MEDIT: Multilingual Text Editing via Instruction Tuning

Vipul Raheja1  Dimitris Alikaniotis1  Vivek Kulkarni1
Bashar Alhafni2∗  Dhruv Kumar3
1Grammarly  2New York University Abu Dhabi
firstname.lastname@grammarly.com

Abstract
We introduce MEDIT, a multilingual extension to COEDIT – the recent state-of-the-art text editing models for writing assistance. MEDIT models are trained by fine-tuning multilingual large, pre-trained language models (LLMs) via instruction tuning. They are designed to take instructions from the user specifying the attributes of the desired text in the form of natural language instructions, such as *Grammatik korrigieren* (German) or 이 텍스트를 단순화 (Korean). We build MEDIT by curating data from multiple publicly available human-annotated text editing datasets for three text editing tasks (Grammatical Error Correction (GEC), Text Simplification, and Paraphrasing) across diverse languages belonging to six different language families. We detail the design and training of MEDIT models and demonstrate their strong performance on many multilingual text editing benchmarks against other multilingual LLMs. We also find that MEDIT generalizes effectively to new languages over multilingual baselines. We publicly release our data, code, and trained models.1

1 Introduction
Large language models (LLMs) have made remarkable progress toward generating fluent and coherent text in a wide variety of tasks and domains to support writing assistance (Brown et al. 2020b; OpenAI 2023; Touvron et al. 2023; *inter alia*). In particular, LLMs have been adapted to perform many complex text editing tasks like GEC (Wu et al., 2023; Coyne and Sakaguchi, 2023; Fang et al., 2023b), text simplification (Baez and Saggion, 2023; Saggion et al., 2022), paraphrasing (Witteveen and Andrews, 2019; Niu et al., 2021), and formality and tone rewriting (Reif et al., 2022; Luo et al., 2023), among others. However, most of these works are restricted to single tasks, with few works adapting LLMs to perform high-quality text editing across multiple tasks (Schick et al., 2023; Raheja et al., 2023; Laban et al., 2023). A lot of these improvements have been driven by fine-tuning large language models (LLMs) with task-specific instruction tuning, resulting in remarkable zero-shot generalization abilities (Sanh et al. 2022; Ouyang et al. 2022; Chung et al. 2022; *inter alia*).

At the same time, significant research effort has been dedicated to leveraging and enhancing the multilingual capabilities of LLMs (Lin et al., 2022). These abilities can be improved using methods such as continued pre-training with abundant monolingual data (Yang et al., 2023b; Cui et al., 2023) or language-specific instruction-tuning (Zhu et al., 2023; Li et al., 2023). However, in the case of continued pre-training, the lack of high-quality multilingual data is a significant challenge. MEDIT can be used to address this challenge by curating data from different languages and using instruction tuning to improve the model’s performance on multilingual tasks.

Multilingual Editing

| Parafrasee la oración: Hoy iré a la escuela a estudiar español. |
| Hoy asistiré a la escuela para aprender español. |
| 풍범 오류 수정: 우리는 어제 행사에서 줄가운 시간을 보내왔습니다. |
| 우리는 어제 행사에서 줄가운 시간을 보냈습니다. |

Cross-lingual Editing

| 文を簡略化してください: Meteorologists often regard a storm surge as the most treacherous facet of a hurricane. |
| A storm surge is considered a hurricane’s most dangerous aspect. |

Figure 1: Examples illustrating multilingual and cross-lingual text editing. The editing instructions are described in bold. Note that the input and output texts are always in the same language. The monolingual vs. cross-lingual setting is determined by comparing the language of the edit instruction to the language of the input text.
Table 1: Set of Languages. The seven languages, along with the ISO-639-1 code and their language family, on which we train and evaluate our models.

<table>
<thead>
<tr>
<th>Language</th>
<th>ISO-639-1</th>
<th>Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>ar</td>
<td>Semitic</td>
</tr>
<tr>
<td>Chinese</td>
<td>zh</td>
<td>Sino-Tibetan</td>
</tr>
<tr>
<td>English</td>
<td>en</td>
<td>Germanic</td>
</tr>
<tr>
<td>German</td>
<td>de</td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>ja</td>
<td>Japonic</td>
</tr>
<tr>
<td>Korean</td>
<td>ko</td>
<td>Koreanic</td>
</tr>
<tr>
<td>Spanish</td>
<td>es</td>
<td>Romance</td>
</tr>
</tbody>
</table>

We address these gaps with mEdIT, a multitask, multilingual extension of CoEdIT (Raheja et al., 2023). mEdIT models can perform text editing operations for three popular tasks: Grammatical Error Correction, Paraphrasing, and Text Simplification, in multilingual and cross-lingual settings (Figure 1) across a diverse set of seven languages, spanning six different language families (Table 1).

To build mEdIT, we fine-tune several multilingual LLMs of varying sizes on carefully curated, largely human-annotated, parallel corpora of over 200k instructional input-output pairs, using publicly available datasets (Table 4) for different text editing tasks. We evaluate the performance of our models extensively on text editing benchmarks in both multilingual and cross-lingual settings to demonstrate their effectiveness.

Figure 2: Data distribution for each of the three tasks and seven languages on which we train. The amount of data is shown in a log scale to aid visualization.

Our contributions are as follows:

- This work, to the best of our knowledge, is the first to investigate multi-task, multilingual text editing via instruction tuning.
- Our models achieve strong performance on multiple text editing tasks across numerous languages and are publicly released for fostering further MTE research.
- Through a comprehensive set of controlled experiments, we provide insights on how model performance on multilingual text editing tasks is affected by various choices like model architecture, model scale, and training data mixtures.

2 Related Work

Multi-lingual LLMs for Text Editing There is an extensive body of prior literature that has leveraged LLMs for various multi-lingual text editing tasks. These works have proposed models for text editing tasks like GEC (Rothe et al., 2021; Sun et al., 2022), paraphrasing (Chowdhury et al., 2022), formality style transfer (Briakou et al., 2021), and text simplification (Mallinson et al., 2020; Martin et al., 2022; Ryan et al., 2023). However, all of these prior approaches have proposed task-specific multi-lingual models. In contrast, we propose a single unified text-editing model for all the considered tasks by leveraging the power of instruction-tuning and task-specific fine-tuning, which enables our multi-lingual models to generalize to multiple text-editing tasks.

Multi-lingual Instruction-Tuning While numerous multi-lingual instruction fine-tuned models like Muennighoff et al. (2023); Wei et al. (2023) and Li et al. (2023) have been developed, they are not focused or tailored for text editing tasks, which we address by task-specific fine-tuning. Specific
to text editing, many prior works have explored instruction tuning capable of performing multiple text editing tasks with a single model, such as GEC, simplification, sentence fusion, style transfer, and paraphrasing, to name a few (Mallinson et al., 2022; Du et al., 2022; Kim et al., 2022; Schick et al., 2023; Raheja et al., 2023). However, while they are able to support multi-task text editing, they are generally mono-lingual and typically restricted to specific languages (predominantly English). Thus, our work addresses this gap by proposing multi-lingual, instruction-tuned models for multiple text editing and revision tasks.

3 MEDIT

3.1 Tasks and Languages
We chose a broad set of languages to ensure coverage and chose text editing tasks that had multilingual, publicly available human-annotated datasets to ensure high data quality. Another criteria was to choose languages at the intersection of the publicly available corpora we could find across a large set of languages for all the tasks we considered. We refer to this as the MEDIT dataset.

Table 1 describes the languages covered in our work, whereas Figure 2 depicts the amounts of training datasets that were available for all tasks and languages we considered. Appendix A details all the training and testing datasets.

3.2 Models
We fine-tune different versions of pre-trained multilingual LLMs (both encoder-decoder/sequence-to-sequence (Seq2Seq) and decoder-only/causal language models (CLM)) on the MEDIT dataset using cross-entropy loss. The details of the MEDIT models are described in § 4.2, whereas the training details are summarized in § 4.3.

4 Experiments

4.1 No-Edits Baseline
We first evaluate a no-edits baseline, where the output is simply a copy of the source input without the instruction. This strategy performs reasonably well on tasks where the target output largely overlaps with the input (e.g., GEC).

4.2 Multilingual LLMs
mT5 (Xue et al., 2021) is a multilingual variant of T5 (Raffel et al., 2020), trained on the mC4 dataset, a multilingual variant of the C4 dataset extended to 101 languages. We experiment with three variants of mT5 – LARGE (770M), XL (3B), and XXL (13B) parameters.

mT0 (Muennighoff et al., 2023) is a family of multilingual Seq2Seq models capable of zero-shot following human instructions in dozens of languages. We use the mT0-LARGE (1.2B), mT0-XL (3.7B), and mT0-XXL (13B) models. These models are constructed by fine-tuning mT5 models on the xP3 cross-lingual task mixture dataset, which consists of multilingual datasets with English prompts. As a result, mT0 models are better suited for following English prompts. We also use the mT0-XXL-MT variant, which is fine-tuned on the xP3mt dataset and is better suited for prompting in non-English.

BLOOMZ (Muennighoff et al., 2023) is a family of multilingual Causal Language Models (CLMs) constructed by fine-tuning BLOOM (Workshop, 2023) on the xP3 dataset. We use BLOOMZ-3b and BLOOMZ-7b1 models for our experiments.

PolyLM (Wei et al., 2023) is a set of multilingual LLMs trained on 640B tokens. We experiment with the PolyLM-MultiAlpaca-13B model, which is PolyLM-13B model fine-tuned on the MultiAlpaca dataset, consisting of 132k samples of multilingual instructions.

Bactrian-X (Li et al., 2023) is a collection of lightweight adapters for LLaMA (7B and 13B) (Touvron et al., 2023) and BLOOM (7B) (Workshop, 2023) on the Bactrian-X dataset, which is a multilingual parallel dataset comprising 3.4 million instruction–response pairs across 52 languages. For simplicity, we only compare against its higher-performant LLaMA-adapted versions.

4.2.1 Large-Pretrained Decoder-only Models
We also conduct zero-shot evaluations against state-of-the-art decoder-only LLMs that have shown impressive multilingual capabilities on a variety of NLP tasks leveraging the power of in-context learning (Lai et al., 2023; OpenAI, 2023).

GPT3.5 (also referred to as ChatGPT),2 is an improved version of GPT3 (Brown et al., 2020a) optimized for chat. We use the gpt-3.5-turbo0613 model from the OpenAI API.3

\footnote{https://openai.com/blog/chatgpt}

\footnote{https://api.openai.com}
**GPT4** (OpenAI, 2023) is the latest iteration of the GPT models and is also optimized for chat. We use the gpt-4-0613 model from the OpenAI API. While we recognize that these models may not be explicitly optimized or trained for multi-lingual settings, considering that they have been trained on massive amounts of web-scale data, these models have been shown to have multi-lingual capabilities (Lai et al., 2023), hence, we consider them as one of our baseline groups.

### 4.3 Training Setup
We perform instruction tuning for all our models by crafting custom prompts for each of the 21 task-language combinations (seven languages, three tasks). Similar to CoEdiIT, for each task-language combination, depending on the number of ways the instructions can be translated without altering the meaning, we write between 14 and 27 instructions by automatically translating each one from English and verifying the accuracy of the translations by asking native language speakers to evaluate and correct them. The total number of task-language instructions is 365, which can be found in Appendix D. We explore three different multilingual and cross-lingual instructional settings, depending on the language of the prompt, where the editing instruction could be in (a) English, (b) the same language as the text being edited (Native), and (c) a random language which may or may not be the same as the language of the text being edited (Random). With this definition, English and Random are cross-lingual text editing tasks, and Native is a multilingual text editing task. In all settings, the input-output pairs are in the same language, but only the language of the instruction changes.

We train all models on 8xA100 80G GPU instances for five epochs. For the PolyLM and Bactrian-X models (>7B parameters), we also use LoRA (Hu et al., 2022) to lower the number of trainable parameters and increase the batch size.

### 4.4 Evaluation
For GEC evaluation, we follow prior work on each language we report on and use the appropriate GEC metric accordingly. Mainly, we use the MaxMatch ($M^2$) Scorer (Dahlmeier and Ng, 2012), ERRANT (Bryant et al., 2017), and GLEU (Napoles et al., 2015, 2016) as our evaluation metrics. The $M^2$ Scorer and ERRANT compare the edits made by a GEC system against annotated reference edits and calculate the precision ($P$), recall ($R$), and $F_{o.5}$ (i.e., weighing precision twice as much as recall). GLEU computes the precision of the n-grams that overlap with the references but not the original texts and penalizes n-grams that overlap with the original texts but not the references.

For simplification, we follow Ryan et al. (2023) and use SARI (Xu et al., 2016a) and BLEU (Papineni et al., 2002) for evaluation. SARI is the average of the F1 score for adding, keeping, and deleting n-grams ($n \in \{1, 2, 3, 4\}$) and has been shown to correlate with human judgments of simplicity (Xu et al., 2016a). BLEU, on the other hand, is a common metric in machine translation and is used as a check for grammatical and meaning preservation. We compute all metrics using the EASSE evaluation suite (Alva-Manchego et al., 2019).

We evaluate paraphrasing on two criteria and metrics: diversity and semantic similarity. For diversity, we use Self-BLEU (Zhu et al., 2018) to measure the diversity of the paraphrases relative to the given source and reference texts. We use Semantic Similarity to measure meaning preservation. Specifically, we use mUSE (Yang et al., 2020) for this, as it is the best-performing multilingual sentence similarity model that supports all the languages in our work. We also considered other notable works that have made significant progress on multilingual sentence similarity, such as Multilingual-SBERT (Reimers and Gurevych, 2020) and LaBSE (Feng et al., 2022). However, we found them unsuitable for our purposes as they were either limited by the languages they support or suffered from lower performance for multilingual sentence similarity.

### 5 Quantitative Results: MTE Quality
We split the models into three main groups. The first group (a) consists of the “no edits” baseline, the second group (b) is the untrained baseline, where models are evaluated in a zero-shot setting without any task-specific fine-tuning, while the third group (c) is our set of multi-lingual models trained on task-specific datasets. For all our experiments, we aggregate the models’ performance by text editing tasks. Specifically, we aggregate the metrics for each task using the harmonic means of its constituents. Specifically, we use (1-Self-BLEU)$^4$ and Semantic Accuracy for Paraphrasing, SARI and BLEU for Simplification, and $F_{0.5}$ and GLEU for GEC. We scale all metrics to lie between

---

$^4$We subtract Self-BLEU from 1 because lower is better in terms of making changes to the source text.
5.1 Baselines

In Figure 3, we report the results of our trained models against various baselines by aggregating the performance on all tasks (as detailed in § 5).

**No Edits (Copy) Baseline** We observe that not making any edits leads to a performance that is on par with the untrained versions of all models, which highlights the limitations of the n-gram overlap-based metrics.

**Untrained Baseline** Similar to Raheja et al. (2023), a core contribution of this work is to push the performance of small- (~1B parameters) to medium-sized (1-15B parameters) LLMs for common text editing tasks across multiple languages. This drives the need for fine-tuning task-specific and language-specific datasets. For this work, we compare our fine-tuned models against their non-fine-tuned counterparts. We find a substantial gap between the untrained models and their trained counterparts, highlighting the impact of task- and language-specific fine-tuning for the tasks.

**English-Only Baseline** In this experiment, we analyze the ability of multilingual LLMs to adapt to different text editing tasks across different languages by fine-tuning them in the most prominent high-resource language (English). We fine-tune all the multi-lingual models on just the English subsets of the training data, as it is the largest in terms of quality and quantity. This experiment tests the need for language-specific data. Similar to the previous result, we observe that the gap between the untrained versions of English-only models is relatively small vs. the ones trained on the full dataset; it increases significantly with language-specific fine-tuning.

**State-of-the-art LLMs** Additionally, we also evaluate against the most powerful commercially available LLMs on their ability to perform MTE. Specifically, we evaluate GPT3.5 and GPT-4 in a zero-shot setting. Although these models have been shown to exhibit strong zero-shot performance on a variety of NLP tasks, we find that the overall performance of both models is close to most untrained baselines. This can be attributed to the rather limited multilingual capabilities of GPT3.5, which often lead to outputs being generated in other languages (English in particular). To some extent, the verbosity of responses is highly detrimental to the performance, especially for GPT4 as it gets penalized by the automatic metrics (especially for GEC).

The rest of this section analyzes different aspects of the quantitative performance of our models.

5.2 Model Performance by Language

In this section, we analyze the performance of different MTE models by language (Figure 5). It is interesting to note that Paraphrasing exhibits a rather steady performance across languages. This can partially be attributed to the weakness of the evaluation metrics as they rely mostly on n-gram overlap, on the multilingual pre-training of the LLMs where they are exposed to medium-large corpora of nearly all the languages, but also that with the increase in the model size and fine-tuning, they
tend to make fewer changes, thus, leading to higher scores. For simplification, the variance in the performance across languages can be attributed to the amount and quality of training data available for each task. For instance, for German, only 1.1k training data points were available, which leads to models not only showing a great improvement in performance with fine tuning, but also a great variance. Similarly, the training data is very noisy for Japanese, which leads to a similar effect. For GEC, we observe that the performance varies a lot by language, indicating the challenging nature of the task. This can be partially attributed to the frequency and the type of errors in each dataset, a phenomenon we see in Arabic datasets. For instance, Arabic datasets contain subtle and frequent errors made by native/L1 speakers, whereas the data for all other languages consists of sparse and infrequency errors made by L2 learners, which could explain the abnormally low quality (on Arabic GEC) of our best-performing model. Moreover, the quality of the training GEC data available also leads to varying performance across languages.

5.3 Language of Instruction

Here, we analyze the effect of the language of the instruction used to instruct the model. As mentioned in § 4.3, we have three configurations for this set of experiments:

English-language Instructions We train the first set of multilingual MEdIT models with just English instructions. These are MTE models capable of performing cross-lingual text editing, trained on data where the instruction is always in English.

Native-language Instructions We train the next set of multi-lingual MEdIT models with instructions in their native language. These models are capable of performing MTE, where the language of edit instructions is the same as the language of texts being edited.

Randomized-language Instructions Finally, we also explore the cross-lingual text editing (Figure 1) abilities of multi-lingual LLMs. To do so, we modify our dataset by appending an edit instruction from a randomly chosen language (different from the language of the edited source-target text pair) and train our models on this cross-lingual-prompted dataset.

Figure 4 shows the effect of instruction language on performance on all three tasks. We note that there is no significant difference in performance between the different settings. This is likely because in each setting, owing to its multilingual instructional pre-training, the model is able to adapt well to the language of the instruction in the fine-tuning phase, hence focusing mostly on the specific tasks.

5.4 Effect of Model Architecture

We present the task-specific results across each model type in Figure 6, observing that CLMs generally are either on par or outperform the rest of the models with GPT3.5, yielding the lowest results.

BLEU relies on n-gram overlap artificially boosting the scores, which highlights the disadvantage of the Copy baseline. Also, GPT3.5 and GPT4 consistently perform poorly in comparison to the rest of the models, which is especially surprising given GPT4’s multilingual capability. In addition, this could be an artifact of the metrics since RLHF can produce excellent results that have little overlap with the references, and future human studies may shed more light on the issue.

CLMs and Seq2Seq models perform similarly on GEC and Paraphrasing, while Seq2Seq performs better on simplification. We posit that this discrepancy happens due to shorter generations from the Seq2Seq models since we observe that the seq2seq models generate significantly shorter sequences than the expected distribution ($D = 0.06, p < 0.001$), increasing the BLEU scores, which are
sensitive to the prediction length, whereas CLMs tend to generate longer sequences ($D = 0.27, p < 0.001$). Looking at the SARI scores, we observe that the two model types do not differ significantly ($p > .05$), indicating similar performance overall.

### 5.5 Effect of Model Scale

In Figure 7, we describe the overall performance of MTE models by size aggregated over the three tasks. It is evident that scaling model size generally increases overall performance significantly, thus reinforcing the effectiveness of model scaling. We also note that all three tasks display similar trends, with relative improvements being the greatest in GEC (28.8%) and Paraphrasing (14.8%).

### 5.6 Effect of Task-specific Data

To understand the effect of task-specific data on model performance, we systematically ablate the proportion of training data for each task. Specifically, we conduct three groups by varying the amounts of training data across a given task between 0%, 10%, 50%, and 100% while keeping the amount of training data across other tasks at 100%, the results of which are shown in Figure 8.

We observe that: (a) Performance on the ablated task generally improves as the amount of training data for that task increases, as expected. As the proportion of training data increases, so does the performance on the specific task. (b) We also note a synergistic relationship between some tasks where training data from one task helps improve performance on a different task. For example, as we add training examples from GEC, we also notice an improvement in model performance on simplification (58.90 vs 45.20) and paraphrasing (65.30 vs 58.34). Similar trends also hold when we add data for the other two tasks. We believe our model (which is inherently multi-task) enables us to leverage such synergy between text editing tasks better, as compared to task-specific models.

---

### Table 2: Zero-shot evaluation results on the language generalization experiments

We present the scores achieved by our best-performing model (Our score) along with the current SOTA results. Wherever possible, we report the metrics reported in the SOTA papers, and if not available, we report commonly used ones by the literature. Note that we focus only on languages that the LLMs have seen during pre-training § 5.7.

<table>
<thead>
<tr>
<th>Task</th>
<th>Language</th>
<th>Dataset</th>
<th>Test Size</th>
<th>Metric</th>
<th>Med/DT</th>
<th>SOTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEC</td>
<td>Romanian</td>
<td>RoGEC</td>
<td>1518</td>
<td>GLEU</td>
<td>45.58</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Hindi</td>
<td>HiWikiEd</td>
<td>13187</td>
<td>ERRANT</td>
<td>32.61</td>
<td>49.4</td>
</tr>
<tr>
<td></td>
<td>Italian</td>
<td>Simpophil</td>
<td>176</td>
<td>SARI</td>
<td>47.84</td>
<td>24.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PaCSSS-IT</td>
<td>63000</td>
<td>BLEU</td>
<td>41.11</td>
<td>36.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hindi</td>
<td>IndicSS</td>
<td>SARI</td>
<td>40.08</td>
<td>19.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rouge-L</td>
<td></td>
<td>45.57</td>
</tr>
<tr>
<td></td>
<td>French</td>
<td>PAWS-X</td>
<td>903</td>
<td>Self-BLEU</td>
<td>69.06</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SA</td>
<td>98.38</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hindi</td>
<td>IndicPara.</td>
<td>Self-BLEU</td>
<td>23.91</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10000</td>
<td>SA</td>
<td>92.06</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>iBLEU</td>
<td>14.2</td>
<td>18.55</td>
</tr>
</tbody>
</table>

---

Footnote: Test datasets for new languages include: RO-GEC: (Cotet et al., 2020), HI-GEC: (Sonawane et al., 2020), IT-SIMP: (Tonelli et al., 2016; Brunato et al., 2016), HI-SIMP: (Kumar et al., 2022), FR-PARA: (Yang et al., 2019), HI-PARA: (Kumar et al., 2022).
5.7 Generalization to new languages

We also explore the capabilities of our MEDIT models on new languages. For every task, we chose two new languages not present in the training set: one related to the language families covered in our training dataset (Table 1) and one belonging to a language family not present in the training dataset. We only considered the languages that the underlying LLM had as part of its pre-training corpora so as to ensure the models had some understanding of the languages in question. Table 2 provides a summary of the languages considered and the datasets they were sourced from, as well as the results of the language generalization experiments. We follow the same metrics for all the tasks as in § 4.4.

MEDIT models are competitive on many unseen languages as compared to the monolingual state-of-the-art, especially on it-Simplification, and hi-GEC and Paraphrasing.7

6 Human Evaluations

We conduct human evaluations of our model outputs by proficient linguists (and native language speakers of the respective languages) on 50 test inputs (per task- language) to ensure they meet the instructional task-specific constraints across the various languages since text editing is often subjective, and automatic metrics are often limited in their effectiveness and accuracy. We evaluate our best-performing MEDIT model8 based on its strong performance (Table 5). Specifically, we conduct a qualitative evaluation where each annotator is shown an instructional input and output from the model and asked to rate the quality of the output on three criteria: fluency (Is the output grammatically correct and sound like it was written by a native speaker of the language?), adequacy (does the output preserve the meaning of the input?), and accuracy (did the model make the desired edits according to the given edit instructions?) of the edited texts, on a Likert scale ranging from Very Bad to Very Good. We collect two annotations for each data point and adjudicate the conflicting judgments with the annotators. The annotation guidelines are provided in Appendix E. Table 3 shows the results of the evaluation. The expert annotators generally rate the model outputs as Good or Very Good across nearly all languages. For Arabic, the preferences are more balanced across the scale and sometimes even leaning towards the Bad or Very Bad across all criteria, which confirms our findings on the automatic metrics as well, indicating that while our model performs very well on languages such as English, German, Chinese, and Spanish, it still has a long way to go in terms of performance for languages such as Arabic in terms of quality, and on Chinese, Japanese, Korean, and Spanish on accuracy.

Table 3: Results of the human evaluation of the model output across three criteria. For each of the criteria, expert human annotators rate the system output, and we note the frequency of their rating (%).

---

7Since no multilingual models have attempted to perform the considered tasks in the respective languages, we are unable to report the metrics on others. We also do not provide any comparisons if the language-specific metrics either do not exist or are reported using different metrics.

8bactrian-x-llama-13b-merged
7 Conclusions

We present MEDIT – an open-sourced dataset and set of multilingual instruction-tuned LLMs capable of following natural language instructions in seven languages to perform various textual editing tasks. It is the first publicly available set of models that heavily outperforms numerous multilingual LLMs on multiple tasks in different languages. Positive feedback from human evaluations shows that MEDIT can assist writers with various aspects of the text revision process at scale by following natural language instructions in multiple languages. Experiments on various multilingual NLP tasks demonstrate that MEDIT models outperform both their corresponding non-fine-tuned and fine-tuned models on other multilingual instruction datasets for text editing, in addition to achieving strong performance on languages unseen during the task-specific fine-tuning. By making our data/models publicly available, we hope to help make advances in multilingual intelligent writing assistants.

Limitations

In this work, we have developed and evaluated instruction-tuned LLMs capable of editing text in multiple languages. However, this work has several limitations that can be improved in future research.

Despite our attempts to cover a diverse set of seven languages, there are still a number of languages that have not been included in our research, largely because of the lack of high-quality, human-annotated text editing data. Our system can be extended to include more languages in the future to better understand the generalization of these models to new languages, and create more accessible and ubiquitous writing assistants.

Secondly, for training and evaluation, we primarily use datasets that are publicly available in specific languages, and sometimes we generate instructions in English and translate them into multiple languages using Google Translate (for example, for simplification tasks). Despite the fact that our approach allows us to support multiple languages with reasonable development costs, data generated and translated might contain unexpected noise. Moreover, they might not best represent expert-annotated edits in different languages. In order to further improve multilingual LLMs for text editing, future research can use human-generated data for training and evaluation.

Thirdly, our system leverages numerous LLMs with billions of parameters. Considering the computing resources required for running and developing these models, replicating the results may prove difficult (which we try to address by sharing our models publicly).

Lastly, our evaluations focus only on the performance of the models on benchmark datasets for text editing, which, in turn, focus primarily on measuring superficial characteristics based on n-gram overlaps. These evaluations are limited as they do not test for the more nuanced aspects of text editing, such as fluency, coherence, and meaning preservation. The lack of human evaluations makes assessing these nuanced characteristics difficult. Future work in this direction could look at robust and scalable evaluation metrics for multilingual text editing.

Ethics Statement

While our models offer several advantages to make intelligent writing assistance more accessible, we do recognize their potential limitations. Since our work mainly focuses on text editing, we are able to avoid many issues involving generating harmful text. Although there is still a possibility of making meaning changes for stylistic tasks due to the lack of user-specific context (Kulkarni and Raheja, 2023), we try to reduce the chance of hallucinations by constraining the generation to strictly editing tasks in order to reduce the chance of adding any new information or perpetuating biases.

Moreover, due to the multilingual settings, there is a risk of our models generating responses that are discriminatory, biased, or contain false information. Hence, our models, when fine-tuned on the text editing datasets, may inadvertently learn or propagate these problematic patterns. To address these concerns and minimize potential harm, we are dedicated to mitigating the risks associated with the use of our models in future research. We strongly advocate for the responsible use of our models to prevent any unintended negative consequences.

Acknowledgments

We sincerely thank Christoph Stuber, David Rojas, Ryan Koo, Samuel Lou, Sofia Mac Gregor, Yuko Okubo, and all the annotators for their immense help, invaluable linguistic expertise, and insightful feedback on the model quality and evaluations. We also thank Max Gubin, Dongyeop Kang, and the anonymous reviewers for their helpful suggestions.
References


Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023b. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation.


Vivek Kulkarni and Vipul Raheja. 2023. Writing assistants should model social factors of language.


Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2016. Glei without tuning.


A Training and Evaluation Datasets

A.1 Grammatical Error Correction

Arabic We report on three publicly available Arabic GEC datasets. The first two come from the QALB-2014 (Mohit et al., 2014) and QALB-2015 (Rozovskaya et al., 2015) shared tasks. The third is the newly created ZAEBUC dataset (Habash and Palfreyman, 2022; Alhafni et al., 2023). QALB-2014 consists of native/L1 user comments from the Aljazeera news website, whereas QALB-2015 consists of essays written by Arabic L2 learners with various levels of proficiency. It is worth noting that the QALB-2015 dataset has two test sets consisting of L1 and L2 data. In this work, we report results on the L1 test set. The ZAEBUC dataset comprises essays written by native Arabic speakers, which were manually corrected. We use the MaxMatch ($M^2$) Scorer (Dahlmeier and Ng, 2012) for the evaluation.

English When it comes to English, we use the Write & Improve + LOCNESS (W&I) corpus released in the Building Educational Applications (BEA) shared task on GEC (Bryant et al., 2019). We also use the NAIST Lang-8 corpus (Tajiri et al., 2012), which is one of the largest and most widely used datasets for English GEC. To test our systems, we use the JFLEG (Napoles et al., 2017) dataset. We use GLEU (Napoles et al., 2015, 2016) for the evaluation.

German For German, we use the Falko-MERLIN corpus (Boyd, 2018), which consists of sentences written by L2 learners that were manually corrected. We use the MaxMatch ($M^2$) Scorer (Dahlmeier and Ng, 2012) for the evaluation.

Spanish For Spanish, we use the publicly available COWS-L2H (Davidson et al., 2020) dataset. COWS-L2H consists of essays written by Spanish L2 learners at the university level in the United States. We use ERRANT (Bryant et al., 2017) for the evaluation.

Chinese For Chinese, we use the data that is part of the NLPCC18 shared task (Zhao et al., 2018). The training data used in the shared task was collected from the NAIST Lang-8 corpus (Tajiri et al., 2012), whereas the test data consists of manually corrected sentences written by Chinese L2 learners. We use GLEU (Napoles et al., 2015, 2016) for the evaluation.

Japanese For Japanese, we use the NAIST Lang-8 corpus (Mizumoto et al., 2011) to train our systems. For evaluation, we use the Japanese L2 TEC-JL dataset (Koyama et al., 2020). We use GLEU (Napoles et al., 2015, 2016) for the evaluation.

Korean We use the recently created Kor-Union dataset (Yoon et al., 2023). Kor-Union was created by collecting and combining GEC data from various sources. This includes essays written by Korean native/L1 speakers and L2 learners. We use the MaxMatch ($M^2$) Scorer (Dahlmeier and Ng, 2012) for the evaluation.

A.2 Paraphrasing

Arabic For training, we use the Arabic SemEval Paraphrasing (ASEP) corpus, which sourced three existing Arabic semantic similarity datasets released during SemEval 2017 Task 1 (Cer et al., 2017), consisting of roughly 1100 sentence pairs. For our purposes, similar to them, we only keep the sentence pairs with a semantic similarity score $\geq$ 3.25, which leads to 603 pairs. We also inverted the pairs for training, leading to a total of 1.2k training pairs. For evaluation, we use the evaluation dataset that was used for SemEval 2017 Track 1, but with the same similarity threshold as the training data, consisting of 67 sentence pairs. This evaluation set consists of sentences from the SNLI Corpus (Bowman et al., 2015) that were human-translated into Arabic, provided by CMU-Qatar by native Arabic speakers with strong English skills.

We also source from the Arabic Question Similarity (Shared Task 8) organized at the Workshop on NLP Solutions for Under-Resourced Languages (NSURL 2019) (Seelawi et al., 2019). The dataset was developed by mawdoo, and consists of 12k pairs for training and 3715 for testing. For both training and evaluation, we filter the semantically similar pairs (similarity score of 1), which leaves us with 10.7k training and 1.7k test pairs.

We also use the Arabic Paraphrasing Benchmark (APB) dataset (Alian et al., 2019), which consists of 1010 Arabic sentence pairs that are collected from different Arabic books. Paraphrasing was performed manually using six transformation procedures (i.e., addition, deletion, expansion, permutation, reduction, and replacement). Similar to other evaluation sets, we only keep the sentence pairs with a semantic similarity score $\geq$ 3.25, which leads to 286 pairs.
Table 4: Datasets used to train and evaluate MEdIT. With the exceptions of Spanish GEC and German Simplification, every other dataset contains >10k examples for all our experiments.

<table>
<thead>
<tr>
<th>Task</th>
<th>Language</th>
<th>Dataset</th>
<th>Split</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>en</td>
<td>BEA (Bryant et al., 2019)</td>
<td>Train</td>
<td>1.1M</td>
</tr>
<tr>
<td></td>
<td>en</td>
<td>JFLEG (Nápoles et al., 2017)</td>
<td>Test</td>
<td>754</td>
</tr>
<tr>
<td>Error</td>
<td>ar</td>
<td>QALB-2014 (Mohit et al., 2014)</td>
<td>Train, Dev</td>
<td>19k, 1k, 968</td>
</tr>
<tr>
<td></td>
<td>QALB-2015 (Rozovskaya et al., 2015)</td>
<td>Test</td>
<td>310, 154, 920</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ZAEBUSC (Habash and Palfreyman, 2022)</td>
<td>Dev</td>
<td>150, 33, 31</td>
<td></td>
</tr>
<tr>
<td></td>
<td>de</td>
<td>Folk-MERLIN (Bloyd, 2018)</td>
<td>Train, Dev, Test</td>
<td>19k, 2.5k, 2.3k</td>
</tr>
<tr>
<td>Correction</td>
<td>es</td>
<td>COWS-L2H (Davison et al., 2020)</td>
<td>Train, Test</td>
<td>398, 85, 86</td>
</tr>
<tr>
<td></td>
<td>ja</td>
<td>Lang-8 (Mizumoto et al., 2011)</td>
<td>Train</td>
<td>1.85M</td>
</tr>
<tr>
<td></td>
<td>ja</td>
<td>TEC-JL (Koyama et al., 2020)</td>
<td>Test</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>ko</td>
<td>Kor-Union (Yoon et al., 2023)</td>
<td>Train, Dev, Test</td>
<td>108.9k, 23.3k, 23.3k</td>
</tr>
<tr>
<td></td>
<td>zh</td>
<td>NLPCC-2018 (Zhao et al., 2018)</td>
<td>Train, Dev, Test</td>
<td>540k, 53.5k, 2k</td>
</tr>
<tr>
<td></td>
<td>en</td>
<td>PAWS (Zhang et al., 2019)</td>
<td>Train, Dev, Test</td>
<td>49k, 8k, 8k</td>
</tr>
<tr>
<td></td>
<td>ar</td>
<td>SemEval 2017 - Task 1 (Cer et al., 2017)</td>
<td>Train, Test</td>
<td>1.2k, 67</td>
</tr>
<tr>
<td></td>
<td>ar</td>
<td>NSURL 2019 - Task 8 (Seelawi et al., 2019)</td>
<td>Train</td>
<td>24k, 3.7k</td>
</tr>
<tr>
<td></td>
<td>fr</td>
<td>NEWSELA-Auto-AR</td>
<td>Dev, Test</td>
<td>94k</td>
</tr>
<tr>
<td></td>
<td>es</td>
<td>PAWS-X (Yang et al., 2019)</td>
<td>Train, Dev, Test</td>
<td>49k, 2k, 2k</td>
</tr>
<tr>
<td></td>
<td>ja</td>
<td>NEWSELA-Auto-ES</td>
<td>Train, Dev, Test</td>
<td>94k</td>
</tr>
<tr>
<td></td>
<td>ko</td>
<td>NEWSELA-Auto-KO</td>
<td>Train, Dev, Test</td>
<td>94k</td>
</tr>
<tr>
<td></td>
<td>zh</td>
<td>NEWSELA-Auto-ZH</td>
<td>Train, Dev, Test</td>
<td>94k</td>
</tr>
<tr>
<td></td>
<td>en</td>
<td>WikiLarge (Zhang and Lapata, 2017)</td>
<td>Train, Dev, Test</td>
<td>296k, 4k, 359</td>
</tr>
<tr>
<td></td>
<td>en</td>
<td>WikiAuto (Jang et al., 2020)</td>
<td>Train</td>
<td>576k, 5k</td>
</tr>
<tr>
<td></td>
<td>en</td>
<td>ASSET (Alva-Manchego et al., 2020)</td>
<td>Train, Dev, Test</td>
<td>94k</td>
</tr>
<tr>
<td>Simplification</td>
<td>ar</td>
<td>ASSET-Auto-AR</td>
<td>Train</td>
<td>2000, 359</td>
</tr>
<tr>
<td></td>
<td>ar</td>
<td>NEWSELA-Auto-AR</td>
<td>Train</td>
<td>94k</td>
</tr>
<tr>
<td></td>
<td>ar</td>
<td>ASSET-Auto-AR</td>
<td>Dev, Test</td>
<td>100, 359</td>
</tr>
<tr>
<td></td>
<td>de</td>
<td>GEOLino (Muller et al., 2020)</td>
<td>Train, Dev, Test</td>
<td>958, 122, 118</td>
</tr>
<tr>
<td></td>
<td>de</td>
<td>TextComplexityDE (Seiffe et al., 2022)</td>
<td>Train</td>
<td>200, 25, 25</td>
</tr>
<tr>
<td></td>
<td>es</td>
<td>NEWSELA-Auto-ES</td>
<td>Train, Dev, Test</td>
<td>94k</td>
</tr>
<tr>
<td></td>
<td>es</td>
<td>ASSET-Auto-ES</td>
<td>Train, Dev, Test</td>
<td>100, 359</td>
</tr>
<tr>
<td></td>
<td>ja</td>
<td>EasyJapanese (Muneyama and Yamamoto, 2018)</td>
<td>Train, Dev, Test</td>
<td>48k, 1k, 1k</td>
</tr>
<tr>
<td></td>
<td>ja</td>
<td>EasyJapanese Extended (Katoh and Yamamoto, 2018)</td>
<td>Train, Dev, Test</td>
<td>34k, 731</td>
</tr>
<tr>
<td></td>
<td>ko</td>
<td>NEWSELA-Auto-KO</td>
<td>Train, Dev, Test</td>
<td>94k</td>
</tr>
<tr>
<td></td>
<td>ko</td>
<td>ASSET-Auto-KO</td>
<td>Train, Dev, Test</td>
<td>100, 359</td>
</tr>
<tr>
<td></td>
<td>zh</td>
<td>NEWSELA-Auto-ZH</td>
<td>Train, Dev, Test</td>
<td>94k</td>
</tr>
<tr>
<td></td>
<td>zh</td>
<td>CSS (Yang et al., 2023a)</td>
<td>Train, Dev, Test</td>
<td>383, 383</td>
</tr>
</tbody>
</table>

English Paraphrase Adversaries from Word Scrambling (PAWS) is a dataset that contains pairs of sentences with a high lexical overlap (Zhang et al., 2019). We use the PAWS dataset for training and evaluation.

German, Spanish, Japanese, Korean, Chinese We use the Cross-lingual Paraphrase Adversaries from Word Scrambling (Yang et al., 2019) dataset (PAWS-X), which was created by translating a subset of the PAWS validation and test sets to six other languages by professional translators. A potential limitation of this approach could be the poor quality of the translation models, which could negatively impact the overall data quality. Therefore, we use the latest Google Translate API\footnote{https://cloud.google.com/translate/docs/advanced/translating-text-v3} to construct the translated data, and further verify the quality of the translated text with human annotators (native speakers) for a subset of the data. We chose the Google API since it performed best amongst the other open-source machine translation engines.

A.3 Simplification

We draw on a variety of existing text simplification datasets in various languages. Table 4 shows the different simplification datasets we draw on in our work and also outlines the training, development, and test settings.

A major issue with text simplification is the absence of publicly available, human-annotated, sentence-level parallel corpora for some of the languages we considered, such as Arabic, Spanish, and Korean. Therefore, we addressed this by translating the Text Simplification datasets for English to these three languages, in which the parallel data is absent. One potential limitation of this approach could be the poor quality of the translation models, which could negatively impact the overall data quality. Therefore, we use the latest Google Translate API\footnote{https://cloud.google.com/translate/docs/advanced/translating-text-v3} to construct the translated data, and further verify the quality of the translated text with human annotators (native speakers) for a subset of the data. We chose the Google API since it performed best amongst the other open-source machine translation engines.
models and APIs we tested\textsuperscript{1011}.

**English** For English, we used Wikilarge (Zhang and Lapata, 2017), WikiAuto (Jiang et al., 2020), and Newsela (Xu et al., 2015) datasets for training. WikiAuto is a neural CRF-aligned corpus of original and simple Wikipedia documents that are automatically aligned to generate sentence pairs, whereas the Newsela (Xu et al., 2015) dataset contains automatically aligned sentence pairs from documents that are generated by professional writers at Newsela for various grade levels. For testing, we use ASSET (Alva-Manchego et al., 2020), which contains ten high-quality human-written simplifications for each of the 2,390 sentences from the TurkCorpus (Xu et al., 2016b).

**German** We use the GEOLino (Mallinson et al., 2020) and TextComplexityDE (Seiffe et al., 2022) datasets for both training and testing. GEOLinoTest contains text about nature, physics, and people from GeoLino, a children’s magazine that was manually simplified by a German linguist to a five to seven-year-old reading level. TextComplexityDE contains 250 complex sentences from German Wikipedia that native speakers manually simplified.

**Japanese** We use the EasyJapanese (Maruyama and Yamamoto, 2018) and EasyJapaneseExtended (Maruyama and Yamamoto, 2018) datasets for training and testing. EasyJapanese contains 50k sentence pairs that were manually created by five students by simplifying text from the Tanaka corpus (Tanaka, 2001). The EasyJapaneseExtended dataset contains an additional 34.4k sentences from the Tanaka corpus with simplifications crowdsourced.

**Arabic, Spanish, Korean** For Arabic, Spanish and Korean, as there were no publicly available sentence-level parallel datasets available, we translated the English simplification datasets. Specifically, we translated the English Newsela dataset for training and ASSET for testing using the Google Translate API, giving us 94k and 359 examples for training and testing, respectively.

**Chinese** We found no publicly available dataset for training Chinese Simplification. Therefore, we again translated the English Newsela training dataset into Chinese. However, for the testing set, we use the CSS (Yang et al., 2023a) dataset. CSS consists of two human-written simplifications for each of the 383 original sentences from the PFR corpus.\textsuperscript{12}

### B Data Preparation

For Seq2Seq models, we prepend the task-specific instructions to the input to the encoder for each language, performing a full parameter update on the entire sequence. We construct the CLM datasets by wrapping each example in model-specific instructions\textsuperscript{13} computing the loss only on the target text; hence, in the Native and Random settings, the model does not optimize for the specific “translated” instructions.

We randomly sample 10k examples from the original datasets for each language-task combination and keep the original validation and test tests (see Table 4). We chose this quantity as a balance between computational cost and qualitative performance based on the insights from (Raheja et al., 2023). Moreover, in our experiments, we did not find a significant impact of increasing the data quantity per task-language combination beyond 10k. In the Spanish GEC task, we only have 398 data points, so this portion of our data is considerably smaller than the rest of the languages. Furthermore, in the GEC and Simplification training data for all languages, we reserve 20% of the set as input-output pairs without any edits to avoid over-corrections by the model.

### C Results

We present the full set of results for our best-performing random-language setting. In addition to the marginally higher performance that all trained models demonstrate in this setting, we choose this as our representative setting due to the fact that it is the most capable setting for our models, allowing them to perform cross-lingual editing.

We present the results for all trained models on all datasets, as well as the No Edits (Copy) baseline and State-of-the-art LLMs. In the interest of space and interpretability, we skip the detailed results for the untrained baseline as we already show them to massively underperform relative to the trained models. We also compare our results to the previously reported SOTA results across tasks and languages:

\textsuperscript{12}https://www.heywhale.com/mw/dataset

\textsuperscript{13}PolyLM and Bactrian-X follow different prompt templates. Details in Appendix D.
**Grammatical Error Correction**  For GEC, we compare against the following SOTA results: Arabic (Alhafni et al., 2023), English (Zhang et al., 2023), German (Fang et al., 2023a), Korean (Yoon et al., 2023), Spanish (Flachs et al., 2021), and Japanese (Koyama et al., 2020).

**Simplification**  For simplification, we take the results from the fine-tuned mT5 models from (Ryan et al., 2023). They compare multiple settings in their work, such as using data from a single training dataset, from a single language, and from multiple languages. We pick the score from any setting that gives the highest score. Thus, the SARI score for German might be picked from one setting while the BLEU score for German from some other setting, and so on. This ensures that we take their strongest models for each use case. We still see that our models perform better than them in most languages and datasets.

**Paraphrasing**  To paraphrase, most of the related works that utilize the PAWS-X dataset (Yang et al., 2019) have used it for paraphrase detection, and not for paraphrase generation. Hence, we are unable to report any comparable SOTA results for this task.

**D Task Verbalizers**  

We present the full list of our manually curated task-specific verbalizers used for training and evaluations in Table 6, Table 7, and Table 8.
The MBZUAI/bactrian-x-llama-7b-merged checkpoint, and bx-llama-7b denotes the MBZUAI/bactrian-x-llama-13b-merged one.

Finally, for Simplification, the first quantity is SARI, and the second one is BLEU. In terms of Table 5:

<table>
<thead>
<tr>
<th>Model</th>
<th>Version</th>
<th>Size</th>
<th>en</th>
<th>de</th>
<th>es</th>
<th>ja</th>
<th>ko</th>
<th>zh</th>
<th>ar</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-4</td>
<td>–</td>
<td>–</td>
<td>41.63</td>
<td>39.25</td>
<td>45.32</td>
<td>50.2</td>
<td>56.80</td>
<td>1.91</td>
<td>1.07</td>
</tr>
<tr>
<td>PolLM</td>
<td>–</td>
<td>–</td>
<td>36.17</td>
<td>32.78</td>
<td>40.87</td>
<td>35.07</td>
<td>36.27</td>
<td>11.66</td>
<td>4.06</td>
</tr>
<tr>
<td>Bactrian-X</td>
<td>–</td>
<td>–</td>
<td>30.57</td>
<td>27.43</td>
<td>34.02</td>
<td>29.18</td>
<td>31.86</td>
<td>3.22</td>
<td>3.33</td>
</tr>
</tbody>
</table>

Table 5: Full set of results on the best-performing setting of 10k random-language-prompted data. For GEC, we report GLEU or F.0.5 depending on the metric as described in Appendix A. For Paraphrasing, the first quantity is Self-BLEU and the second one is the accuracy of semantic similarity as calculated by mUSE (explained in § 4.4). Finally, for Simplification, the first quantity is SARI, and the second one is BLEU.
<table>
<thead>
<tr>
<th>Language</th>
<th>Verbalizers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>ًلَمْ يُصَلِّي الَّذِينَ قَضَيْنَا عَلَيْهِمُ الْأَمْرَ فَلَمْ يُصَلِّي الَّذِينَ قَضَيْنَا عَلَيْهِمُ الْأَمْرَ</td>
</tr>
<tr>
<td>Chinese</td>
<td>文法</td>
</tr>
<tr>
<td>English</td>
<td>Fix grammar</td>
</tr>
<tr>
<td>German</td>
<td>Grammar korrigieren</td>
</tr>
<tr>
<td>Japanese</td>
<td>この文の文法を修正してください</td>
</tr>
<tr>
<td>Korean</td>
<td>편집문법</td>
</tr>
<tr>
<td>Spanish</td>
<td>Corrige grámatica</td>
</tr>
</tbody>
</table>

Table 6: Grammatical Error Correction instruction verbalizers. For every language, we craft 27 GEC-specific instructions, increasing their diversity when the model is trained. For this and subsequent tables, we verify the validity of the instructions with native language speakers (§ 4.3).
<table>
<thead>
<tr>
<th>Language</th>
<th>Verbalizers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>简化語句</td>
</tr>
<tr>
<td></td>
<td>Reescribe esta oración para simplificar</td>
</tr>
<tr>
<td></td>
<td>Simplificar la oración</td>
</tr>
<tr>
<td></td>
<td>Simplificar este párrafo</td>
</tr>
<tr>
<td></td>
<td>Simplificar esta oración con un enfoque más simple</td>
</tr>
<tr>
<td>Chinese</td>
<td>簡述</td>
</tr>
<tr>
<td></td>
<td>重述</td>
</tr>
<tr>
<td></td>
<td>簡述</td>
</tr>
<tr>
<td></td>
<td>簡述</td>
</tr>
<tr>
<td>German</td>
<td>Vereinfachen Sie den Satz</td>
</tr>
<tr>
<td></td>
<td>Vereinfachen Sie diesen Satz um, damit er einfacher ist</td>
</tr>
<tr>
<td></td>
<td>Vereinfachen Sie diesen Satz</td>
</tr>
<tr>
<td></td>
<td>Vereinfachen Sie diesen Satz</td>
</tr>
<tr>
<td>Japanese</td>
<td>文を簡略化してください</td>
</tr>
<tr>
<td></td>
<td>文をより簡単な言語に書き換えてください</td>
</tr>
<tr>
<td></td>
<td>文をより簡単に書き換えてください</td>
</tr>
<tr>
<td></td>
<td>文をより簡単に書き換えてください</td>
</tr>
<tr>
<td>Korean</td>
<td>문장이 간결하게</td>
</tr>
<tr>
<td></td>
<td>문장이 간결하게</td>
</tr>
<tr>
<td></td>
<td>문장이 간결하게</td>
</tr>
<tr>
<td></td>
<td>문장이 간결하게</td>
</tr>
<tr>
<td>Spanish</td>
<td>Simplifica la oración</td>
</tr>
<tr>
<td></td>
<td>Simplifica la oración con una reducción más simple</td>
</tr>
<tr>
<td></td>
<td>Simplificar la oración</td>
</tr>
<tr>
<td></td>
<td>Simplificar la oración con una reducción más simple</td>
</tr>
</tbody>
</table>

Table 7: Simplification instruction verbalizers. For the simplification task, we generate 19 instructions per language, taking care to not change the meaning of the instruction. For more information see § 4.3 and Table 6.
<table>
<thead>
<tr>
<th>Language</th>
<th>Verbalizers</th>
</tr>
</thead>
</table>
| Arabic   | أن كتبت هذه الجملة  
اكتب إعادة سياقاتها من الجملة  
أعد كتابة هذه الجملة  
الرجع  
أعد كتابة جميع سياقاتها  
أعد كتابة هذا الجملة |
| Chinese  | 解释一下句子  
为这句话写一个解释  
写这句话的解释  
重写这句话  
解释一下这段文字  |
| English  | Paraphrase the sentence  
Paraphrase this sentence  
Rewrite the sentence with different wording  
Reword this text  
Reword this text |
| German   | Umschreiben Sie den Satz  
Umschreiben Sie diesen Satz um  
Formulieren Sie diesen Satz um  
Umschreiben Sie diesen Text  
Umschreiben Sie eine Umschreibung für den Satz  |
| Japanese | 文書長く読まれたくない  
この文書は読まれたくない  
別の表現で文書を書き直してください  
この文書を書き直してください  
このテキストを言葉換えください  
文書を使い換えてください  
別の表現を使用してください  
このテキストを書き直してください |
| Korean   | 문자를 줄여야요  
문장을 줄여서 쓰시다  
문장을 줄여서 쓰시겠어요  
문장을 줄여서 쓰시겠어요  
문장을 줄여서 쓰시겠어요  
문장은 줄여서 쓰시겠어요  
문장은 줄여서 쓰시겠어요  
문장은 줄여서 쓰시겠어요  
문장은 줄여서 쓰시겠어요  
문장은 줄여서 쓰시겠어요  
문장은 줄여서 쓰시겠어요  |
| Spanish  | Paráfrase la oración  
Paráfrase esta oración  
Reescribe la oración con una redacción diferente  
Reword this oración  
Paráfrase este texto  |

Table 8: Paraphrasing instruction verbalizers. In this case, we create 14 instructions per language so as not to alter the meaning of the task. For more information see § 4.3 and Table 6.
E  Human Evaluation Annotation
Guidelines
The human experts were asked to rate the **fluency**, **adequacy**, and **accuracy** separately, following the guidelines in Figure 9, Figure 10 and Figure 11, respectively.

**Figure 9: Annotation guidelines for human evaluations for Fluency.**

Given the instruction to edit the text and the text to be edited, rate the output on the following dimensions:

- **Fluency**: is the output valid, free from errors, and like how a native speaker of the language would write?
  - **Very Good**: The output is of high quality, valid, and correct, like a native speaker.
  - **Good**: The output is acceptable with minor errors.
  - **Average**: The output is relevant but has significant errors.
  - **Bad**: The output is subpar quality.
  - **Very Bad**: The output is unusable.

**Figure 10: Annotation guidelines for human evaluations for Adequacy.**

Given the instruction to edit the text and the text to be edited, rate the output on the following dimensions:

- **Adequacy**: does the output preserve the meaning of the original sentence?
  - **Very Good**: The output fully preserves the meaning of the text.
  - **Good**: The output is semantically similar to the input with minor errors.
  - **Average**: The output is semantically similar to the input but has significant errors.
  - **Bad**: The output is barely similar to the input.
  - **Very Bad**: The output has opposite meaning to the input.

**Figure 11: Annotation guidelines for human evaluations for Accuracy.**

Given the instruction to edit the text and the text to be edited, rate the output on the following dimensions:

- **Accuracy**: how well do the edits made in the output follow the given instructions?
  - **Very Good**: The output follows the instructions exactly.
  - **Good**: The output generally follows the instructions with minor errors.
  - **Average**: The instructions are partially followed but has errors.
  - **Bad**: The output did not follow the instructions but did not significantly make the text unusable.
  - **Very Bad**: The instructions were completely ignored, and wrong edits were made to make the text unusable.