.7	25.0	**26.2**	26.1	9.3	26.1	25.5
yo$^{\text{GTr}}$	neither	+0.6	34.6	33.8	35.4	35.2	35.8	29.2	**36.8**	26.7	27.6	21.3	32.6
tn$^{\sim\text{st}}$	neither	+0.2	52.5	50.8	51.9	52.7	53.2	50.1	51.7	53.3	36.8	**55.6**	53.0
ar-M.$^{\sim\text{ar}}$	neither	+0.1	40.1	41.8	39.1	40.2	40.9	40.5	40.8	40.4	41.0	-	**43.0**
am$^{\text{GTr}}$	neither	+0.1	34.0	33.2	33.0	34.1	33.5	31.6	32.6	**35.8**	29.6	34.7	30.3
ig$^{\text{GTr}}$	neither	-0.1	47.2	47.1	47.0	47.1	47.8	43.9	46.2	**47.8**	46.2	47.6	46.6
so$^{\text{GTr}}$	neither	-0.1	49.7	46.2	50.0	49.6	49.1	48.7	49.8	50.3	**50.8**	50.6	48.6
arz$^{\sim\text{ar}}$	neither	-0.1	48.6	46.1	48.9	48.5	47.8	49.6	**50.3**	48.8	49.7	-	49.6
kl	SMOLSENT	-0.3	40.6	39.7	30.2	40.3	41.2	42.2	**43.1**	41.2	41.6	42.9	-
sa	SMOLSENT	-0.3	33.0	32.9	26.7	32.7	33.4	31.8	33.0	32.1	32.0	**35.2**	29.0
ay	SMOLSENT	-0.5	32.7	31.8	24.2	32.2	32.4	33.4	33.2	32.9	30.0	**34.7**	31.7
sn$^{\text{GTr}}$	neither	-0.5	50.5	48.3	50.2	50.0	50.3	46.8	48.8	**51.8**	50.3	49.2	48.2
efi	BOTH	-0.6	14.7	14.5	14.3	14.1	14.2	**15.3**	15.1	15.0	2.2	-	-
ss$^{\sim\text{zu}}$	neither	-0.6	50.2	48.8	48.2	49.6	50.3	49.6	51.2	51.8	46.0	**56.3**	48.1
yue$^{\sim\text{zh}}$	neither	-0.7	26.8	25.8	25.1	26.1	26.1	28.2	28.2	27.5	**31.6**	25.9	22.6
bci	BOTH	-0.7	23.2	22.2	21.8	22.5	22.9	17.1	20.7	27.6	1.0	**29.8**	-
ndc-Z.$^{\sim\text{sn}}$	neither	-1.0	29.2	27.9	28.9	28.2	28.3	27.6	28.0	**29.6**	28.6	29.5	-
es$^{\text{GTr}}$	neither	-1.1	62.4	61.3	51.6	61.3	61.3	-	-	63.0	-	**63.5**	61.8
sat	SMOLSENT	-1.3	32.4	30.9	1.0	31.1	30.8	34.7	36.0	**36.3**	1.8	35.7	-
rw$^{\text{GTr}}$	neither	-1.4	45.2	43.1	43.0	43.8	43.8	43.2	44.0	45.1	44.7	**48.8**	43.4
nd$^{\sim\text{zu}}$	neither	-1.4	43.9	41.5	42.6	42.5	43.2	42.3	42.9	**44.5**	43.6	-	-
sw$^{\text{GTr}}$	neither	-1.5	66.7	64.6	64.2	65.2	64.8	64.0	65.5	**67.2**	66.5	65.3	60.5
mg$^{\text{GTr}}$	neither	-1.9	52.8	48.9	51.6	50.9	51.5	52.4	52.5	**53.3**	52.2	52.6	52.1
qu	SMOLSENT	-1.9	34.7	32.8	30.4	32.8	33.0	35.3	35.1	34.0	22.0	**36.3**	27.9
zu$^{\text{GTr}}$	neither	-2.0	58.3	56.4	55.3	56.3	56.1	54.1	55.5	**58.5**	57.5	57.6	57.6
lus	SMOLSENT	-2.1	42.6	39.7	38.0	40.5	41.4	40.6	41.5	**43.8**	33.5	42.6	39.0
scn$^{\sim\text{it}}$	neither	-2.2	52.4	47.3	50.0	50.2	50.8	49.9	51.4	52.1	49.5	**53.3**	51.0
nso$^{\sim\text{st}}$	neither	-2.3	46.8	42.8	43.6	44.5	44.6	46.9	47.7	**48.1**	46.9	47.6	45.5
xh$^{\text{GTr}}$	neither	-2.3	53.9	49.7	50.7	51.6	51.8	51.6	52.3	53.9	53.7	**54.8**	51.2
ne$^{\text{GTr}}$	neither	-2.7	54.3	51.8	51.7	51.6	52.1	52.4	52.5	52.4	52.7	**54.9**	45.2
pa-A.$^{\sim\text{pa}}$	neither	-3.0	38.1	35.8	0.3	35.1	35.7	41.6	41.2	36.7	37.3	**43.5**	-
aeb$^{\sim\text{ar}}$	neither	-3.3	46.5	41.9	42.3	43.2	43.4	45.9	46.8	47.6	**49.2**	-	43.8
ha$^{\text{GTr}}$	neither	-3.4	**54.5**	50.7	49.9	51.1	51.5	50.9	51.0	54.1	53.9	53.8	53.9

lang	cat	Δ_{FT}	G2F	+sS	+sD	+sB	+sG	Cld	+RAG	G2P	GPT4o	GTr	NLLB
$ts^{\sim zu}$	neither	-3.6	50.6	47.2	46.4	47.0	48.1	49.7	50.1	51.6	49.0	**52.9**	51.3
$rn^{\sim rw}$	neither	-3.9	44.5	39.3	40.5	40.6	40.9	43.1	43.4	**46.2**	44.3	45.4	45.0
af^{GTr}	neither	-4.2	71.9	68.8	67.9	67.7	68.3	71.7	**72.5**	72.1	71.8	71.5	68.6
bo	SMOLSENT	-4.3	41.3	36.7	34.7	37.0	37.3	42.6	42.1	**43.3**	19.8	41.8	36.9
ny^{GTr}	neither	-5.1	55.0	47.5	50.5	49.9	49.7	53.0	53.1	55.3	53.9	**55.8**	50.3
$pcm^{\sim en}$	neither	-6.5	47.9	43.5	39.4	41.4	41.6	51.3	45.7	49.8	**56.0**	-	-
tcy	SMOLDOC	-6.8	34.7	22.0	28.1	27.9	28.8	28.2	29.3	36.7	21.6	**39.1**	-
ktu	BOTH	-9.4	56.6	59.3	40.4	47.2	51.3	45.8	48.4	57.8	22.3	**64.3**	-

Table G.1: Full results (0-shot) For the en→xx direction. Languages in the Intersect subset (supported by all models) are shown first, and then all other languages. The Δ_{FT} compares the base model and the model finetuned on BOTH, to give an idea of how effective the SMOL datasets are for that language. The CAT column indicates which SMOL datasets support this language. The superscript GTr indicates a language supported by Google Translate; a superscript like $\sim$ xx means that this language is closely related to that Google-Translate-supported language.

Abbreviations: This table needed some squishing to fit. Language varieties whose script/region is different from the CLDR default would have the ISO-15924 script code in the BCP-47 code, like MNI-MTEI or BER-LATN; in this table we have abbreviated them to the first letter thereof (MNI-M or BER-L). Similarly, we have abbreviated:
SMOLSENT → SS
SMOLDOC → SD
BOTH → SB
BOTH+GATITOS → SG
GEMINI 2.0-{FLASH, PRO} → G2.0-{F,P}
GOOGLE TRANSLATE → GTR.

H Complete Per-Language details: the Big-SMOL table

A summary of all SMOL language pairs and coarse-grained information about them can be seen in Table H.1. Numbers are given in terms of examples; keep in mind that a single example in SMOLDOC is a document, whereas in SMOLSENT it is a sentence.

Lang. pair	target language name	ISO 15924 Script	Continent	trg.chars	S.DOC	S.SENT
en_yo	Yoruba	Latn	Africa	780k	584	863
en_sw	Swahili	Latn	Africa	699k	584	863
en_ha	Hausa	Latn	Africa	696k	584	863
en_grt-Latn	Garo (Latin script)	Latn	Asia	591k	457	0
en_trp	Kokborok	Latn	Asia	581k	457	0
en_mg	Malagasy	Latn	Africa	580k	391	863
en_xsr-Tibt	Sherpa (Tibetan script)	Tibt	Asia	569k	457	0
en_om	Oromo	Latn	Africa	542k	391	863
en_sd-Deva	Sindhi (Devanagari script)	Deva	Asia	525k	456	0
en_ccp-Latn	Chakma (Latin script)	Latn	Asia	521k	457	0
en_spv	Sambalpuri	Orya	Asia	508k	457	0
en_doi	Dogri	Deva	Asia	503k	454	0
en_xnr	Kangri	Deva	Asia	503k	457	0
en_mjl	Mandeali	Deva	Asia	496k	457	0
en_lif-Limb	Limbu (Limbu script)	Limb	Asia	494k	457	0
en_ne	Nepali	Deva	Asia	494k	456	0
en_kru	Kurukh	Deva	Asia	492k	457	0
en_hoc-Wara	Ho (Warang Chiti script)	Wara	Asia	492k	457	0
en_bra	Braj	Deva	Asia	491k	457	0
en_bns	Bundeli	Deva	Asia	490k	456	0
en_mag	Magahi	Deva	Asia	488k	456	0
en_wbr	Wagdi	Deva	Asia	488k	455	0
en_bfy	Bagheli	Deva	Asia	487k	457	0
en_unr-Deva	Mundari (Devanagari script)	Deva	Asia	485k	457	0
en_mtr	Mewari	Deva	Asia	480k	457	0
en_tcy	Tulu	Knda	Asia	480k	451	0
en_ahr	Ahirani	Deva	Asia	479k	457	0
en_ig	Igbo	Latn	Africa	474k	391	863
en_dhd	Dhundari	Deva	Asia	465k	456	0
en_bfq	Badaga	Taml	Asia	464k	457	0
en_kfy	Kumaoni	Deva	Asia	462k	457	0
en_bgq	Bagri	Deva	Asia	462k	457	0
en_scl	Shina	Arab	Asia	460k	457	0
en_am	Amharic	Ethi	Africa	443k	584	863
en_lep	Lepcha	Lepc	Asia	441k	456	0
en_st	Sesotho	Latn	Africa	412k	260	863
en_sgj	Surgujia	Deva	Asia	395k	356	0
en_so	Somali	Latn	Africa	392k	260	862
en_ny	Chichewa	Latn	Africa	386k	260	863
en_sn	Shona	Latn	Africa	382k	260	863
en_rw	Kinyarwanda	Latn	Africa	378k	260	863
en_zu	Zulu	Latn	Africa	373k	260	863
en_lg	Luganda	Latn	Africa	369k	260	863
en_xh	Xhosa	Latn	Africa	368k	260	863
en_ln	Lingala	Latn	Africa	365k	260	863
en_noe	Nimadi	Deva	Asia	342k	315	0
en_luo	Luo	Latn	Africa	340k	260	863
en_bm	Bambara	Latn	Africa	337k	260	863
en_ak	Twi	Latn	Africa	328k	260	863
en_sjp	Surjapuri	Deva	Asia	327k	299	0
en_wo	Wolof	Latn	Africa	321k	260	863
en_ff	Fulani	Latn	Africa	320k	260	862
sw_ar	Arabic	Arab	Asia	274k	330	0
en_ar-MA	Morrocan Arabic	Arab	Africa	273k	260	863
en_arz	Egyptian Arabic	Arab	Africa	265k	260	863
am_ar	Arabic	Arab	Asia	265k	329	0
en_nso	Sepedi	Latn	Africa	243k	130	863
en_ti	Tigrinya	Ethi	Africa	231k	260	863
en_af	Afrikaans	Latn	Africa	219k	130	863
en_ber-Latn	Tamazight (Latin Script)	Latn	Africa	206k	130	862
en_ber	Tamazight (Tifinagh Script)	Tfng	Africa	206k	130	862
en_ee	Ewe	Latn	Africa	202k	130	863
en_pcm	Nigerian Pidgin	Latn	Africa	195k	130	864
en_yue	Cantonese	Hant	Asia	195k	584	863
en_kri	Krio	Latn	Africa	188k	130	863
en_tn	Tswana	Latn	Africa	182k	66	863
en_ve	Venda	Latn	Africa	167k	66	863
en_bm-Nkoo	NKo	Nkoo	Africa	167k	66	863
en_bem	Bemba (Zambia)	Latn	Africa	166k	66	863
en_ts	Tsonga	Latn	Africa	165k	66	863
en_tum	Tumbuka	Latn	Africa	164k	66	863
en_ss	Swati	Latn	Africa	163k	66	863
en_ktu	Kituba (DRC)	Latn	Africa	162k	66	863
en_nr	South Ndebele	Latn	Africa	159k	66	863
en_fon	Fon	Latn	Africa	157k	66	863
en_ndc-ZW	Ndau	Latn	Africa	156k	66	863
en_kg	Kongo	Latn	Africa	154k	66	863
en_dov	Dombe	Latn	Africa	153k	66	863
en_nd	North Ndebele	Latn	Africa	150k	66	863
en_ki	Kikuyu	Latn	Africa	149k	66	863
en_lu	Kiluba (Luba-Katanga)	Latn	Africa	148k	66	863
en_efi	Efik	Latn	Africa	147k	66	863
en_cgg	Kiga	Latn	Africa	147k	66	863
en_din	Dinka	Latn	Africa	145k	66	863
en_rn	Rundi	Latn	Africa	144k	66	863
en_tiv	Tiv	Latn	Africa	141k	66	863
en_kr	Kanuri	Latn	Africa	139k	66	863

Lang. pair	target language name	ISO 15924 Script	Continent	trg.chars	S.DOC	S.SENT
en_alz	Alur	Latn	Africa	139k	66	863
en_mfe	Mauritian Creole	Latn	Africa	137k	66	863
en_dyu	Dyula	Latn	Africa	136k	66	863
en_ach	Acholi	Latn	Africa	135k	66	863
en_dje	Zarma	Latn	Africa	135k	66	863
en_aa	Afar	Latn	Africa	133k	66	863
en_bci	Baoulé	Latn	Africa	131k	66	863
en_sus	Susu	Latn	Africa	128k	66	863
en_gaa	Ga	Latn	Africa	126k	66	863
en_mos	Mooré	Latn	Africa	125k	66	863
en_aeb	Tunisian Arabic	Arab	Africa	115k	66	862
en_lij	Ligurian	Latn	Europe	114k	25	863
en_apd	Sudanese Arabic	Arab	Africa	112k	66	855
en_ayl	Libyan Arabic	Arab	Africa	109k	66	863
en_scn	Sicilian	Latn	Europe	102k	100	0
sw_zh	Mandarin Chinese	Hans	Asia	101k	330	0
en_kl	Kalaallisut	Latn	Americas	97k	0	863
am_zh	Mandarin Chinese	Hans	Asia	96k	329	0
en_es	Spanish	Latn	Europe	88k	0	863
en_sat	Santali (Ol Chiki script)	Olck	Asia	83k	0	863
en_bo	Tibetan	Tibt	Asia	82k	0	863
en_lus	Mizo	Latn	Asia	82k	0	863
en_gn	Guarani	Latn	Americas	82k	0	863
en_ay	Aymara	Latn	Americas	82k	0	863
en_sat-Latn	Santali (Latin Script)	Latn	Asia	81k	0	863
en_hac	Hawrami	Arab	Asia	77k	0	863
en_glk	Gilaki	Arab	Asia	77k	0	863
en_ckb	Sorani	Arab	Asia	77k	0	863
en_is	Icelandic	Latn	Europe	77k	0	863
en_sa	Sanskrit	Deva	Asia	77k	0	863
en_qu	Quechua	Latn	Americas	74k	0	863
en_brx	Bodo (India)	Deva	Asia	74k	0	863
en_ks	Kashmiri	Arab	Asia	73k	0	863
en_pa-Arab	Lahnda Punjabi (Pakistan)	Arab	Asia	73k	0	863
en_mni-Mtei	Meiteilon (Manipuri)	Mtei	Asia	71k	0	863
en_ks-Deva	Kashmiri (Devanagari script)	Deva	Asia	65k	0	863

Table H.1: Details on all SMOL language pairs, sorted by the total number of characters in the target side (col. 5). The last two columns are the number of examples per language pair; keep in mind that an example for SMOLSENT is a sentence pair but for SMOLDOC is a document/paragraph. Language pairs are only listed in the direction in which they were translated, so no xx→en pairs are present.

I Data sample

I.1 Sample datum from SmolSent


```
{'id': 381,
 'sl': 'en',
 'tl': 'luo',
 'is_src_orig': True,
 'src': 'Rih, a deaf former soldier, plots rebellion while married to a queer,
    teenage god.',
 'trg': 'Rih, mane en jalweny ma Radin, ochano balo ka koni to okendo ng'ano manigi
    kido mar chuech kamare, nyasaye ma en ojana.'
}
```


I.2 Sample datum from SmolDoc


```
{
'id': 'topic_587__weyiwiniwaaotiwenwy',
 'sl': 'en',
 'tl': 'pcm',
 'is_src_orig': True,
 'factuality': 'ok', # this is a story so there is no factual claim that could be
    wrong
 'srcs': ['"What the hell are you doing, you idiot?!"',
        '"Excuse me?"',
        '"You cut me off! You almost made me crash!"',
        '"I'm sorry, I didn't mean to. I was just trying to get around that slow-
           moving truck."',
        '"Well, you could have at least used your turn signal!"',
        '"I did use my turn signal!"',
        '"No, you didn't! You just pulled right out in front of me!"',
        '"I'm telling you, I used my turn signal!"',
```

```
        ’"Whatever. You’re still a terrible driver."’,
        ’"And you’re a jerk!"’,
        ’"At least I know how to drive!"’,
        ’"Oh, yeah? Well, I’m a better writer than you are!"’,
        ’"That’s debatable."’,
        ’"It’s not debatable! I’m Ernest Hemingway!"’,
        ’"Who?"’,
        ’"Ernest Hemingway! The greatest writer of all time!"’,
        ’"Never heard of him."’,
        ’"Well, you’ve heard of me now!"’,
        ’"Yeah, I heard of you."’],
 ’trgs’: [’"Wetin di hell dey do, yu idiot?!"’,
        ’"Ekskuse mi?"’,
        ’"Yu komot mi! Yu almost make mi krash!"’,
        ’"I dey sorry, I nor wont do am. I just dey try get around dat truk wey slow
            ."’,
        ’"Well, yu for don yus yor turn sign!"’,
        ’"I yus mai turn sign!"’,
        ’"No, yu nor turn am! Yu just turn rite in front of mi!"’,
        ’"I dey tell yu, I yus mai turn sign!"’,
        ’"Wateva. Yu still bi one tribol driva."’,
        ’"And yu bi jerk!"’,
        ’"At least I sabi hau to drive!"’,
        ’"Oh, yeah? Well, I bi ogbonge writa pass yu!"’,
        ’"Wi fit dibate dat."’,
        ’"nortin to dibate! I bi Ernest Hemingway!"’,
        ’"Who?"’,
        ’"Ernest Hemingway! De writa of all taim wey grate pass!"’,
        ’"Neva hear am."’,
        ’"Well, yu don hear mi nau!"’,
        ’"Na so, I don hear yu."’]
}
```