Comparing Approaches to Automatic Summarization in Less-Resourced Languages

Chester Palen-Michel and Constantine Lignos
Michtom School of Computer Science, Brandeis University
{cpalenmichel, lignos}@brandeis.edu

Abstract

Automatic text summarization has achieved high performance in higher-resourced languages like English, but comparatively less attention has been given to summarization in less-resourced languages. This work compares a variety of approaches to summarization from zero-shot prompting of LLMs large and small to fine-tuning smaller models like mT5 with and without three data augmentation approaches and multilingual transfer. We also explore an LLM translation pipeline approach, translating from the source language to English, summarizing and translating back. Evaluating with five different metrics, we find that there is variation across LLMs in their performance at similar model sizes, that our multilingual fine-tuned mT5 baseline outperforms most other approaches including zero-shot LLM performance for most metrics, and that LLM as judge may be unreliable on less-resourced languages.

Keywords: Summarization, less-resourced languages, low resource, multilingual, evaluation

1. Introduction

Automatic text summarization in higher-resourced languages like English has achieved high scores in automated metrics (Al-Sabahi et al., 2018; Liu et al., 2022; Zhang et al., 2020a). However, for many less-resourced languages, the task remains challenging. While there are datasets that cover multilingual summarization in less-resourced languages (Giannakopoulos et al., 2015, 2017; Palen-Michel and Lignos, 2023; Hasan et al., 2021), these datasets often still have relatively few examples compared to their higher-resourced counterparts.

To better understand which approaches work best with less-resourced languages, we conduct a comparative study of a variety of approaches to automatic summarization. Specifically, we compare zero-shot prompting with three smaller-scale LLMs (Mixtral 8x7B, Llama 3 8B, Aya 101). Given that LLMs' pretraining data tends to be dominated by higher-resourced languages, we also experiment with fine-tuning smaller mT5 in a variety of settings. We fine-tune mT5 per-individual language and with all available language data combined for multilingual transfer as baselines. Multilingual transfer has proven to be a useful strategy for less-resourced languages (Wang et al., 2021); however, other works have shown that multilingual models have limits and given enough data, fully monolingual models can perform better (Virtanen et al., 2019; Tanvir et al., 2021).

We further explore fine-tuning mT5 with synthetic data generated by leveraging extra Wikipedia data using three different approaches shown in Figure 1. While prior work has focused on comparing multilingual summarization models which take advantage of multilingual transfer with fine-tuning on a single language (Palen-Michel and Lignos, 2023; Hasan et al., 2021) or the use of synthetic data for a single language only (Parida and Motlicek, 2019), this work compares the performance of multilingual pre-trained models fine-tuned using data for a single language with fine-tuning that uses the combination of synthetic and real data from all languages.

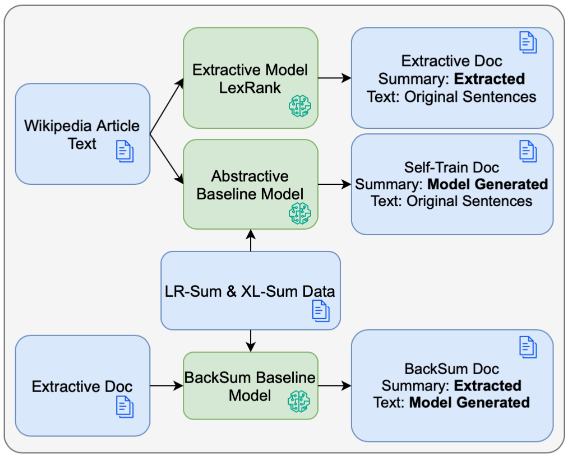

Figure 1: Methodology for generating additional training examples from Wikipedia articles

We then conduct additional experiments with three larger LLMs (Gemma 3 27B, Llama 3.3 70B, Aya Expanse 32B). We also try a pipeline approach with these larger LLMs, translating to English, summarizing in English, and translating back to the target language. We primarily evaluate with ROUGE scores and BERTScore, but with increased attention on LLM as judge evaluation (Fu et al., 2024; Kim et al., 2024; Pombal et al., 2025) we also conduct some experiments with M-Prometheus, an open multilingual LLM trained for evaluation.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3402–3422
11–16 May 2026.

Our contributions are the following:

1. A comparison of various approaches to summarization in less-resourced languages including: fine-tuning mT5 in a per-individual language and multilingual setting with and without three data augmentation strategies, zero-shot LLM inference with smaller LLMs and comparatively larger LLMs, and a pipeline approach translating from the original language to English then summarizing and translating back to the target language with LLMs.

2. A comparison of popular summarization evaluation approaches including ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, and reference-free LLM as judge using M-Prometheus, which demonstrates that different evaluation methods yield somewhat different views of which models perform best.

3. An analysis of the English content produced by LLMs when producing summaries for less-resourced languages.

We conclude that there is some variation across LLMs in their performance at similar model sizes and that zero-shot LLM performance significantly lags the multilingual baseline for most metrics. We also find that data augmentation for individual language fine-tuning for mT5 showed improvement over baselines, but does not outperform fine-tuning mT5 in a multilingual transfer scenario.

Because improved methods for evaluating summarization continue to be developed and explored and because we welcome participatory research (Caselli et al., 2021), where speakers of languages have the opportunity to collaborate on the design of NLP tools, we release all candidate summaries generated as part of this work for future summarization evaluation research.[1]

2. Background

The two main approaches to automatic summarization have been extractive and abstractive methods. Extractive models select important sentences in the source article to use as summaries (Luhn, 1958; Radev et al., 2001; Christian et al., 2016). Abstractive models typically cast the problem as a sequence-to-sequence problem and apply a neural language model (Rush et al., 2015; See et al., 2017; Hsu et al., 2018; Zhang et al., 2020a). Abstractive neural models typically require larger amounts of training data to train.

2.1. Related Work

Prior work on multilingual summarization has largely focused on newswire text from higher resourced languages or covers more languages but with very limited data (Scialom et al., 2020; Giannakopoulos et al., 2015, 2017). Some of the languages in our study have little to no work in summarization, like Armenian (Avetisyan and Broneske, 2023). Others, like Georgian, have been studied in cross-lingual summarization (Turcan et al., 2022), but appear to be underexplored for monolingual summarization. There is a recent effort to create a Kurdish summarization dataset (Badawi, 2023). The Global Voices summarization dataset (Nguyen and Daumé III, 2019) contains some examples of Macedonian. MassiveSumm (Varab and Schluter, 2021) has greater coverage of languages, but is automatically created, recall-oriented, and has complex redistribution requirements, so we did not make use of it in this work.

Large language models (LLMs) have been shown to perform comparably to human summaries for English (Zhang et al., 2024). However, using LLMs for less-resourced languages is less well-studied. We select three reasonably well performing smaller LLMs and three more recent larger LLMs largely because of their widespread adoption and benchmark performance.[2]

Regarding data augmentation, the most similar prior work to our approaches includes Parida and Motlicek (2019), which used a similar approach to what we refer to as "back-summarization," but they apply it only to German. The approach is also similar to the concept of back-translation (Sennrich et al., 2016) for machine translation where inference is done in the opposite direction to create additional synthetic labeled data. Another approach we use, self-training, has been used in other previous work with other tasks and datasets (Du et al., 2021; Karamanolakis et al., 2021; Meng et al., 2021).

2.2. Summarization Evaluation

Evaluating the quality of a summary is inherently difficult due to there being multiple ways to express similar content and the subjectiveness of the task. Summarization was historically scored using ROUGE-1, ROUGE-2, and ROUGE-L metrics (Lin, 2004). Model-based metrics such as BERTScore (Zhang et al., 2020b) are also used, although works such as Sun et al. (2022) have found issues in bias with BERTScore. There have been many proposed methods of evaluating summarization systems (Darrin et al., 2024; Vasilyev et al., 2020; Fu et al., 2024).

Human evaluation is often conducted, but it is time consuming, expensive, and cognitively demanding (Iskender et al., 2021; Lin, 2004) and can also be inconsistent, with some work showing inter-annotator discrepancy (Kryscinski et al., 2019). The challenge of conducting human evaluation is even more prominent in work in multiple

[1] https://github.com/cpalenmichel/model_inferred_summaries_less-resourced_langauges

[2] More details on LLM selection are in Appendix I.

languages, particularly less-resourced languages, where human judges may be difficult to recruit.

3. Datasets

For experiments, we use LR-Sum (Palen-Michel and Lignos, 2023) and XL-Sum (Hasan et al., 2021). LR-Sum contains summarization data for 40 languages, many of which are also less-resourced. LR-Sum is built using the description field from the Multilingual Open Text corpus (Palen-Michel et al., 2022) and is similar in approach and content to XL-Sum (Hasan et al., 2021), which contains BBC news articles in 44 languages. For this study, we focused on a small set of languages from LR-Sum and XL-Sum which had the fewest number of examples in the corpus, while also choosing languages for linguistic diversity. The goal was to select a relatively diverse set of less-resourced languages.

As seen in Table 15 in Appendix H, many of the languages we work with have fewer than 1,000 summarization examples, which presents a challenge for neural abstractive summarization systems, which typically require large amounts of training data. With the exception of Burmese and Pashto, the languages we worked with did not have overlap between XL-Sum and LR-Sum. While there is little summarization training data for these languages, there is unlabeled text data available in Wikipedia. However, many Wikipedia articles for less-resourced languages are quite short in length.

We used Segment Any Text (Frohmann et al., 2024) to perform sentence segmentation of the Wikipedia articles to filter out documents which have fewer than 5 sentences. Wikipedia articles that have fewer than 5 sentences tend to be incomplete, lists, or definitions, and do not appear to be useful as additional summarization data. After filtering out Wikipedia articles shorter than five sentences long, for many of the languages there is substantially less data available than may appear in raw counts of Wikipedia articles. Specifically, Khmer surprisingly has nearly the same amount of training examples available in LR-Sum as there are suitable Wikipedia articles.

4. Methodology

4.1. Fine-tuning mT5 with Data Augmentation

We explore three approaches for using Wikipedia articles as extra synthetic training data for summarization. The summarization task can be considered document and summary pairs, $\langle D, S \rangle$, where documents consist of sentences and summaries consist of sentences $D = \{d_1, d_2, d_3 ...\}, S = \{s_1, s_2, s_3 ...\}$. Generating augmented data then consists of creating novel D' and/or S' as additional training pairs. In this case we apply the augmentation strategies to Wikipedia, which does not have existing summaries. Portions of example summaries and documents created from each strategy are shown in Table 10 in Appendix E.

The approach to creating these extra synthetic training documents is shown in Figure 1. We train a baseline multilingual sequence-to-sequence abstractive model using mT5 (Xue et al., 2021). For experiments, we do this on a per-language basis and also in a multilingual way with upsampling to ensure a balance of different languages is seen.

We use the same set of hyperparameters across all experiments. All models used mT5-base as the underlying pre-trained model. All models were trained for 3 epochs with 100 warmup steps. We used a label smoothing factor of 0.1, a beam size of 4, weight decay of 0.01, a max target length of 512, a max source length of 1024, an effective batch size of 32 and a learning rate of 5e-4. Hyperparameters were chosen largely following those suggested in XL-Sum (Hasan et al., 2021) and LR-Sum (Palen-Michel and Lignos, 2023). For upsampling of multilingual fine-tuning, we use an upsampling factor of .5 following Hasan et al. (2021) and conduct the multilingual training using the codebase from Hasan et al. (2021).[3]

Extractive Training For augmented data, first we use the LexRank (Erkan and Radev, 2004) extractive summarization algorithm as implemented in the lexrank Python package[4] to create summaries. We chose LexRank since it was reported as the highest performing extractive method by Palen-Michel and Lignos (2023). LexRank was set to select two sentences since most of the newswire summaries in LR-Sum and XL-Sum are roughly two sentences long. We then directly use these extracted summaries as target summaries alongside the original Wikipedia text. The extractive summary is composed of sentences chosen from the document so the new example is $\langle D, S' \rangle$ where $S' = \{d_n, d_m\}$.

Self-Training Second, after fine-tuning a multilingual abstractive sequence-to-sequence model using mT5 as the underlying model, we use it to generate summaries on Wikipedia articles. These generated summaries and the original Wikipedia text are used for the self-training experiment. Again the summary is new $\langle D, S' \rangle$ but here the sentences are model generated $S' = \{x_1, x_2, ...\}$.

Back-Summarization Third, we train a model that when given a summary generates the article

[3] https://github.com/csebuetnlp/xl-sum/tree/master/seq2seq

[4] https://github.com/crabcamp/lexrank

associated with the summary. We apply this back-summarization model to the LexRank extracted summaries of Wikipedia articles, $S' = \{d_n, d_m\}$, to get a generated document $D' = \{y_1, y_2, ...\}$ and use the extracted summary as the summary. For Back-Summarization, the summary and document are both automatically generated, $\langle D', S' \rangle$.

Individual Models with Augmented Data For each of the three data augmentation approaches, we train on a concatenation of the original training dataset with up to 6k of the synthetic training examples. We refer to this as "individual" because models are trained on individual languages (i.e. they are not multilingual models). We choose to use only a subset of available Wikipedia articles in part to have a better balance of synthetic data and real data and also in part for faster experiments due to resource constraints. For individual model experiments, we focus on just the smallest 7 languages from LR-Sum: Sorani Kurdish (ckb), Haitian Creole (hat), Armenian (hye), Georgian (kat), Khmer (khm), Kurmanji Kurdish (kmr), and Macedonian (mkd).

Multilingual Models with Augmented Data We fine-tune three versions of mT5 for each of the data augmentation approaches with a combination of all the XL-Sum and LR-Sum training data with the addition of the augmented data. When training the multilingual model, upsampling is done by language. This increases the diversity of the examples seen during training for less-resourced languages, but not their frequency. We focus on 18 languages for multilingual model experiments which represent the smaller languages of LR-Sum and XL-Sum.

4.2. LLM Experiments

Smaller LLMs We run inference with a set of comparatively smaller LLMs, Mixtral 8x7B (Jiang et al., 2024), Llama 3 (8B) (Dubey et al., 2024), and Aya 101 (12.9B) (Üstün et al., 2024). We use Ollama for inference for Mixtral and Llama 3.

As Aya 101 was not supported by Ollama, we used Hugging Face Transformers (Wolf et al., 2020) and BitsAndBytes[5] to load it in 8-bit quantization.[6] We set no_repeat_ngrams to 3, truncation to True, and max_length to 256. Prompts are described in the Appendix D. Because we found that the LLMs generated a significant amount of English, we use CLD3 (Salcianu et al., 2016) to detect the mean proportion of English summaries generated by the

[5]https://github.com/bitsandbytes-foundation/bitsandbytes

[6]Models run with Ollama also used quantization by default.

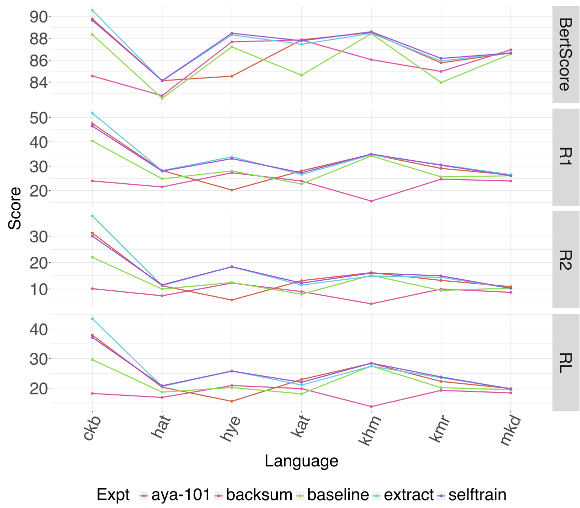

Figure 2: Scores for augmentation approaches with individual-language fine-tuning using mT5. Aya 101 performance is added for comparison with LLM performance.

LLMs. We evaluated CLD3's performance on random Unicode characters to ensure that it did not automatically misclassify unfamiliar characters as English.

5. Results

Individual Models As shown in Figure 2,[7] and in more detail in Appendix A Table 5, all languages have higher ROUGE scores with the inclusion of additional synthetic training data than fine-tuning mT5 with just the training data of an individual language. Sorani Kurdish (ckb), Kurmanji Kurdish (kmr), and Armenian (hye) in particular have the most substantial increases in ROUGE scores from the baseline. Armenian using the backsum approach is the only language that has a worse score when using augmented data.

Of the different strategies for making use of the additional Wikipedia articles, none stands out as being particularly stronger than the others across all languages. Self-training seems to have better scores for ROUGE-2 and ROUGE-L when it outperforms the other methods, but the difference tends to be small with the exception of Kurmanji Kurdish. Khmer (khm) had the smallest amount of augmented data since the Khmer Wikipedia articles were quite small and had a relatively small increase in scores.

[7]While conventionally line plots are not used when the x-axis is categorical, we provide a single line per method in addition to points to make cross-lingual trends easier to visualize.

	Best Aug. Indiv.			Aya 101			Multilingual Baseline		
Lang.	R1	R2	RL	R1	R2	RL	R1	R2	RL
ckb	**52.6**	**38.1**	**44.4**	23.9	10.1	18.2	45.2	27.0	34.7
hat	28.7	11.2	20.9	21.5	7.4	16.9	**29.0**	**12.1**	**21.6**
hye	34.5	18.3	25.9	27.3	12.2	20.9	**34.9**	**19.2**	**26.5**
kat	28.1	13.1	22.9	23.9	9.0	19.8	**29.2**	**14.4**	**24.2**
khm	35.1	16.0	28.5	15.6	4.3	13.8	**38.2**	**19.0**	**31.3**
kmr	30.5	15.0	23.9	24.6	10.0	19.3	**33.7**	**18.0**	**27.0**
mkd	27.0	10.1	20.4	23.9	8.7	18.4	**27.3**	**11.4**	**20.9**

Table 1: Comparison of the best augmented individual model with Aya 101 and the baseline multilingual model

Multilingual Models Although augmented training of individual models performs better than training individual models without augmented data, the scores of individual models are still lower than multilingual fine-tuning of mT5 with a combination of all LR-Sum and XL-Sum training data as seen in Table 1. However, these individual models do outperform the best performing smaller LLM, Aya 101. Hasan et al. (2021) and Palen-Michel and Lignos (2023) found multilingual models to generally perform better than individually trained models. We compare the performance of the best augmented training approach with the reported multilingual model scores from LR-Sum.

The results of including augmented data in the fine-tuning of the multilingual model, shown in Figure 3 and in more detail in Table 6 in Appendix A, do not demonstrate a clear improvement over the baseline. For Amharic, Sorani, Georgian, Pashto, and Somali, the Back-Summarization approach performs somewhat better. Self-training tends to have the same or lower ROUGE Scores for all languages and test set varieties tested. Compared to the best performing LLM with prompting, both the multilingual fine-tunings of mT5 with and without augmentation have higher scores.

Smaller LLMs Despite the impressive summarization capabilities of LLMs as discussed in Zhang et al. (2024), the LLMs we explored here performed underwhelmingly. As seen in Table 2, Mixtral performed the worst while Llama3 tended to perform somewhat better than Mixtral, and Aya 101 performed best of the LLMs examined. [8] We observed that Mixtral and Llama3 tend to produce a significant amount of English. The proportion of English appears to be highest when the target language has a non-Roman script. For example, for Amharic, Georgian, Khmer, Lao, and Burmese, these models all produced English with mean proportions over 30%.

We manually reviewed the responses of LLMs when they generated English output despite being given non-English articles and being prompted to respond in the target language. Sometimes the English response is an apology message explaining that it cannot perform the task; other times it is a plausible English summary of the target article. In some cases the model asks to see the text despite having already been shown the text, and sometimes the model begins in the target language but eventually switches to English.

6. Additional Experiments

This work was developed over years by a research group with relatively limited computational resources. During development of this work, we gained access to GPUs with more memory, new LLMs were released, and better LLM as judge approaches were developed. These factors led us to perform an additional set of experiments that builds upon the results in the previous section but explores using larger LLMs and using LLMs in additional ways, namely for evaluation and for a translate-summarize-translate pipeline.[9] Now, having compared mT5 fine-tuning methods with smaller LLMs and identified that the multilingual baseline mT5 model and Aya 101 have the best ROUGE and BERTScore performance, we turn to using larger LLMs and evaluating with M-Prometheus.

6.1. Methods

Larger LLMs We additionally run inference using three moderately sized LLMs, Gemma 3 27B (Gemma Team et al., 2025), Aya Expanse 32B (Dang et al., 2024), and Llama 3.3 70B. For these runs we use VLLM for inference. The prompts we used request that the model generate a summary in two sentences. A two sentence summary was requested to be similar to the reference summaries for the datasets used given that most of the reference summaries are roughly two sentences long. The specific prompts used are detailed in Appendix D.

Translate-Summarize-Translate (TST) with LLMs We experiment with a pipeline approach to generating summaries in less-resourced languages where an LLM is first prompted to translate the article into English, and then summarize in two sentences the English-translated article. Finally, the model is prompted to translate back to the target language. We run experiments using Gemma 3 27B and Aya Expanse 32B since these two LLMs had better performance in preliminary

[8] Rouge-1 and Rouge-2 are included in Appendix A in Table 7.

[9] We separate this work into a different section, as these experiments were able to be run with more languages and with additional evaluation metrics.

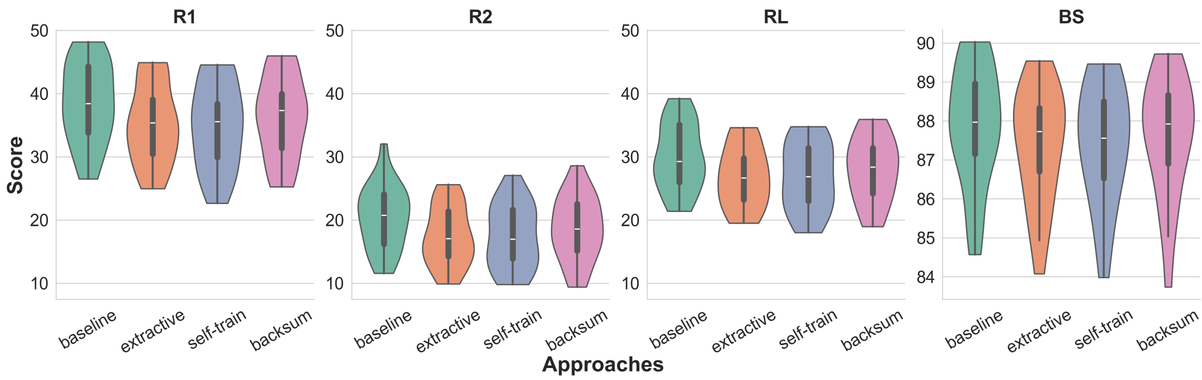

Figure 3: Distribution of scores for augmentation approaches with multilingual fine-tuning across all languages and datasets.

		Mixtral			LLama 3			Aya 101		
Dataset	Lang.	% Eng.	RL	BERTScore	% Eng.	RL	BERTScore	% Eng.	RL	BERTScore
LR-Sum	amh	64.6	6.96$^{\pm0.02}$	80.15$^{\pm0.01}$	38.6	12.49$^{\pm0.02}$	81.83$^{\pm0.01}$	0.0	**14.63**$^{\pm0.03}$	**83.29**$^{\pm0.01}$
LR-Sum	ckb	55.8	6.28$^{\pm0.01}$	78.26$^{\pm0.01}$	16.8	**29.79**$^{\pm0.04}$	**86.93**$^{\pm0.01}$	0.0	18.20$^{\pm0.03}$	84.56$^{\pm0.01}$
LR-Sum	hat	2.6	13.36$^{\pm0.01}$	82.55$^{\pm0.00}$	0.2	**18.65**$^{\pm0.02}$	**83.87**$^{\pm0.00}$	0.6	16.86$^{\pm0.02}$	82.77$^{\pm0.01}$
LR-Sum	hye	60.7	7.17$^{\pm0.01}$	81.16$^{\pm0.02}$	4.0	14.09$^{\pm0.01}$	85.21$^{\pm0.00}$	0.0	**20.89**$^{\pm0.02}$	**87.67**$^{\pm0.01}$
LR-Sum	kat	75.4	4.83$^{\pm0.01}$	81.07$^{\pm0.01}$	45.0	10.69$^{\pm0.01}$	83.18$^{\pm0.01}$	0.0	**19.85**$^{\pm0.02}$	**87.84**$^{\pm0.01}$
LR-Sum	khm	76.2	4.77$^{\pm0.01}$	80.29$^{\pm0.01}$	37.9	8.77$^{\pm0.01}$	81.80$^{\pm0.00}$	0.1	**13.77**$^{\pm0.02}$	**86.05**$^{\pm0.00}$
LR-Sum	kmr	5.4	12.53$^{\pm0.01}$	83.64$^{\pm0.00}$	0.0	18.76$^{\pm0.02}$	**85.28**$^{\pm0.00}$	0.2	**19.25**$^{\pm0.02}$	84.97$^{\pm0.01}$
LR-Sum	lao	63.9	4.85$^{\pm0.01}$	79.61$^{\pm0.01}$	49.2	10.07$^{\pm0.01}$	81.64$^{\pm0.00}$	0.2	**20.86**$^{\pm0.02}$	**87.40**$^{\pm0.00}$
LR-Sum	mkd	6.6	8.24$^{\pm0.01}$	83.52$^{\pm0.01}$	0.3	14.85$^{\pm0.01}$	86.18$^{\pm0.00}$	0.0	**18.40**$^{\pm0.02}$	**86.95**$^{\pm0.01}$
LR-Sum	mya	58.6	4.09$^{\pm0.00}$	79.19$^{\pm0.01}$	54.0	8.88$^{\pm0.01}$	82.00$^{\pm0.00}$	0.1	**13.79**$^{\pm0.01}$	**85.67**$^{\pm0.00}$
LR-Sum	pus	47.4	9.40$^{\pm0.01}$	80.78$^{\pm0.01}$	3.2	22.57$^{\pm0.01}$	85.85$^{\pm0.00}$	0.0	**28.69**$^{\pm0.01}$	**87.25**$^{\pm0.00}$
LR-Sum	sna	3.0	11.73$^{\pm0.01}$	82.13$^{\pm0.00}$	0.0	15.70$^{\pm0.02}$	83.30$^{\pm0.00}$	0.2	**15.74**$^{\pm0.02}$	**83.48**$^{\pm0.01}$
LR-Sum	som	13.9	12.54$^{\pm0.03}$	83.38$^{\pm0.01}$	1.8	**22.05**$^{\pm0.03}$	**86.00**$^{\pm0.01}$	0.0	20.98$^{\pm0.04}$	85.60$^{\pm0.01}$
XL-Sum	amh	67.4	5.62$^{\pm0.01}$	79.54$^{\pm0.01}$	43.7	11.90$^{\pm0.01}$	82.01$^{\pm0.00}$	0.0	**20.88**$^{\pm0.02}$	**85.21**$^{\pm0.01}$
XL-Sum	gla	2.0	13.58$^{\pm0.01}$	83.60$^{\pm0.00}$	0.2	18.54$^{\pm0.01}$	84.67$^{\pm0.00}$	0.0	**26.79**$^{\pm0.02}$	**86.56**$^{\pm0.00}$
XL-Sum	ibo	13.8	13.24$^{\pm0.01}$	82.47$^{\pm0.01}$	0.0	19.55$^{\pm0.01}$	85.09$^{\pm0.00}$	0.0	**29.39**$^{\pm0.02}$	**87.33**$^{\pm0.00}$
XL-Sum	mya	53.3	8.86$^{\pm0.01}$	81.45$^{\pm0.01}$	38.7	14.92$^{\pm0.01}$	83.18$^{\pm0.00}$	0.0	**27.09**$^{\pm0.02}$	**88.05**$^{\pm0.01}$
XL-Sum	orm	16.9	10.38$^{\pm0.01}$	81.23$^{\pm0.00}$	0.7	13.91$^{\pm0.01}$	82.47$^{\pm0.01}$	1.6	**19.48**$^{\pm0.01}$	**83.28**$^{\pm0.00}$
XL-Sum	pus	40.6	10.15$^{\pm0.01}$	81.40$^{\pm0.01}$	4.0	21.37$^{\pm0.01}$	85.63$^{\pm0.00}$	0.0	**30.55**$^{\pm0.01}$	**87.86**$^{\pm0.00}$
XL-Sum	sin	59.4	5.10$^{\pm0.01}$	78.82$^{\pm0.03}$	15.7	14.60$^{\pm0.01}$	84.73$^{\pm0.00}$	0.0	**22.87**$^{\pm0.03}$	**88.16**$^{\pm0.01}$
XL-Sum	som	17.2	11.04$^{\pm0.01}$	82.58$^{\pm0.00}$	0.8	17.61$^{\pm0.01}$	84.74$^{\pm0.00}$	0.0	**22.10**$^{\pm0.02}$	**85.92**$^{\pm0.00}$
XL-Sum	yor	17.3	13.79$^{\pm0.01}$	82.85$^{\pm0.00}$	0.9	19.47$^{\pm0.01}$	85.56$^{\pm0.00}$	0.1	**26.92**$^{\pm0.02}$	**86.91**$^{\pm0.00}$

Table 2: LLM performance across languages measured in Rouge-L and BERTScore along with the percentage of generated summary containing English found by language id. All results are the mean of 500 bootstrap samples with standard error reported

TST experiments. The specific prompts used are detailed in Appendix D.

6.2. Results

Larger LLMs When comparing relatively larger LLMs of similar sizes, shown in Table 3, Aya Expanse and Llama 3.3 tend to outperform Gemma, but neither completely outperforms the other for all languages. All smaller LLMs performed worse than larger LLMs with the exception of Llama 3's performance on Sorani Kurdish and Aya 101, which has better performance in some languages than larger LLMs. We see in Figure 4 that Aya Expanse tends to have scores below the multilingual fine-tuned mT5 baseline model, with the exception of M-Prometheus scores which tend to favor Aya Expanse 32B. Aya 101 and Aya Expanse are comparable in most cases, but it varies based on metric and language.

Translate-Summarize-Translate (TST) Overall, the TST pipeline results (Table 4) demonstrated worse performance across all metrics than the highest performing simple zero-shot prompting.[10] By examining the output, we observed that although smaller LLMs often produced output which was entirely an English response, larger LLMs sometimes produced additional English commentary in addition to a summary in the target language. We postprocessed the LLM output to include only the target summary by filtering out extra English com-

[10]Standard error included in Appendix A Table 8.

		Aya Expanse 32B		Gemma 3		Llama 3.3	
Dataset	Lang.	RL	BERTScore	RL	BERTScore	RL	BERTScore
LR-Sum	amh	$20.54^{\pm0.03}$	$85.14^{\pm0.01}$	$16.88^{\pm0.03}$	$84.03^{\pm0.02}$	**21.83**$^{\pm0.03}$	**85.68**$^{\pm0.01}$
LR-Sum	ckb	**21.94**$^{\pm0.02}$	**85.18**$^{\pm0.00}$	$15.81^{\pm0.01}$	$83.48^{\pm0.01}$	$21.26^{\pm0.02}$	$84.69^{\pm0.01}$
LR-Sum	hat	**20.48**$^{\pm0.02}$	**84.25**$^{\pm0.00}$	$17.02^{\pm0.01}$	$83.39^{\pm0.00}$	$20.09^{\pm0.02}$	$84.12^{\pm0.00}$
LR-Sum	hye	$19.76^{\pm0.01}$	$86.91^{\pm0.00}$	$19.67^{\pm0.02}$	$87.70^{\pm0.00}$	**20.48**$^{\pm0.02}$	**87.76**$^{\pm0.00}$
LR-Sum	kat	$17.75^{\pm0.02}$	$86.99^{\pm0.00}$	$17.05^{\pm0.01}$	$86.16^{\pm0.02}$	**20.12**$^{\pm0.02}$	**87.63**$^{\pm0.00}$
LR-Sum	khm	$13.03^{\pm0.01}$	$85.33^{\pm0.00}$	$13.10^{\pm0.01}$	$85.50^{\pm0.01}$	**15.02**$^{\pm0.01}$	**85.91**$^{\pm0.00}$
LR-Sum	kmr	$15.45^{\pm0.02}$	$82.80^{\pm0.01}$	$15.54^{\pm0.01}$	$84.45^{\pm0.01}$	**19.70**$^{\pm0.02}$	**85.88**$^{\pm0.00}$
LR-Sum	lao	$21.36^{\pm0.01}$	$87.06^{\pm0.00}$	$12.46^{\pm0.01}$	$84.69^{\pm0.00}$	**23.52**$^{\pm0.01}$	**87.78**$^{\pm0.00}$
LR-Sum	mkd	$16.07^{\pm0.01}$	$85.87^{\pm0.00}$	$14.94^{\pm0.01}$	$86.24^{\pm0.01}$	**16.35**$^{\pm0.01}$	**86.70**$^{\pm0.00}$
LR-Sum	mya	**23.88**$^{\pm0.02}$	**87.49**$^{\pm0.00}$	$11.63^{\pm0.01}$	$82.75^{\pm0.02}$	$18.08^{\pm0.02}$	$87.16^{\pm0.00}$
LR-Sum	pus	$24.82^{\pm0.01}$	$86.13^{\pm0.00}$	$21.05^{\pm0.01}$	$85.67^{\pm0.00}$	**25.03**$^{\pm0.01}$	**87.11**$^{\pm0.00}$
LR-Sum	sna	$16.58^{\pm0.02}$	$83.89^{\pm0.00}$	$14.73^{\pm0.01}$	$83.26^{\pm0.00}$	**16.72**$^{\pm0.02}$	**83.48**$^{\pm0.00}$
LR-Sum	som	**23.83**$^{\pm0.03}$	$86.71^{\pm0.01}$	$19.62^{\pm0.03}$	$85.85^{\pm0.01}$	$23.35^{\pm0.03}$	**86.88**$^{\pm0.01}$

Table 3: Performance of larger LLMs with simple prompting for summaries in 2 sentences

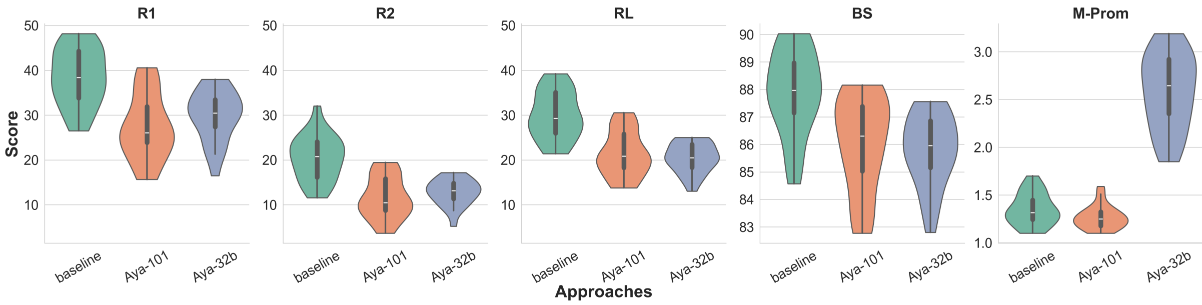

Figure 4: Comparison of Aya 101, Aya Expanse, and Multilingual Transfer MT5 finetuned baseline

	Aya Exp. TST		Gemma TST		Best Non-TST	
Lang	RL	BS	RL	BS	RL	BS
amh	10.79	80.86	15.29	83.98	**21.83**	**85.68**
ckb	13.62	82.95	12.44	82.18	**21.94**	**85.18**
hat	12.98	82.62	14.57	83.15	**20.48**	**84.25**
hye	11.68	83.64	14.84	86.49	**20.48**	**87.76**
kat	11.86	84.88	12.61	85.96	**20.12**	**87.63**
khm	3.49	79.13	10.83	84.96	**15.02**	**85.91**
kmr	9.77	82.07	11.69	84.18	**19.70**	**85.88**
lao	6.41	81.26	13.25	85.81	**23.52**	**87.78**
mkd	12.00	84.27	13.15	86.00	**16.35**	**86.70**
mya	8.10	82.15	8.80	84.88	**23.88**	**87.49**
pus	14.83	82.33	17.65	85.37	**25.03**	**87.11**
sna	8.46	81.66	9.66	81.98	**16.72**	**83.48**
som	13.44	83.47	15.61	85.19	**23.83**	**86.71**

Table 4: Results for Translate-Summarize-Translate Pipeline approach with ROUGE-L (RL) and BERTScore (BS).

mentary using a combination of pattern matching and language ID. This English commentary is broken down into categories in Appendix A Figure 6. Gemma produced the most additional commentary, nearly always mentioning a translation when using the TST approach and sometimes providing a translation in the simple summarization prompt approach. Languages like Pashto, Sorani Kurdish, and Amharic show higher rates of the LLM providing unprompted transliteration and pronunciation.

LLM as Judge We use M-Prometheus (Pombal et al., 2025) for LLM as judge evaluation of summaries. M-Prometheus is a multilingual version of Prometheus (Kim et al., 2024), which is an open-sourced LLM specifically tuned for text generation evaluation. M-Prometheus is based on Qwen 2.5 (Yang et al., 2024) which claims support for roughly 30 languages. M-Prometheus is tested on roughly 30 languages, most of which are higher resourced. The criteria and prompts provided for evaluation are described in the Appendix Section D.2.

We find M-Prometheus scores favor larger LLM output such as that of Aya Expanse 32B (as seen in Figure 4). We examined the distribution of summaries' ROUGE-L scores from Llama 3.3, comparing M-Prometheus with BERTScore and ROUGE-1. We show the distribution of summaries' ROUGE-L scores from Llama 3.3 compared with M-Prometheus scores in Figure 5, and BERT Score and ROUGE-1 in Appendix G, Figures 7 and 8.

We observed that M-Prometheus scores generally do not appear to increase with other scoring metrics and that there is a larger variance for less-resourced languages, while M-Prometheus

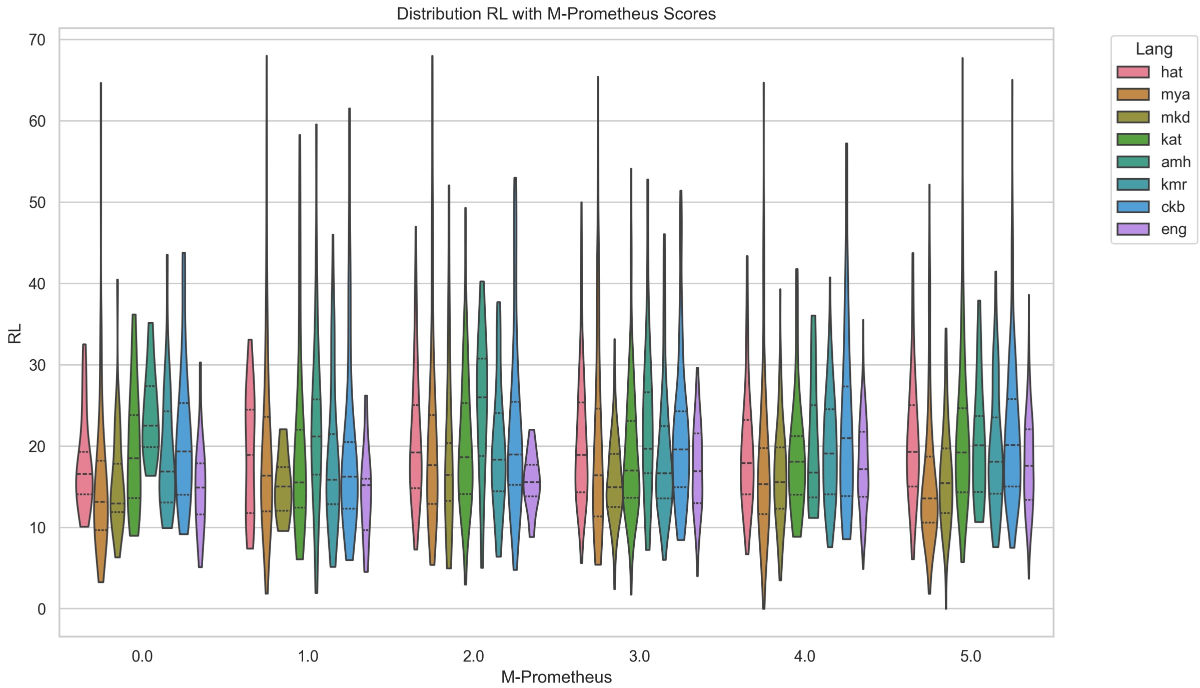

Figure 5: Distribution of summary scores for Llama 3.3 comparing scoring methods RL and M-Prometheus

scores on English summaries appear to have a better alignment with reference-based metrics and less variance. This suggests that M-Prometheus may be more reliable at evaluating on higher-resource than less-resourced languages. There is some slight pattern of increase between M-Prometheus scores 1-3, but the distribution of other scoring metrics appears to be slightly lower generally for M-Prometheus scores of 4 and 5.

7. Discussion

Which approaches work best overall for summarization in less-resourced languages? The TST pipeline approach did not outperform zero-shot LLM prompting. We compare the best zero-shot LLM of smaller models, larger models with the best performing fine-tuning of mT5 in Figure 4. Multilingual T5 without augmentation out-performed LLMs for all metrics except M-Prometheus.

How well do different metrics measure summarization performance? While our aim is not to conduct a comprehensive assessment of summarization metrics, we do observe some differences in metrics. BERTScore, RL, and R1 can show some nuances in performance when comparing languages, but general trends appear to be mostly consistent across these metrics. Meanwhile, reference-free LLM evaluation with M-Prometheus tends to favor the output of larger LLMs. It is possible that M-Prometheus is better equipped to evaluate nuances in summarization that reference-based summarization misses. However, it appears M-Prometheus, being largely trained on six languages and evaluated on a set of roughly 30 higher-resourced languages, has a bias towards larger LLMs trained similarly on mostly higher-resourced languages and we found evidence it may be less reliable in measuring less-resourced languages.

Which LLMs output the most English and what kind of English output is it? Smaller LLMs tend to have English responses that refuse to complete the task or summarize in English. Of the smaller models, Aya 101 avoids English responses best. We find that larger, more recent LLMs instead add extra English text in their response. Gemma 3 tends to produce extra English comments the most of models examined, and the type of comments varies some by language.

Single language or multilingual-fine-tuning with synthetic data? We find that generally, multilingual fine-tuning still works best with most of the languages we examined, even when synthetic data is added. However there is evidence from performance in Sorani Kurdish that the best performing augmentation approach with individual language fine-tuning can outperform the multilingual fine-tuning baseline by a significant margin.

What augmentation approach works best? For individual models, we found that each data augmentation approach showed an increase in ROUGE scores over the baseline, but there was not one approach that proved to be definitely better than any other across languages. For multilingual models, back-summarization appeared to be the most competitive augmentation strategy, but the baseline without augmentation performed better for many languages.

8. Conclusion

We have compared a variety of different approaches to summarization with less-resourced languages. We find that LLMs—along with other strategies including data augmentation and a translation pipeline—still under-perform a fine-tuned mT5 multilingual baseline when using traditional reference-based metrics. We report that one of the biggest challenges with using LLMs for summarization in less-resourced languages is that they frequently output English instead of or in addition to the desired language. We also report that the LLM judge M-Prometheus shows a preference for LLM-generated summaries and appears less reliable when evaluating summaries in less-resourced languages than in high-resourced languages. While human evaluation of summaries is costly and especially challenging, a promising direction for future work would be human examination of scoring metrics and model outputs specifically for less-resourced languages that may not be well-represented in evaluation models' training data.

Limitations

As some LLMs in our experiments were loaded with quantization due to resource constraints, it is possible they could have attained higher performance if the non-quantized model could have been used. Marchisio et al. (2024) demonstrated that quantization can have more prominent impacts on human evaluation than automatic metrics and that not all languages are impacted equally with multilingual models. Unfortunately, we were resource constrained and could only make of the LLMs in a quantized setting.

An important limitation to this work is that the evaluation is done entirely with automated metrics. Limitations to summarization metrics are known and human evaluation is preferred. However, human evaluation can be expensive and especially difficult for less-resourced languages due to the added difficulty in recruitment of annotators and quality control with a team of speakers of a diverse set of languages. We have done our best to report reasonable evaluation metrics and release our model generated summaries for further evaluation in future work by speakers of these languages. We unfortunately did not have the funding required to perform a substantial human evaluation even for a subset of the languages we study in this work.

Our work is limited to claims about the particular models and datasets studied. While we examined multiple languages, data augmentation strategies, and LLMs, it is possible that the findings we observed here may not be the same as those on a different set of languages or different LLMs. However, we believe that our experiments and observations are still informative and of interest to the research community.

Ethical Considerations and Broader Impact

Like any text generation model, automatic summarization is based on statistical properties of language and is likely to sometimes generate statements that may be false. The models and approaches described in this work are primarily for research purposes and summaries generated by these models are only intended to be used to aid human creation of summaries and should be viewed with skepticism regarding their factual content.

The datasets used in this work are free and openly available to the public. While we did not collaborate directly with speakers of the languages studied in this work, we make our model outputs publicly available and welcome collaborations with speakers of the languages studied in order to further investigate approaches to summarization in these languages.

Acknowledgments

This work was supported by the grant *Improving Relevance and Recovery by Extracting Latent Query Structure* from eBay to Brandeis University.

Bibliographical References

Kamal Al-Sabahi, Zhang Zuping, and Mohammed Nadher. 2018. A hierarchical structured self-attentive model for extractive document summarization (HSSAS). *IEEE Access*, 6:24205–24212.

Hayastan Avetisyan and David Broneske. 2023. Large language models and low-resource languages: An examination of Armenian NLP. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 199–210, Nusa Dua, Bali. Association for Computational Linguistics.

Soran Badawi. 2023. KurdSum: A new benchmark dataset for the kurdish text summarization. *Natural Language Processing Journal*, 5:100043.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen

Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open LLM leaderboard (2023-2024). https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics.

Hans Christian, Mikhael Pramodana Agus, and Derwin Suhartono. 2016. Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, 7(4):285–294.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya Expanse: Combining research breakthroughs for a new multilingual frontier.

Maxime Darrin, Philippe Formont, Jackie Cheung, and Pablo Piantanida. 2024. COSMIC: Mutual information for task-agnostic summarization evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12696–12717, Bangkok, Thailand. Association for Computational Linguistics.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open LLM leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug

Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

George Giannakopoulos, John M Conroy, Jeff Kubina, Peter A Rankel, Elena Lloret, Josef Steinberger, Marina Litvak, and Benoit Favre. 2017. MultiLing 2017 overview. *MultiLing 2017*, page 1.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Phan Viet Hoang. 2020. Khmer natural language processing toolkit. `https://github.com/VietHoang1512/khmer-nltk`.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 132–141, Melbourne, Australia. Association for Computational Linguistics.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Giannis Karamanolakis, Subhabrata Mukherjee, Guoqing Zheng, and Ahmed Hassan Awadallah. 2021. Self-training with weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 845–863, Online. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the*

2024 Conference on Empirical Methods in Natural Language Processing, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. How does quantization affect multilingual LLMs? *arXiv preprint arXiv:2407.03211*.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10367–10378, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chester Palen-Michel and Constantine Lignos. 2023. LR-sum: Summarization for less-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6829–6844, Toronto, Canada. Association for Computational Linguistics.

Shantipriya Parida and Petr Motlicek. 2019. Abstract text summarization: A low resource challenge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5994–5998, Hong Kong, China. Association for Computational Linguistics.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André F. T. Martins. 2025. M-prometheus: A suite of open multilingual LLM judges.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Dragomir R Radev, Sasha Blair-Goldensohn, and Zhu Zhang. 2001. Experiments in single and multidocument summarization using MEAD. In *First document understanding conference*, pages 1–7.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Alex Salcianu, Andy Golding, Anton Bakalov, Chris Alberti, Daniel Andor, David Weiss, Emily Pitler, Greg Coppola, Jason Riesa, Kuzman Ganchev, Michael Ringgaard, Nan Hua, Ryan McDonald, Slav Petrov, Stefan Istrate, and Terry Koo. 2016. Compact Language Detector v3 (CLD3). `https://github.com/google/cld3`.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. 2021. EstBERT: A pretrained language-specific BERT for Estonian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 11–19, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436.

Elsbeth Turcan, David Wan, Faisal Ladhak, Petra Galuscakova, Sukanta Sen, Svetlana Tchistiakova, Weijia Xu, Marine Carpuat, Kenneth Heafield, Douglas Oard, and Kathleen McKeown. 2022. Constrained regeneration for cross-lingual query-focused extractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2668–2680, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for finnish.

Danqing Wang, Jiaze Chen, Hao Zhou, Xipeng Qiu, and Lei Li. 2021. Contrastive aligned joint learning for multilingual summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2739–2750, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Language Resource References

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Khanh Nguyen and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China. Association for Computational Linguistics.

Chester Palen-Michel, June Kim, and Constantine Lignos. 2022. Multilingual open text release 1: Public domain news in 44 languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2080–2089, Marseille, France. European Language Resources Association.

Chester Palen-Michel and Constantine Lignos. 2023. LR-sum: Summarization for less-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6829–6844, Toronto, Canada. Association for Computational Linguistics.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A. Results Tables

We report ROUGE-1, ROUGE-2, ROUGE-L and BERTScore for experiments with individual models in Table 5 and multilingual models in Table 6. All results are mean results and include standard error which was computed using bootstrap sampling with 500 resamples (Tibshirani and Efron, 1993).

B. Tokenizers

For the all experiments in this work we used the mT5 tokenizer to compute ROUGE scores. For computing novelty and length, we made use of some language specific tokenizers rather than rely on subword tokenization. For Haitian Creole, Georgian, Macedonian, and both varieties of Kurdish, we used utoken[11]. For Armenian, we used Stanza (Qi et al., 2020). and we used khmernltk (Hoang, 2020) for Khmer. The tokenizers used in this work matter both for caluculating ROUGE scores and for determining the mean novelty score. For non-latin scripts, using the rouge package in huggingface's evaluate[12] can result in zero or near zero scores for non-latin script languages without explicitly supplying a tokenizer.

C. Analysis

We conducted further analysis of generated summaries using bigrams to compute mean novelty and also include the mean length of summaries. We include them here due to space constraints in the paper. Table 9 shows the mean novelty scores for summaries computed using bigrams.

D. LLM Prompts

For LLM experiments we use the following prompts. For Aya-101, we use

`“Write a summary for the following article in «LANGUAGE». \n «ARTICLE_TEXT»”`,

where «ARTICLE_TEXT» is replaced with the text of the article to be summarized and «LANGUAGE» is replaced with the desired language. For Mixtral and Llama3, we used:

`“Write a summary for the following article in «LANGUAGE». Write the summary in «LANGUAGE». Do not provide a translation or explain anything. Only provide the summary, do not provide any other information except for the summary in «LANGUAGE». Summarize this article:\n«ARTICLE_TEXT» ”`.

D.1. Larger LLMs

Summarization Prompt: "Summarize the following article using only 2 sentences: «ARTICLE_TEXT»"

Translation prompt: "Translate the following text into «LANG»: «ARTICLE_TEXT»"

D.2. mPrometheus LLM as Judge

"criteria": "Does the model provide a summary of the input article text that has decent seman-

[11] https://github.com/uhermjakob/utoken

[12] https://github.com/huggingface/evaluate

		Baseline				Extractive				Self-train				Backsum			
Dataset	Lang.	R1	R2	RL	BERTScore	R1	R2	RL	BERTScore	R1	R2	RL	BERTScore	R1	R2	RL	BERTScore
lr-sum	ckb	40.33$^{\pm0.04}$	22.00$^{\pm0.05}$	29.65$^{\pm0.05}$	88.34$^{\pm0.01}$	**51.78**$^{\pm0.05}$	**37.69**$^{\pm0.06}$	**43.52**$^{\pm0.06}$	**90.54**$^{\pm0.01}$	46.45$^{\pm0.05}$	30.04$^{\pm0.06}$	37.18$^{\pm0.05}$	89.64$^{\pm0.01}$	47.49$^{\pm0.04}$	31.18$^{\pm0.06}$	37.95$^{\pm0.05}$	89.75$^{\pm0.01}$
lr-sum	hat	24.79$^{\pm0.03}$	9.93$^{\pm0.02}$	18.65$^{\pm0.02}$	82.53$^{\pm0.00}$	28.18$^{\pm0.03}$	11.25$^{\pm0.03}$	20.54$^{\pm0.03}$	84.11$^{\pm0.01}$	27.96$^{\pm0.03}$	**11.53**$^{\pm0.03}$	**20.80**$^{\pm0.03}$	**84.14**$^{\pm0.00}$	**28.20**$^{\pm0.03}$	11.20$^{\pm0.03}$	20.33$^{\pm0.03}$	**84.14**$^{\pm0.00}$
lr-sum	hye	28.03$^{\pm0.02}$	12.41$^{\pm0.02}$	20.20$^{\pm0.02}$	87.22$^{\pm0.00}$	**33.84**$^{\pm0.03}$	18.39$^{\pm0.03}$	**25.85**$^{\pm0.03}$	88.30$^{\pm0.01}$	33.09$^{\pm0.03}$	**18.41**$^{\pm0.03}$	25.76$^{\pm0.03}$	**88.45**$^{\pm0.01}$	20.16$^{\pm0.01}$	5.76$^{\pm0.01}$	15.55$^{\pm0.01}$	84.55$^{\pm0.00}$
lr-sum	kat	22.68$^{\pm0.02}$	7.99$^{\pm0.02}$	18.08$^{\pm0.02}$	84.63$^{\pm0.01}$	26.58$^{\pm0.03}$	11.46$^{\pm0.03}$	21.07$^{\pm0.03}$	87.45$^{\pm0.01}$	27.28$^{\pm0.04}$	12.20$^{\pm0.04}$	21.99$^{\pm0.04}$	87.77$^{\pm0.01}$	**28.09**$^{\pm0.04}$	**13.13**$^{\pm0.04}$	**22.94**$^{\pm0.04}$	**87.85**$^{\pm0.01}$
lr-sum	khm	34.21$^{\pm0.04}$	15.15$^{\pm0.04}$	27.49$^{\pm0.04}$	88.42$^{\pm0.01}$	34.62$^{\pm0.03}$	14.87$^{\pm0.04}$	27.41$^{\pm0.03}$	88.44$^{\pm0.01}$	**34.95**$^{\pm0.04}$	16.03$^{\pm0.04}$	**28.45**$^{\pm0.04}$	**88.59**$^{\pm0.01}$	34.84$^{\pm0.04}$	**16.16**$^{\pm0.04}$	28.40$^{\pm0.04}$	88.52$^{\pm0.01}$
lr-sum	kmr	25.59$^{\pm0.02}$	9.41$^{\pm0.02}$	20.17$^{\pm0.02}$	83.96$^{\pm0.00}$	**30.56**$^{\pm0.03}$	14.40$^{\pm0.04}$	23.53$^{\pm0.03}$	85.92$^{\pm0.01}$	30.40$^{\pm0.04}$	**14.94**$^{\pm0.04}$	**23.80**$^{\pm0.04}$	**86.17**$^{\pm0.01}$	29.07$^{\pm0.03}$	13.22$^{\pm0.04}$	22.28$^{\pm0.04}$	85.76$^{\pm0.01}$
lr-sum	mkd	26.04$^{\pm0.03}$	10.28$^{\pm0.03}$	19.47$^{\pm0.03}$	86.56$^{\pm0.00}$	**26.52**$^{\pm0.03}$	10.21$^{\pm0.03}$	19.79$^{\pm0.03}$	**86.69**$^{\pm0.00}$	26.06$^{\pm0.03}$	10.24$^{\pm0.03}$	19.74$^{\pm0.03}$	86.65$^{\pm0.00}$	26.41$^{\pm0.03}$	**10.84**$^{\pm0.03}$	**19.85**$^{\pm0.03}$	86.67$^{\pm0.00}$

Table 5: Results of data augmentation experiments for individual models for each language. Each score is computed from a single run.

		Baseline				Extractive				Self-Train				Backsum			
Dataset	**Lang.**	**R1**	**R2**	**RL**	**BERTScore**	**R1**	**R2**	**RL**	**BERTScore**	**R1**	**R2**	**RL**	**BERTScore**	**R1**	**R2**	**RL**	**BERTScore**
lr-sum	amh	33.64$^{\pm0.05}$	15.50$^{\pm0.06}$	24.42$^{\pm0.05}$	86.19$^{\pm0.01}$	32.76$^{\pm0.04}$	14.78$^{\pm0.05}$	22.65$^{\pm0.04}$	85.87$^{\pm0.01}$	33.27$^{\pm0.05}$	15.39$^{\pm0.05}$	23.24$^{\pm0.04}$	86.15$^{\pm0.01}$	**34.62**$^{\pm0.05}$	**17.14**$^{\pm0.05}$	**24.21**$^{\pm0.05}$	**86.34**$^{\pm0.01}$
lr-sum	ckb	48.18$^{\pm0.05}$	32.04$^{\pm0.06}$	39.22$^{\pm0.06}$	**90.03**$^{\pm0.01}$	43.92$^{\pm0.04}$	25.60$^{\pm0.06}$	33.28$^{\pm0.05}$	89.04$^{\pm0.01}$	44.40$^{\pm0.05}$	27.06$^{\pm0.06}$	34.78$^{\pm0.05}$	89.28$^{\pm0.01}$	**45.98**$^{\pm0.04}$	**28.60**$^{\pm0.06}$	**35.93**$^{\pm0.05}$	89.47$^{\pm0.01}$
lr-sum	hat	**30.49**$^{\pm0.03}$	**13.62**$^{\pm0.03}$	**23.21**$^{\pm0.03}$	**84.57**$^{\pm0.01}$	27.71$^{\pm0.03}$	10.62$^{\pm0.03}$	20.06$^{\pm0.03}$	84.08$^{\pm0.00}$	27.27$^{\pm0.03}$	10.84$^{\pm0.03}$	19.83$^{\pm0.03}$	83.98$^{\pm0.01}$	26.10$^{\pm0.03}$	9.40$^{\pm0.03}$	18.96$^{\pm0.02}$	83.73$^{\pm0.00}$
lr-sum	hye	**35.83**$^{\pm0.03}$	**20.47**$^{\pm0.03}$	**27.73**$^{\pm0.03}$	**89.02**$^{\pm0.00}$	32.34$^{\pm0.03}$	17.40$^{\pm0.03}$	24.40$^{\pm0.03}$	88.16$^{\pm0.01}$	29.99$^{\pm0.03}$	15.30$^{\pm0.03}$	22.93$^{\pm0.03}$	87.94$^{\pm0.01}$	31.63$^{\pm0.03}$	16.75$^{\pm0.03}$	24.02$^{\pm0.03}$	88.05$^{\pm0.01}$
lr-sum	kat	**30.95**$^{\pm0.04}$	**15.56**$^{\pm0.04}$	**25.46**$^{\pm0.04}$	**88.38**$^{\pm0.01}$	29.08$^{\pm0.04}$	14.22$^{\pm0.04}$	23.53$^{\pm0.04}$	87.89$^{\pm0.01}$	27.28$^{\pm0.04}$	12.23$^{\pm0.04}$	22.30$^{\pm0.04}$	87.70$^{\pm0.01}$	29.51$^{\pm0.03}$	14.80$^{\pm0.04}$	24.27$^{\pm0.04}$	88.05$^{\pm0.01}$
lr-sum	khm	**41.30**$^{\pm0.05}$	23.36$^{\pm0.06}$	35.12$^{\pm0.05}$	89.90$^{\pm0.01}$	36.63$^{\pm0.04}$	16.79$^{\pm0.05}$	29.83$^{\pm0.04}$	88.90$^{\pm0.01}$	37.81$^{\pm0.05}$	19.02$^{\pm0.05}$	31.42$^{\pm0.05}$	89.33$^{\pm0.01}$	37.40$^{\pm0.04}$	18.08$^{\pm0.05}$	30.71$^{\pm0.04}$	89.20$^{\pm0.01}$
lr-sum	kmr	**34.25**$^{\pm0.04}$	18.62$^{\pm0.05}$	27.49$^{\pm0.04}$	86.92$^{\pm0.01}$	28.29$^{\pm0.03}$	12.77$^{\pm0.03}$	21.48$^{\pm0.03}$	85.61$^{\pm0.01}$	24.92$^{\pm0.04}$	11.53$^{\pm0.04}$	19.67$^{\pm0.04}$	85.80$^{\pm0.01}$	29.48$^{\pm0.03}$	13.11$^{\pm0.04}$	22.58$^{\pm0.04}$	85.92$^{\pm0.01}$
lr-sum	lao	**44.11**$^{\pm0.02}$	23.89$^{\pm0.03}$	36.98$^{\pm0.02}$	89.59$^{\pm0.00}$	30.03$^{\pm0.02}$	13.64$^{\pm0.02}$	25.16$^{\pm0.02}$	88.02$^{\pm0.00}$	29.88$^{\pm0.02}$	13.93$^{\pm0.02}$	25.32$^{\pm0.02}$	88.04$^{\pm0.00}$	31.26$^{\pm0.02}$	14.91$^{\pm0.02}$	26.48$^{\pm0.02}$	88.23$^{\pm0.00}$
lr-sum	mkd	**27.70**$^{\pm0.03}$	11.57$^{\pm0.03}$	21.39$^{\pm0.03}$	87.44$^{\pm0.01}$	25.58$^{\pm0.03}$	9.89$^{\pm0.03}$	19.50$^{\pm0.03}$	86.73$^{\pm0.01}$	25.99$^{\pm0.03}$	10.39$^{\pm0.03}$	19.87$^{\pm0.03}$	86.70$^{\pm0.01}$	26.07$^{\pm0.03}$	10.02$^{\pm0.03}$	19.63$^{\pm0.03}$	86.89$^{\pm0.01}$
lr-sum	mya	**47.77**$^{\pm0.02}$	26.52$^{\pm0.02}$	37.94$^{\pm0.02}$	88.11$^{\pm0.00}$	35.30$^{\pm0.03}$	19.69$^{\pm0.03}$	29.42$^{\pm0.03}$	88.37$^{\pm0.01}$	37.39$^{\pm0.03}$	21.89$^{\pm0.04}$	31.71$^{\pm0.03}$	88.75$^{\pm0.01}$	37.32$^{\pm0.03}$	21.60$^{\pm0.03}$	31.63$^{\pm0.03}$	88.76$^{\pm0.01}$
lr-sum	pus	**45.44**$^{\pm0.01}$	**24.67**$^{\pm0.01}$	**35.00**$^{\pm0.01}$	88.47$^{\pm0.00}$	44.46$^{\pm0.01}$	23.38$^{\pm0.01}$	34.10$^{\pm0.01}$	88.17$^{\pm0.00}$	43.94$^{\pm0.01}$	22.97$^{\pm0.01}$	33.65$^{\pm0.01}$	88.14$^{\pm0.00}$	44.71$^{\pm0.01}$	23.70$^{\pm0.01}$	34.32$^{\pm0.01}$	88.32$^{\pm0.00}$
lr-sum	sna	**26.50**$^{\pm0.03}$	**13.46**$^{\pm0.03}$	**21.75**$^{\pm0.03}$	**85.19**$^{\pm0.01}$	24.96$^{\pm0.03}$	11.78$^{\pm0.03}$	19.94$^{\pm0.03}$	84.93$^{\pm0.01}$	22.65$^{\pm0.03}$	9.80$^{\pm0.03}$	18.01$^{\pm0.03}$	84.84$^{\pm0.01}$	25.25$^{\pm0.03}$	11.80$^{\pm0.03}$	20.45$^{\pm0.03}$	85.03$^{\pm0.01}$
lr-sum	som	36.90$^{\pm0.05}$	19.05$^{\pm0.05}$	28.60$^{\pm0.05}$	87.35$^{\pm0.01}$	35.44$^{\pm0.05}$	17.31$^{\pm0.05}$	26.88$^{\pm0.05}$	86.71$^{\pm0.01}$	36.87$^{\pm0.05}$	18.47$^{\pm0.05}$	28.36$^{\pm0.05}$	87.17$^{\pm0.01}$	**38.33**$^{\pm0.05}$	**20.27**$^{\pm0.05}$	**30.30**$^{\pm0.05}$	**87.67**$^{\pm0.01}$
xlsum	amh	39.50$^{\pm0.02}$	21.93$^{\pm0.02}$	29.93$^{\pm0.02}$	87.62$^{\pm0.00}$	39.64$^{\pm0.02}$	21.69$^{\pm0.02}$	29.66$^{\pm0.02}$	87.57$^{\pm0.00}$	38.49$^{\pm0.02}$	20.84$^{\pm0.02}$	29.13$^{\pm0.02}$	87.40$^{\pm0.00}$	**40.46**$^{\pm0.02}$	**22.89**$^{\pm0.02}$	**30.62**$^{\pm0.02}$	**87.80**$^{\pm0.00}$
xlsum	gla	**36.94**$^{\pm0.02}$	**17.94**$^{\pm0.02}$	**27.93**$^{\pm0.02}$	**87.08**$^{\pm0.00}$	35.85$^{\pm0.02}$	16.29$^{\pm0.02}$	26.45$^{\pm0.02}$	86.69$^{\pm0.00}$	34.29$^{\pm0.02}$	15.56$^{\pm0.02}$	25.91$^{\pm0.02}$	86.46$^{\pm0.00}$	35.89$^{\pm0.02}$	17.10$^{\pm0.02}$	27.13$^{\pm0.02}$	86.95$^{\pm0.00}$
xlsum	ibo	**41.48**$^{\pm0.03}$	**21.00**$^{\pm0.02}$	**30.24**$^{\pm0.02}$	**87.82**$^{\pm0.00}$	31.70$^{\pm0.03}$	14.39$^{\pm0.02}$	23.08$^{\pm0.02}$	86.84$^{\pm0.00}$	31.24$^{\pm0.03}$	13.89$^{\pm0.02}$	23.38$^{\pm0.02}$	86.82$^{\pm0.00}$	38.04$^{\pm0.03}$	19.09$^{\pm0.03}$	28.14$^{\pm0.02}$	87.24$^{\pm0.01}$
xlsum	mya	**45.23**$^{\pm0.02}$	**25.74**$^{\pm0.02}$	**35.35**$^{\pm0.02}$	**89.75**$^{\pm0.00}$	44.47$^{\pm0.02}$	24.86$^{\pm0.02}$	34.48$^{\pm0.02}$	89.54$^{\pm0.00}$	44.00$^{\pm0.02}$	24.31$^{\pm0.02}$	34.24$^{\pm0.02}$	89.44$^{\pm0.00}$	44.95$^{\pm0.02}$	25.36$^{\pm0.02}$	35.26$^{\pm0.02}$	89.72$^{\pm0.00}$
xlsum	orm	33.23$^{\pm0.02}$	15.34$^{\pm0.02}$	25.24$^{\pm0.02}$	85.59$^{\pm0.00}$	33.13$^{\pm0.02}$	15.02$^{\pm0.02}$	24.91$^{\pm0.02}$	85.45$^{\pm0.00}$	32.90$^{\pm0.02}$	14.46$^{\pm0.02}$	24.50$^{\pm0.01}$	85.28$^{\pm0.00}$	**33.63**$^{\pm0.02}$	**15.83**$^{\pm0.02}$	**25.51**$^{\pm0.02}$	**85.60**$^{\pm0.00}$
xlsum	pus	**45.64**$^{\pm0.01}$	24.13$^{\pm0.01}$	35.10$^{\pm0.01}$	88.80$^{\pm0.00}$	44.92$^{\pm0.01}$	23.75$^{\pm0.01}$	34.62$^{\pm0.01}$	88.68$^{\pm0.00}$	44.56$^{\pm0.01}$	23.06$^{\pm0.01}$	34.22$^{\pm0.01}$	88.55$^{\pm0.00}$	45.47$^{\pm0.01}$	**24.28**$^{\pm0.01}$	**35.25**$^{\pm0.01}$	**88.92**$^{\pm0.00}$
xlsum	sin	**38.67**$^{\pm0.03}$	**24.23**$^{\pm0.03}$	**30.37**$^{\pm0.03}$	**89.62**$^{\pm0.00}$	37.43$^{\pm0.03}$	22.73$^{\pm0.03}$	28.57$^{\pm0.02}$	89.28$^{\pm0.00}$	37.15$^{\pm0.03}$	22.81$^{\pm0.03}$	29.26$^{\pm0.02}$	89.46$^{\pm0.00}$	38.40$^{\pm0.02}$	24.02$^{\pm0.03}$	30.03$^{\pm0.02}$	89.56$^{\pm0.00}$
xlsum	som	38.15$^{\pm0.02}$	19.08$^{\pm0.02}$	28.11$^{\pm0.02}$	87.56$^{\pm0.00}$	37.30$^{\pm0.02}$	18.13$^{\pm0.02}$	27.85$^{\pm0.02}$	87.35$^{\pm0.00}$	38.15$^{\pm0.02}$	18.40$^{\pm0.02}$	27.81$^{\pm0.02}$	87.43$^{\pm0.00}$	**38.46**$^{\pm0.02}$	**19.25**$^{\pm0.02}$	**28.60**$^{\pm0.02}$	**87.63**$^{\pm0.00}$
xlsum	yor	**44.34**$^{\pm0.02}$	**21.69**$^{\pm0.02}$	**32.30**$^{\pm0.02}$	**88.44**$^{\pm0.00}$	43.66$^{\pm0.02}$	20.44$^{\pm0.02}$	31.21$^{\pm0.02}$	88.25$^{\pm0.00}$	43.31$^{\pm0.02}$	20.81$^{\pm0.02}$	31.41$^{\pm0.02}$	88.35$^{\pm0.00}$	43.95$^{\pm0.02}$	21.54$^{\pm0.02}$	32.15$^{\pm0.02}$	88.38$^{\pm0.00}$

Table 6: Results of multilingual models for different data augmentation approaches. Standard error reported using bootstrapping with 500 samples.

		Mixtral			LLama 3			Aya-101		
Dataset	Lang.	% Eng.	R1	R2	% Eng.	R1	R2	% Eng.	R1	R2
LR-Sum	amh	64.6	10.24$^{\pm0.04}$	2.59$^{\pm0.02}$	38.6	**20.04**$^{\pm0.04}$	5.49$^{\pm0.03}$	0.0	19.69$^{\pm0.04}$	**6.80**$^{\pm0.03}$
LR-Sum	ckb	55.8	8.34$^{\pm0.02}$	2.37$^{\pm0.01}$	16.8	**39.40**$^{\pm0.04}$	**24.87**$^{\pm0.05}$	0.0	23.92$^{\pm0.03}$	10.08$^{\pm0.03}$
LR-Sum	hat	2.6	19.61$^{\pm0.02}$	9.15$^{\pm0.01}$	0.2	**26.90**$^{\pm0.02}$	**12.69**$^{\pm0.02}$	0.6	21.47$^{\pm0.03}$	7.39$^{\pm0.02}$
LR-Sum	hye	60.7	10.02$^{\pm0.02}$	3.45$^{\pm0.01}$	4.0	21.85$^{\pm0.02}$	7.96$^{\pm0.01}$	0.0	27.29$^{\pm0.03}$	**12.21**$^{\pm0.02}$
LR-Sum	kat	75.4	5.83$^{\pm0.01}$	1.54$^{\pm0.01}$	45.0	13.18$^{\pm0.02}$	3.67$^{\pm0.01}$	0.0	**23.88**$^{\pm0.03}$	**8.97**$^{\pm0.03}$
LR-Sum	khm	76.2	5.23$^{\pm0.01}$	0.34$^{\pm0.00}$	37.9	11.57$^{\pm0.01}$	1.57$^{\pm0.01}$	0.1	**15.65**$^{\pm0.02}$	**4.33**$^{\pm0.01}$
LR-Sum	kmr	5.4	16.29$^{\pm0.02}$	8.87$^{\pm0.01}$	0.0	**26.24**$^{\pm0.02}$	**12.30**$^{\pm0.02}$	0.2	24.65$^{\pm0.03}$	9.96$^{\pm0.02}$
LR-Sum	lao	63.9	5.54$^{\pm0.01}$	0.80$^{\pm0.00}$	49.2	12.46$^{\pm0.01}$	3.11$^{\pm0.01}$	0.2	**24.36**$^{\pm0.02}$	**10.24**$^{\pm0.01}$
LR-Sum	mkd	6.6	11.38$^{\pm0.01}$	4.54$^{\pm0.01}$	0.3	20.99$^{\pm0.02}$	**8.98**$^{\pm0.01}$	0.0	**23.91**$^{\pm0.02}$	8.69$^{\pm0.02}$
LR-Sum	mya	58.6	4.65$^{\pm0.01}$	0.58$^{\pm0.00}$	54.0	10.77$^{\pm0.01}$	1.72$^{\pm0.00}$	0.1	**15.83**$^{\pm0.01}$	**3.65**$^{\pm0.01}$
LR-Sum	pus	47.4	12.63$^{\pm0.01}$	4.73$^{\pm0.01}$	3.2	34.32$^{\pm0.01}$	15.99$^{\pm0.01}$	0.0	**37.32**$^{\pm0.01}$	**18.25**$^{\pm0.01}$
LR-Sum	sna	3.0	16.29$^{\pm0.02}$	8.48$^{\pm0.01}$	0.0	**21.06**$^{\pm0.02}$	**10.68**$^{\pm0.02}$	0.2	18.89$^{\pm0.02}$	7.21$^{\pm0.02}$
LR-Sum	som	13.9	17.32$^{\pm0.04}$	8.89$^{\pm0.03}$	1.8	**30.91**$^{\pm0.04}$	**15.61**$^{\pm0.03}$	0.0	27.27$^{\pm0.04}$	10.73$^{\pm0.04}$
XL-Sum	amh	67.4	7.38$^{\pm0.01}$	1.72$^{\pm0.01}$	43.7	18.95$^{\pm0.01}$	5.30$^{\pm0.01}$	0.0	**27.09**$^{\pm0.02}$	**12.39**$^{\pm0.02}$
XL-Sum	gla	2.0	20.15$^{\pm0.01}$	9.63$^{\pm0.01}$	0.2	27.29$^{\pm0.02}$	11.96$^{\pm0.01}$	0.0	**35.56**$^{\pm0.02}$	**16.24**$^{\pm0.02}$
XL-Sum	ibo	13.8	18.60$^{\pm0.02}$	8.31$^{\pm0.01}$	0.0	28.64$^{\pm0.02}$	14.00$^{\pm0.01}$	0.0	**40.58**$^{\pm0.02}$	**19.44**$^{\pm0.02}$
XL-Sum	mya	53.3	11.36$^{\pm0.02}$	3.33$^{\pm0.01}$	38.7	20.22$^{\pm0.02}$	5.86$^{\pm0.01}$	0.0	**32.69**$^{\pm0.03}$	**17.30**$^{\pm0.02}$
XL-Sum	orm	16.9	14.99$^{\pm0.01}$	5.83$^{\pm0.01}$	0.7	21.04$^{\pm0.02}$	8.28$^{\pm0.01}$	1.6	**25.04**$^{\pm0.02}$	**9.37**$^{\pm0.01}$
XL-Sum	pus	40.6	14.21$^{\pm0.01}$	5.23$^{\pm0.01}$	4.0	34.14$^{\pm0.01}$	14.31$^{\pm0.01}$	0.0	**39.99**$^{\pm0.01}$	**19.38**$^{\pm0.01}$
XL-Sum	sin	59.4	6.85$^{\pm0.02}$	2.00$^{\pm0.01}$	15.7	22.09$^{\pm0.02}$	9.05$^{\pm0.01}$	0.0	**29.04**$^{\pm0.03}$	**16.06**$^{\pm0.03}$
XL-Sum	som	17.2	16.89$^{\pm0.01}$	6.47$^{\pm0.01}$	0.8	27.75$^{\pm0.01}$	9.58$^{\pm0.01}$	0.0	**29.61**$^{\pm0.02}$	**11.35**$^{\pm0.02}$
XL-Sum	yor	17.3	20.08$^{\pm0.01}$	8.35$^{\pm0.01}$	0.9	29.67$^{\pm0.02}$	12.88$^{\pm0.01}$	0.1	**37.33**$^{\pm0.02}$	**15.13**$^{\pm0.02}$

Table 7: LLM performance across languages measured in ROUGE-1 and ROUGE-2 along with the percentage of generated summary containing English found by language ID. All results are a a mean of 500 bootstrap samples with standard error reported. Rouge-L and BERTScore are included in the main paper, but ROUGE-1 and ROUGE-2 were omitted due to space constraints.

tic coverage, factuality, is consistent with the original article, is informative, coherent, fluent, concise and written in the language the article is written in?", "score1_description": "The model neglects to provide a summary or the summary is not in the intended language.", "score2_description": "The model provides a response but it is not a good summary. The response is factually inaccurate, not very informative, or not very fluent.", "score3_description": "The model provides a summary but it is lacking in some of the desired qualities of a good summary.", "score4_description": "The model provides a reasonable summary of the input text that includes most of the desired qualities of a good summary.", "score5_description": "The model provides an excellent summary that meets

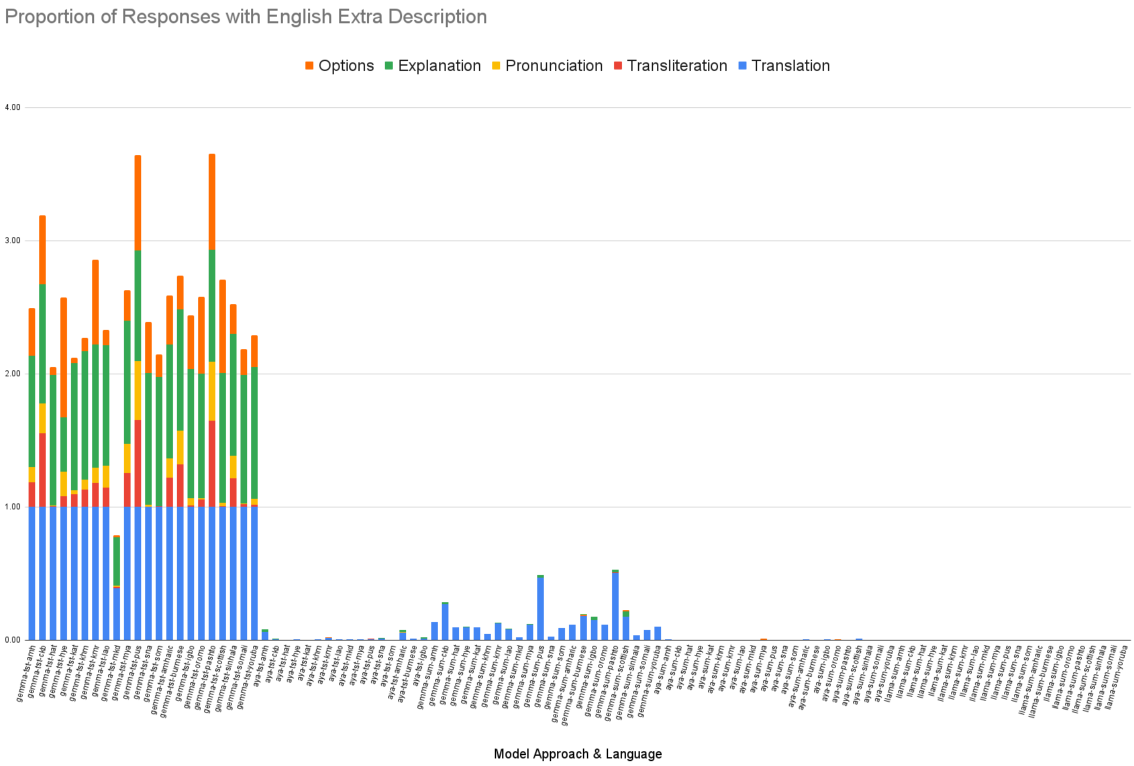

Figure 6: Proportion of summaries with extra English text generated for each model and approach by language by category of Larger LLMs and Translate-Summarize-Translate approach

	Aya TST		Gemma TST		Best Non-TST LLM	
Lang	RL	BERTScore	RL	BERTScore	RL	BERTScore
amh	10.79$^{\pm0.01}$	80.86$^{\pm0.00}$	15.29$^{\pm0.02}$	83.98$^{\pm0.01}$	**21.83**$^{\pm0.03}$	**85.68**$^{\pm0.01}$
ckb	13.62$^{\pm0.01}$	82.95$^{\pm0.00}$	12.44$^{\pm0.01}$	82.18$^{\pm0.00}$	**21.94**$^{\pm0.02}$	**85.18**$^{\pm0.00}$
hat	12.98$^{\pm0.01}$	82.62$^{\pm0.00}$	14.57$^{\pm0.01}$	83.15$^{\pm0.00}$	**20.48**$^{\pm0.02}$	**84.25**$^{\pm0.00}$
hye	11.68$^{\pm0.01}$	83.64$^{\pm0.00}$	14.84$^{\pm0.01}$	86.49$^{\pm0.00}$	**20.48**$^{\pm0.02}$	**87.76**$^{\pm0.00}$
kat	11.86$^{\pm0.01}$	84.88$^{\pm0.00}$	12.61$^{\pm0.01}$	85.96$^{\pm0.00}$	**20.12**$^{\pm0.02}$	**87.63**$^{\pm0.00}$
khm	3.49$^{\pm0.00}$	79.13$^{\pm0.00}$	10.83$^{\pm0.01}$	84.96$^{\pm0.00}$	**15.02**$^{\pm0.01}$	**85.91**$^{\pm0.00}$
kmr	9.77$^{\pm0.01}$	82.07$^{\pm0.00}$	11.69$^{\pm0.01}$	84.18$^{\pm0.00}$	**19.70**$^{\pm0.02}$	**85.88**$^{\pm0.00}$
lao	6.41$^{\pm0.00}$	81.26$^{\pm0.00}$	13.25$^{\pm0.00}$	85.81$^{\pm0.00}$	**23.52**$^{\pm0.01}$	**87.78**$^{\pm0.00}$
mkd	12.00$^{\pm0.01}$	84.27$^{\pm0.00}$	13.15$^{\pm0.01}$	86.00$^{\pm0.00}$	**16.35**$^{\pm0.01}$	**86.70**$^{\pm0.00}$
mya	8.10$^{\pm0.01}$	82.15$^{\pm0.00}$	8.80$^{\pm0.00}$	84.88$^{\pm0.00}$	**23.88**$^{\pm0.02}$	**87.49**$^{\pm0.00}$
pus	14.83$^{\pm0.00}$	82.33$^{\pm0.00}$	17.65$^{\pm0.00}$	85.37$^{\pm0.00}$	**25.03**$^{\pm0.01}$	**87.11**$^{\pm0.00}$
sna	8.46$^{\pm0.01}$	81.66$^{\pm0.00}$	9.66$^{\pm0.01}$	81.98$^{\pm0.00}$	**16.72**$^{\pm0.02}$	**83.48**$^{\pm0.00}$
som	13.44$^{\pm0.01}$	83.47$^{\pm0.00}$	15.61$^{\pm0.02}$	85.19$^{\pm0.01}$	**23.83**$^{\pm0.03}$	**86.71**$^{\pm0.01}$

Table 8: Results for Translate-Summarize-Translate Pipeline approach

all of the requested criteria of a good summary."

E. Example Model Output

We show example augmentation data in Table 10 and examples of LLMs generating English output in Table 11.

F. Novelty and Length

How extractive or abstractive are the multilingual fine-tuned MT5 summaries? While models trained on synthetic data have an advantage in ROUGE score over the baselines trained on only the human written summaries, it is possible that summaries produced by these models are still lacking in certain ways despite having higher scores. In particular, models trained on Extract-Train or Back-

Lang.	LR Sum	Base-line	Extract-Train	Self-Train	Back-Sum
ckb	63.94	9.16	11.15	5.17	5.72
hat	50.46	12.74	9.64	5.58	6.53
hye	69.08	17.78	21.68	16.95	97.33
kat	44.02	32.14	11.04	4.22	6.56
khm	25.81	67.44	68.92	68.48	66.78
kmr	41.26	43.38	10.77	8.90	4.06
mkd	66.74	12.99	18.70	11.87	12.22

Table 9: Mean novelty scores using bigrams

Sum data are trained on summaries generated from extractive models. One concern could be that these models only learn to copy material from the text rather than synthesizing a novel summary. We further probe this issue by computing mean novelty scores for each summary. This score is the percentage of tokens that do not appear in the article text. We compute this novelty score using tokenizers described in Appendix B.

As seen in Table 13, the test set reference summaries have somewhat high novelty. Each individual model generally has lower mean novelty than the test set. We may have expected model trained on extractive summaries to be generally less novel than those trained on self-training; however this does not appear to be the case. This also shows a hint at why Armenian has low ROUGE scores for the back-sum approach. With such a high mean novelty score, there is evidence the model is generating a larger number of irrelevant words.

We show the novelty scores for the multilingual models in Table 12. Similarly we see lower mean novelty from the multilingual models with augmented training data than the reference summaries.

G. Evaluation Figures

We present additional evaluation figures here. Figure 7 and Figure 8 compare M-Prometheus scores with BERTScore and ROUGE scores.

H. Datasets

Table 15 shows counts of documents for each dataset.

I. LLM Selection

Llama 3 8b Instruct (Dubey et al., 2024) and Mixtral 8x7b (Jiang et al., 2024) are two open source LLMs which we use for benchmarking LLM performance in this work because of their reasonable performance on other benchmarks (Beeching et al., 2023; Fourrier et al., 2024) and ease of use for inference using Ollama.[13] While Llama 3 and Mixtral perform reasonably on English, Aya-101 (Üstün et al., 2024) is an LLM trained on much more multilingual data. We use these three LLMs off-the-shelf with simple prompts to compare our fine-tuned mT5 experiments. For larger LLMs we use Gemma-3 27B (Gemma Team et al., 2025), Aya-Expanse 32B (Dang et al., 2024), and Llama 3.3 70B. These models have reported reasonable performance and claim multilingual coverage.

[13] https://ollama.com

	Extractive	Self-Train	BackSum
Summary	Ji girîngtirîn pêşangehên wênekêşiya takekesî ku ji aliyê Moîn Haşemînesab ve hatine organîzekirin, em dikarin behsa pêşangeha Wêne û Muzîk "Notên Bêdawî" û pêşangeha wêne û muzîkê "Bîst û Yek" bikin...	Mêvanê vê xelekê ji bernama Deng û Reng, Moîn Haşimî Neseb e	Ji girîngtirîn pêşangehên wênekêşiya takekesî ku ji aliyê Moîn Haşemînesab ve hatine organîzekirin, em dikarin behsa pêşangeha Wêne û Muzîk "Notên Bêdawî" û pêşangeha wêne û muzîkê "Bîst û Yek" bikin...
Document	Jînenîgarî. Moîn Haşimî Neseb bi çêkirina kurtefîlma "Dîktasyon" û beşdarbûna di çendîn festîvalên navneteweyî yên sala 2021ê de, xelata baştirîn kurtefîlma Festîvala New Yorkê wergirt. Fîlm di Festîvala Fîlman a Glasgowê (TMFF)...	Jînenîgarî. Moîn Haşimî Neseb bi çêkirina kurtefîlma "Dîktasyon" û beşdarbûna di çendîn festîvalên navneteweyî yên sala 2021ê de, xelata baştirîn kurtefîlma Festîvala New Yorkê wergirt. Fîlm di Festîvala Fîlman a Glasgowê (TMFF)...	Roja Pêncşemê li bajarê New Yorkê, ji aliyê Moîn Haşemînesab ve hatine organîzekirin, em dikarin behsa pêşangeha Wêne û Muzîk "Notên Bêdawî"...

Table 10: Examples of Extractive, Self-Training, and Back-Summarization approaches to synthetic data creation using an example from Kurmanji Kurdish

Model	Output
Mixtral	I'm sorry, I'm having difficulty understanding the text you provided. It appears to be written in Lao language and contains some unusual characters. Could you please rephrase or translate the question into English so I can better understand and provide an accurate response?
Mixtral	Overall, Ambassador Goldberg's visit to Luang Prabang reflects the U.S. government's ongoing commitment to supporting sustainable development, cultural preservation, and consular services in Laos, while also fostering stronger bilateral relations between the two countries.
Llama3	I'd be happy to help you with that! However, I need the article text to write a summary in Lao. Please provide the article text, and I'll do my best to summarize it for you in Lao.
Llama3	The Hua Seng Hung Company is one of the companies that have received an investment from the United States. This company has a lot of potential and it's expected to grow rapidly. The company is involved in many fields such as real estate, finance, and technology...

Table 11: Examples of LLM English output when prompted to summarize non-English news articles

	Novelty				Length			
Lang.	Reference	Extractive	Self-Train	BackSum	Reference	Extractive	Self-Train	BackSum
amh	49.8	7.2	9.5	12.7	25.1	16.6	15.8	15.5
ckb	38.9	1.6	1.6	2.0	23.3	26.1	25.0	26.1
hat	18.1	1.0	1.6	2.6	26.7	27.4	26.5	25.8
hye	35.4	2.9	7.2	4.0	24.6	18.5	17.3	18.5
kat	22.3	2.3	6.4	3.4	14.7	15.8	14.5	16.0
khm	7.8	9.3	9.0	8.8	31.6	64.8	59.4	60.6
kmr	16.3	0.5	25.0	1.4	20.2	22.4	16.1	22.7
lao	22.5	13.2	12.3	12.4	28.4	30.0	29.6	29.8
mkd	31.4	2.3	3.4	3.0	20.0	20.3	20.2	20.8
mya	20.0	8.4	9.1	9.1	35.4	33.7	32.6	33.1
pus	21.1	3.1	3.0	3.7	33.2	25.2	25.3	25.1
sna	31.7	1.6	13.3	1.9	17.6	18.4	17.5	18.2
som	26.0	2.8	4.4	5.8	24.7	25.0	23.0	21.6

Table 12: Mean novelty scores and mean lengths of generated summaries by the multilingual models

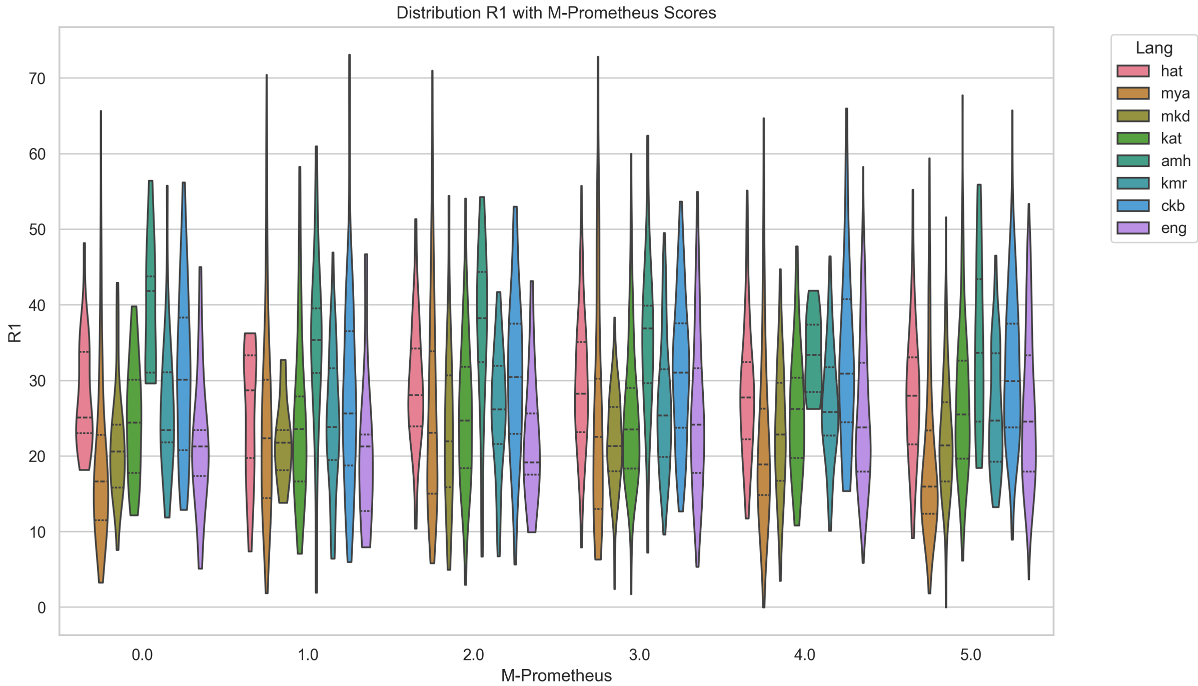

Figure 7: Distribution of summary scores for Llama 3.3 comparing scoring methods ROUGE-1 and M-Prometheus

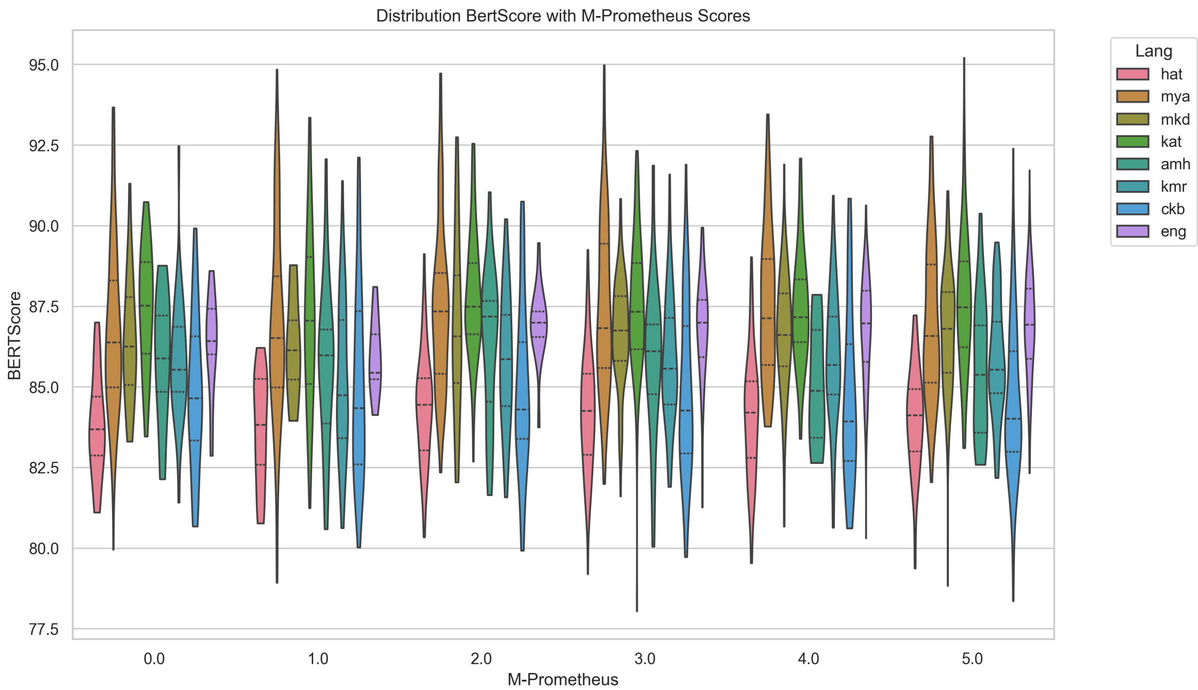

Figure 8: Distribution of summary scores for Llama 3.3 comparing scoring methods BERTScore and M-Prometheus

	Ref.	Base-line	Extract-Train	Self-Train	Back-Sum
ckb	38.9	3.0	4.6	1.7	2.0
hat	18.1	6.7	2.6	1.4	2.0
hye	35.4	5.7	6.8	4.5	66.0
kat	22.3	19.1	4.2	1.4	1.9
khm	7.8	6.6	7.6	6.9	6.2
kmr	16.3	15.8	3.9	4.1	1.2
mkd	31.4	4.7	6.4	3.7	3.9

Table 13: Mean Novelty for summaries generated by individual models and the summaries of the test set (LR-Sum)

Lang.	Ref.	Base-line	Extract-Train	Self-Train	Back-Sum
ckb	23.3	25.1	27.9	25.0	26.8
hat	26.7	20.9	31.4	26.9	29.4
hye	24.6	22.2	19.9	16.9	16.8
kat	14.7	17.9	16.7	14.7	15.4
khm	31.6	71.2	74.9	69.0	71.1
kmr	20.2	15.9	27.4	21.2	22.6
mkd	20.0	21.6	21.2	19.9	21.5

Table 14: Mean lengths for summaries generated by individual models in number of tokens

Language	ISO 639-3	Lang. Family	Train LR-Sum	Train XL-Sum	Train Combined	Wikipedia	Wikipedia Length Filtered
Amharic	amh	Afro-Asiatic	0	5,761	5,761	13,906	7,125
Armenian	hye	Indo-European	920	0	920	303,036	287,288
Burmese	mya	Sino-Tibetan	7,921	4,569	12,490	109,310	17,080
Georgian	kat	Kartvelian	511	0	511	169,602	148,785
Haitian Creole	hat	French Creole	452	0	452	70,159	57,953
Igbo	ibo	Niger-Congo (Volta-Niger)	0	4,183	4,183	22,908	20,496
Khmer	khm	Austro-asiatic	3,888	0	3,888	11,994	4,323
Kurmanji Kurdish	kmr	Indo-Iranian	791	0	791	63,076	36,657
Lao	lao	Kra-Dai	11,964	0	11,964	5,014	3,407
Macedonian	mkd	Indo-European (Slavic)	1,223	0	1,223	139,559	122,754
Oromo	orm	Afro-Asiatic (Cushitic)	0	6,063	6,063	1,970	1,195
Pashto	pus	Indo-Iranian	14,353	16,854	31,207	20,529	15,308
Scottish Gaelic	gla	Indo-European (Celtic)	1,313	0	1,313	15,979	12,398
Shona	sna	Niger-Congo (Bantu)	383	0	383	11,621	9,963
Sinhala	sin	Indo-Iranian	0	3,249	3,249	23,065	16,782
Somali	som	Afro-Asiatic (Cushitic)	0	5,962	5,962	9,021	6,540
Sorani Kurdish	ckb	Indo-Iranian	1,230	0	1,230	52,024	35,098
Yoruba	yor	Niger-Congo (Volta-Niger)	6,350	0	6,350	33,819	7,960

Table 15: Language families and size of training data of LR-Sum and XL-Sum and available additional data from Wikipedia articles in number of documents before and after filtering for documents with more than 5 sentences