# RewriteLM: An Instruction-Tuned Large Language Model for Text Rewriting

**Lei Shu** [*1] **Liangchen Luo** [1] **Jayakumar Hoskere** [1] **Yun Zhu** [1] **Yinxiao Liu** [1] **Simon Tong** [1] **Jindong Chen** [1] **Lei Meng** [*1]

## Abstract

Large Language Models (LLMs) have demonstrated impressive zero-shot capabilities in long-form text generation tasks expressed through natural language instructions. However, user expectations for long-form text rewriting is high, and unintended rewrites ("hallucinations") produced by the model can negatively impact its overall performance. Existing evaluation benchmarks primarily focus on limited rewriting styles and sentence-level rewriting rather than long-form open-ended rewriting.We introduce OPENREWRITEEVAL, a novel benchmark that covers a wide variety of rewriting types expressed through natural language instructions. It is specifically designed to facilitate the evaluation of open-ended rewriting of long-form texts. In addition, we propose a strong baseline model — RewriteLM, an instruction-tuned large language model for long-form text rewriting. We develop new strategies that facilitate the generation of diverse instructions and preference data with minimal human intervention. We conduct empirical experiments and demonstrate that our model outperforms the current state-of-the-art LLMs in text rewriting. Specifically, it excels in preserving the essential content and meaning of the source text, minimizing the generation of "hallucinated" content, while showcasing the ability to generate rewrites with diverse wording and structures.

## 1. Introduction

Text rewriting plays an essential role in a wide range of professional and personal written communications. It can be conceptualized as a form of controllable text generation (Zhang et al., 2022a) , where a specified textual input is modified based on the user's requirement. Several categories of text rewriting have been extensively researched, such as paraphrasing (Siddique et al., 2020; Xu et al., 2012), style transfer (Riley et al., 2020; Zhang et al., 2020; Reif et al., 2021), and sentence fusion (Mallinson et al., 2022).

Recent advances in Large Language Models (LLMs) have shown impressive zero-shot capabilities in a wide range of text generation tasks expressed through natural language instructions (Chung et al., 2022). However, user expectation for text rewriting is high and any unintended edits by the model negatively impact the user's satisfaction. Given that the LLMs can be hard to control (Qin et al., 2023) and prone to generating "hallucinated" content (Ji et al., 2023), we propose methods to ensure that the model is properly trained and tested using instruction datasets that are both diverse and representative.

To this end, we introduce a new benchmark OPENREWRITEEVAL and collect human-generated text rewrites with natural language instructions. Unlike the previous benchmarks for text rewriting, which had restricted types of rewrites (Reif et al., 2021; Mallinson et al., 2022) or mostly were focused on sentence-level rewriting (Riley et al., 2020; Siddique et al., 2020; Mallinson et al., 2022), our benchmark is designed for research on long-form text input and covers a wide variety of rewriting types expressed through natural language instructions.

We also present a strong baseline model — RewriteLM, an instruction-tuned large language model for long-form text rewriting. Similar to InstructGPT (Ouyang et al., 2022), we perform supervised fine-tuning, reward model training, and reinforcement learning (RL) steps to train our model: (1) fine-tuning on Wiki-based instruction dataset and synthetic instruction dataset; (2) training the reward model on human and synthetic preference data; and (3) optimizing a policy by training it against a reward model using reinforcement learning.

InstructGPT heavily relies on human labelers to collect instruction data (for supervised fine-tuning) and preference data (for reward training). We instead develop the following strategies to minimize human intervention in the data collection process: (1) For instruction data generation, we extract long-form, high quality edits with substantial changes from

[*]Equal contribution [1]Google Research. Correspondence to: Lei Shu <leishu@google.com>, Lei Meng <leimeng@google.com>.

*Proceedings of the 40th International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. 

| | |
|---|---|
| Source Text | The way in which we feed our children determines how healthy they are. It also determines how nicely they develop and develop. Children need to be fed a selection of foods each day in order to make sure their physical, emotional, and mental health. All foods are from 1 of the fundamental food teams. Milk and cheese arrive from the dairy team, for instance, and green beans arrive from the vegetable team. Bread arrives from the grain team and beef comes from the meat team. Chocolate arrives from the body fat and sugars team. Our children need so many servings for each day from each of these meals teams to preserve great health. |
| Instruction | Rewrite the text so that it is easy to understand. |
| Target Text | What we give our kids to eat affects how healthy they are and how they grow. Kids need to eat different types of food each day to stay physically, emotionally, and mentally healthy. All foods belong to one of the five food groups: dairy, vegetables, grains, meat and beans, and fruits. Milk and cheese are dairy foods, green beans are vegetables, bread is a grain, beef is a meat, and chocolate is a fat and sugar food. To stay healthy, kids need to eat a certain number of servings from each food group every day. |
| Instruction | Summarize the text. |
| Target Text | Feeding children a variety of foods from the five fundamental food groups (dairy, vegetables, grains, meat, and fats/sweets) is essential for their overall health and development, including their physical, emotional, and mental well-being. |

*Figure 1.* Rewrite examples. The input for RewriteLM contains an instruction and a source text. The target is a rewrite. We present one source text and two instructions here for illustration purpose.

Wiki as our main natural source (Schick et al., 2022) of the source and target pairs, and use their associated comments as proxies for instructions. (2) To increase the amount and the diversity of the instruction data beyond Wikipedia, we leverage Chain-of-thoughts (CoT) and the capability of the LLMs to generate instructions and target text for various input text. (3) To generate more preference data, we sample multiple LLM model outputs and rank them using a human-designed heuristic ranker. We then fine-tune the pretrained language models on the collected data to produce RewriteLM models.

We conduct empirical studies to evaluate the model performance on the OpenRewriteEval benchmark. The results show that even current state-of-the-art pretrained LLMs have poor performance on open-ended rewriting tasks. LLMs fine-tuned on general-purpose instruction datasets like Flan-PaLM (Chung et al., 2022) and Alpaca (Taori et al., 2023) have better performance compared with the pretrained foundation models, but still have room for improvement. The proposed RewriteLMs, including Rewrite-PaLM and Rewrite-PaLM 2, both outperform their corresponding foundation models by a significant margin. They also outperform other instruction-tuned LLMs, showcasing the effectiveness of the generated training data. Applying reinforcement learning on top of the supervised tuned Rewrite-PaLM 2 further improves its performance, resulting in a new state-of-the-art model Rewrite-$\text{RL}_{\text{r/w}}$-PaLM 2 for text rewriting.

Our main contributions can be summarized as follows:

- A new benchmark, OpenRewriteEval, designed for research on long-form text input and covering a wide variety of rewriting types expressed through natural language instructions, such as formality, expansion, conciseness, paraphrasing, tone and style transfer. Unlike previous benchmarks, which were primarily focused on limited rewrite styles and sentence-level rewriting, our benchmark is specifically designed to facilitate open-ended rewriting of long-form text. To the best of our knowledge, no such dataset has existed previously.

- A strong baseline model, RewriteLM, has demonstrated superior performance compared to the state-of-the-art LLMs for text rewriting, especially in long-form. We developed new strategies to generate diverse, long-form rewriting instruction datasets and preference data that can be used to enhance the editing and rewriting capabilities of LLMs. Our results demonstrate that our model has strong performance in rewriting with diverse language and structures while preserving the essential meaning of the original input. Importantly, our model effectively mitigates the issue of “hallucinations”, which is commonly observed in existing LLMs.

## 2. Related Work

**Text Editing**. The majority of the research on rewriting currently focuses on a particular set of editing tasks at the sentence level, such as paraphrase (May, 2021), style transfer (Tikhonov et al., 2019), spelling correction (Napoles et al., 2017), formalization (Rao & Tetreault, 2018), simplification (Xu et al., 2016) and elaboration (Iv et al., 2022). (Faltings et al., 2020) trained an editing model to follow instructions using Wikipedia data. However, their focus was solely on edits limited to a single sentence. PEER (Schick et al., 2022) can follow human-written instructions for updating text in any domain, but is still limited by the edit types available on Wikipedia. Moreover, it was only evaluated on a small set of edit types from a human-defined instruction evaluation benchmark (Dwivedi-Yu et al., 2022).

**Instruction Tuning**. Instruction tuning has shown to improve model performance and generalization to unseen tasks (Chung et al., 2022; Sanh et al., 2022). InstructGPT (Ouyang et al., 2022) extends instruction tuning further with reinforcement learning with human feedback (RLHF), which heavily relies on human labelers to collect instruction data and model output rankings for training. The focus of these works was primarily on extensively researched tasks and benchmarks, which do not include open-ended text rewriting.

**Data Augmentation via LLM**. A common data augmentation approach involves utilizing trained LLMs to generate more data, which is subsequently incorporated as training data to enhance the model's performance (He et al., 2019; Xie et al., 2020; Huang et al., 2022). PEER (Schick et al., 2022) leverage LLMs to infill missing data and then use this synthetic data to train other models. Self-Instruct (Wang et al., 2022a; Taori et al., 2023) improves its ability to accurately follow instructions by bootstrapping off its own generated outputs. Our work builds upon similar ideas and leverages the power of LLMs to enhance existing datasets and generate additional synthetic datasets.

## 3. Methods

In this section, we discuss the training data (Section 3.1) and the training procedure (Section 3.2) for the proposed RewriteLM models.

### 3.1. Training Dataset

#### 3.1.1. Wiki Instruction Dataset

We examine Wiki revisions and extract long-form, high quality edits that contain substantial changes. We also use the associated edit summary of the revision as a proxy for the instructions. We describe edit extraction, edit filtering, and instruction improvement in details:

- **Edit Extraction:** We initiate the instruction tuning data collection process by gathering Wikipedia revision history, where each revision record includes the original text, revision differences, and an edit summary written by the revision author. We extract text block differences between each consecutive snapshots of a Wikipedia article and the associated edit summary, following the approach in Schick et al. (2022). In the rest of the section, we may use the terms *source text*, *target text* and *comment* to denote the text before revision, the text after revision and the edit summary of a revision record, respectively.

- **Edit Filtering:** In order to create long-form, high-quality edits with substantial changes, we remove revision records that meet any of the following criteria: (i) the edit summary indicates low-quality content of a snapshot, such as containing ”revert” or ”vandalism” keywords; (ii) the edit summary contains keywords indicating a format-only change (*e.g.*, bold-facing or hyperlinks), which is not a focus of this work; (iii) the source text contains two or fewer sentences.

- **Instruction Improvement:** The raw comment may not directly meet our data requirements, which can be empty, contain irrelevant descriptions to the revision, or not describe the editing behavior (*e.g.*, only describes the deficiencies of source text). We take the following steps to enhance the quality of the instructions: (i) Extract revision records where the edit summary starts with a verb describing an edit intent (*e.g.*, “make the text easier to read”); (ii) Fine-tune an LLM to generate comments from `<source>-<target>` text pairs as well as learn to control the length and specificity of the instructions. We use the heuristic that if a comment mentions a word from the edit then it is a detailed instruction. (iii) Generate detailed comments for all `<source>-<target>` pairs using the model trained in the previous steps.

#### 3.1.2. Synthetic Instruction Dataset

The Wiki instruction dataset is limited by the available edit types found on Wikipedia. To collect a more diverse and representative instruction dataset, we first use chain-of-thoughts prompting and few-shot prompting to generate instructions, and then generate the target text from a general purpose LLM model:

- **Instruction generation:** By applying a 3-shot chain-of-thought (CoT) prompting method to text inputs from any domain (see Figure 2), we can leverage the knowledge acquired by the LLM during pre-training. This enables the LLM to produce more diverse instructions beyond Wiki edit types. CoT contains two QA stages:

**Model Input**

```
Q1: What kind of text is the following {Nothing hurts more than the distance that we have created between us. I
sometimes wish you were by my side and make my life complete. It's becoming hard for me to live without you.}?
A1: a romantic longing note
Q2: What is a relevant writing prompt or edit instruction for text {Nothing hurts more than the distance that we
have created between us. I sometimes wish you were by my side and make my life complete. It's becoming hard for me
to live without you.}?
A2: make this more poetic

...

Q1: What kind of text is the following {Providing tools that deliver a consistent employee experience anywhere and
at any time can bridge the gap between teams, no matter where they're working from. Employees also need tools that
simply help them get their work done. Being able to easily move between applications and quickly access
information from any device is key. Creating a dedicated Space for shared projects is one way to provide
distributed teams with a central hub to search and share files across applications, assign tasks, brainstorm, and
give feedback at any time, with a persistent record of all interactions. Similarly, the ability to quickly pivot
to a Meet call from Gmail, Chat, or Docs ensures they don't lose momentum when collaborating from anywhere.}
```

**Model Output**

```
A1: A boring sales pitch
Q2: What is a relevant writing prompt or edit instruction?
A2: make it more engaging
```

*Figure 2.* Chain-of-thought (CoT) approach to generating rewrite instructions. The answer to the second question in the output is the generated instruction.

**Text description** (answering "What kind of text is the following") and **Instruction generation** (answering "What is a relevant writing prompt or edit instruction for text"). The answer to the second question is the generated instruction.

- **Target generation:** Given the source text and the generated instructions, we generate the model outputs with a general purpose LLM and filter them in a post-processing step (see Section 3.1.3).

#### 3.1.3. HEURISTIC POST-PROCESSING

In order to improve the quality of the instruction datasets, we do the following post-processing: (1) In general, rewriting should preserve the overall meaning of the text, and thus, we employ Natural Language Inference (NLI; See Section 5.1) to detect "hallucinations" from the source to the target text and vice versa. If the "hallucination" is in the target text and fixable using simple heuristic rules, we remove the "hallucination" from the target text and keep the instance. (2) For any other detected "hallucination", we filter the instance. (3) If the difference between the source and target texts is unexpectedly small, we also filter the instance.

### 3.2. Modeling

**Supervised Fine-Tuning (SFT)**. Given a pretrained language model $M_{\text{base}}$, we fine-tune it using the instruction tuning dataset discussed in Section 3.1, producing a model $M_{\text{SFT}}$. We employ the decoder-only Transformer architecture for our experiments, details of which are explained in Section 5. For both models, the input is formed by concatenating `<instruction>` and `<source>` with a newline, while the output is `<target>`.

**Reward Modeling (RM)**. We leverage human preference and synthetic preference datasets to train our reward models. The synthetic preference dataset is generated as follows:

- **Collecting Model Responses**. We sample `<instruction>-<source>` pairs from the synthetic instruction dataset (Section 3.1.2) and obtain outputs `<target>` from both $M_{\text{SFT}}$ and $M_{\text{base}}$.

- **Heuristic Ranking**. To rank the outputs of the models, we employ the methodology outlined in Section 3.1.3. We utilize NLI scores to assess content preservation and the degree of hallucination in the generated text. Additionally, we calculate the edit distance between `<source>` and `<target>`, considering small edit distances as penalties. Furthermore, we consider the length ratio between `<target>` and `<source>` to determine whether there has been an appropriate text expansion or compression. Specifically, for tasks that require adding new information, the length ratio should be greater than 1, whereas for tasks that emphasize conciseness, the length ratio should be less than 1. The target with a higher rank is labeled as the "good target" ($r_{\text{good}}$), while the other is referred to as the "bad target" ($r_{\text{bad}}$).

The reward model employs a transformer-based architecture with a linear regression output layer. The input consists of concatenated `<instruction>`, `<source>`, and `<target>` tokens, while the output is a scalar score. We use a good target $r_{\text{good}}$ and a bad target $r_{\text{bad}}$ for comparison

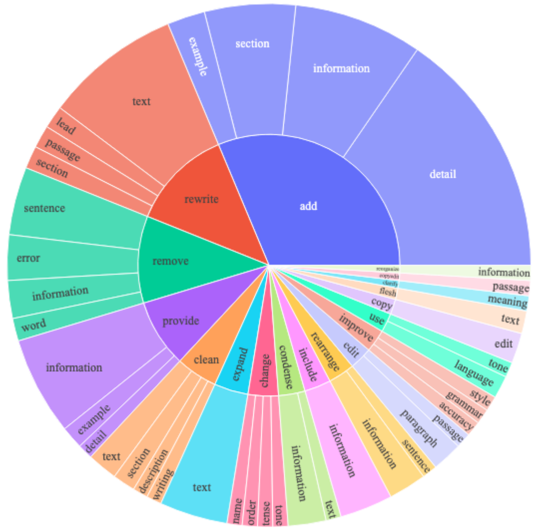



*Figure 3.* The plot of instructions in OPENREWRITEEVAL.

during training. The reward model scores both inputs, and computes the predicted score difference $\hat{r} = \sigma(r_{\text{good}} - r_{\text{bad}})$. The training loss is the entropy of the normalized score difference $\hat{r}$.

**Reinforcement Learning**. Finally, we further optimize the supervised fine-tuned model $M_{\text{SFT}}$ by employing reinforcement learning (Ouyang et al., 2022), guided by the scores provided by the fine-tuned reward model $R_{\text{SFT}}$. This process results in the final model, $M_{\text{rewrite}}$.

## 4. Evaluation Framework

### 4.1. OpenRewriteEval — A New Benchmark for Text Rewriting

To facilitate the evaluation of open-ended rewriting, we have curated a new dataset called OPENREWRITEEVAL, which focuses on open instructions, long-form text, and large edits. Each example in the dataset consists of a three-tuple (`<instruction>`, `<source>`, `<target>`).

OPENREWRITEEVAL consists of six datasets $D_{\text{Formality}}$, $D_{\text{Paraphrase}}$, $D_{\text{Shorten}}$, $D_{\text{Elaborate}}$, $D_{\text{MixedWiki}}$ and $D_{\text{MixedOthers}}$. See Table 1 and Figure 3 for more details about dataset size, data source, and instruction examples. For $D_{\text{Formality}}$, $D_{\text{Paraphrase}}$, and $D_{\text{Shorten}}$, we use a fixed set of instruction. For the rest of the datasets, we asked human annotators to attach appropriate instructions to each source text and then rewrite them accordingly. Appendix A.2 provides detailed guidelines for the rewrite annotations. Table 5 in Appendix A.4 provides information on the size of each task and the average word-level lengths of instructions, source text, and target text.

### 4.2. Automatic Evaluation Metrics

We employ various metrics to evaluate the model's performance including

- **NLI** (Bowman et al., 2015) and **Reversed NLI** (*i.e.*, reverse the premise and the hypotheses) score over the source-prediction pair. NLI and Reversed NLI scores illustrate the model prediction's content presentation and factuality quality. We use the off-the-shelf NLI predictor introduced by (Honovich et al., 2022).
- **Edit Distance Raito (Edit Ratio)**. Edit distance (Ristad & Yianilos, 1998) measures the word-level textural difference between two pieces of text. We report the relative edit distance between the prediction and source text, *i.e.*, dividing the edit distance by the length of the source text. The edit ratio represents the proportion of the source text that has been modified. It is undesirable if the edit distance is small because this indicates the prediction is primarily identical to the source text. Ideally, we expect to see this value to be neither excessively high (indicating the entire content has been changed) nor excessively low (indicating that only minor rewriting occurred thereby diminishing the perceived effectiveness of the system).
- **SARI** (Xu et al., 2016) is an n-gram based metric measures how a close a prediction is relative to the source text and the reference text by rewarding words added, kept, or deleted. SARI computes the arithmetic mean of n-gram F1-scores for each of the three operations.
- **GLEU** (Napoles et al., 2015) measures the precision of the n-grams in the model's prediction that match the reference. It is a variant of BLEU (Papineni et al., 2002).

| Dataset | Size | Data Source | Instruction Examples |
|---|---|---|---|
| $D_{\text{Formality}}$ | 200 | See Appendix A.1 | Too conversational, rephrase it to be more formal?<br>Make the text more formal.<br>Rephrase it to be more formal? |
| $D_{\text{Paraphrase}}$ | 102 | See Appendix A.1 | Paraphrase this.<br>Reword this text.<br>Use different wording. |
| $D_{\text{Shorten}}$ | 102 | See Appendix A.1 | Make wording more concise.<br>Improve accuracy, clarity, and conciseness of language.<br>Rephrase for clarity and conciseness. |
| $D_{\text{Elaborate}}$ | 102 | See Appendix A.1 | Elaborate on advantages of JavaScript.<br>Add more details about fighting styles.<br>Describe more about what the third page does. |
| $D_{\text{MixedWiki}}$ | 606 | Wiki | Attempt to make the text sound less like an advertisement.<br>Change to have a consistent past tense throughout the paragraph.<br>Rewrite text in the present tense.<br>Give a detailed and concise description of the Wollyleaf bush.<br>Rewrite for clarity and encyclopedic tone. |
| $D_{\text{MixedOthers}}$ | 517 | C4, Human | Make it more personal and friendly.<br>Rewrite to haiku.<br>Change the name to Horton Beach throughout the text.<br>Make it more motivational for parents of age 50.<br>Create bullet points from text. |
| All | 1629 | | |

*Table 1.* The data statistics and instruction samples of OPENREWRITEEVAL.

GLEU is customized to penalize only the changed n-grams in the targets, as unmodified words do not necessarily need to be penalized in the rewriting task.

- **Update-ROUGE (Updated-R)** (Iv et al., 2022) measures the recall of n-grams between the model's prediction and the references. It is a modified version of ROUGE (Lin & Hovy, 2003). Updated-R specifically computes ROUGE-L on the updated sentences rather than the full text.

When evaluating quality, it is desirable to have a higher value of NLI. Additionally, a higher Edit Ratio within a reasonable range is preferred. However, it's important to note that considering these metrics independently is insufficient. In some cases, predictions with a low edit ratio may still have high NLI scores. Conversely, a large edit ratio can contain hallucinations if the NLI scores are low. Additionally, higher values of SARI, GLEU, and Update-ROUGE indicate that the predictions are more similar to the gold reference text.

**Automatic Side-by-Side Evaluation (AutoSxS)**. We also employ the largest PaLM 2 model to automatically decide which of two given models perform better at following the given instructions. We present a few exemplars to the PaLM 2 model, instructing it on how to select a particular model and providing reasons for choosing that specific model. See Appendix A.5 for prompting details.

## 5. Experiments and Results

This section provides an overview of our experimental settings, baselines, and result analysis. Detailed information about the hyperparameters can be found in Appendix A.3.

### 5.1. Baselines

We use the following baseline models for quality comparison in the later sections:

- **PaLM** (Chowdhery et al., 2022) is a large, densely activated transformer-based language model that can generate text in an open-ended fashion.
- **PaLM 2** (Passos et al., 2023), is an advanced language model which surpasses its predecessor PaLM in terms of multilingual and reasoning abilities while being more computationally efficient. It is a Transformer-based model that underwent training using a blend of objectives.
- **LLaMA** (Touvron et al., 2023) is an efficient, open-source foundation language model.
- **Flan-PaLM** (Chung et al., 2022) is fine-tuned on a large variety of tasks and chain-of-thought data using PaLM as the base model.
- **Alpaca** (Taori et al., 2023) is a language model that

| | | Edit Ratio | NLI (s-p) | NLI (p-s) | SARI | GLEU | Update-R |
|---|---|---|---|---|---|---|---|
| Pretrained LLMs | | | | | | | |
| PaLM (Chowdhery et al., 2022) | 62B | 0.31 | 0.25 | 0.11 | 28.24 | 0.74 | 11.99 |
| PaLM 2 (Passos et al., 2023) | M | **1.22** | 0.63 | 0.37 | 28.62 | 0.48 | 8.14 |
| LLaMA (Touvron et al., 2023) | 65B | 0.71 | 0.83 | 0.83 | 27.98 | 2.10 | 21.35 |
| Instruction-Tuned LLMs | | | | | | | |
| Alpaca (Taori et al., 2023) | 13B | 0.11 | 0.90 | 0.85 | 36.12 | 6.81 | 34.88 |
| Vicuna (Chiang et al., 2023) | 13B | 0.23 | 0.89 | 0.77 | 39.05 | 6.84 | 33.31 |
| Flan-PaLM (Chung et al., 2022) | 62B | 0.12 | 0.58 | 0.42 | 24.52 | 1.87 | 6.23 |
| RewriteLMs | | | | | | | |
| Rewrite-PaLM | 62B | 0.14 | 0.88 | 0.76 | 37.02 | 7.40 | 36.68 |
| Rewrite-PaLM 2 | M | 0.25 | 0.93 | 0.79 | 40.92 | **9.64** | 39.36 |
| Rewrite-RL-PaLM 2 | M | 0.27 | 0.94 | 0.81 | **40.97** | 9.43 | 39.36 |
| Rewrite-RL$_{r/w}$-PaLM 2 | M | 0.29 | **0.96** | **0.87** | 40.66 | **9.64** | **40.10** |

*Table 2.* Model Performance on OpenRewriteEval. Edit distance ratio (Edit Ratio) between the model prediction and the source text; NLI score with source as premise and model prediction as hypothesis (NLI s-p) and vice versa (NLI p-s); SARI, GLEU and Updated-ROUGE (Updated-R) between the gold reference and the model prediction are reported here.

is fine-tuned from LLaMA using 52,000 instruction-following demonstrations.

- **Vicuna** (Chiang et al., 2023) is an open-source chatbot trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT[1].

We follow the same zero-shot prompt setting for all the baseline models as Schick et al. (2022). The pre-trained models without any instruction tuning generally exhibit slightly lower performance in following instructions compared to the instruction tuned models under zero-shot scenario.

### 5.2. Results on OpenRewriteEval Benchmark

The automatic evaluation results for the OpenRewriteEval dataset are presented in Table 2. Rewrite-PaLM and Rewrite-PaLM 2 are supervised fine-tuned versions (as discussed in Section 3.2) based on PaLM, and PaLM 2, respectively. Rewrite-RL-PaLM 2 and Rewrite-RL$_{r/w}$-PaLM 2 are reinforcement learning models tuned over Rewrite-PaLM 2. The reward model from the former does not use our synthetic preference dataset (as discussed in Section 3.2), whereas the reward model from the latter incorporates it.

As shown in Table 2, our RL tuned model Rewrite-RL$_{r/w}$-PaLM 2 has the highest scores in almost all the metrics (i.e., NLI scores, SARI, GLEU, and Update-R). This indicates that our model is good at generating outputs faithful to the original input, while other models might generate more "hallucinations". For edit ratio, Rewrite-RL$_{r/w}$-PaLM 2 has a better score than all the models except PaLM 2. Pre-trained models such as PaLM 2 without any instruction tuning are prone to generating "hallucinations", resulting in a significantly high edit ratio score (*i.e.*, 1.22). Therefore, our model is good at keeping all the essential content and meaning of the source text, while also being able to rewrite with varied language and structures. Given that Rewrite-RL$_{r/w}$-PaLM 2 consistently outperforms Rewrite-RL-PaLM 2 across nearly all metrics, this strongly suggests the effectiveness and value of employing synthetic preference data.

See Appendix A.6 for more metrics on OpenRewriteEval dataset (see Table 9) and a breakdown by each subgroup (see Tables 10, 11, 12, 13, 14, 15).

### 5.3. Results on EditEval

We also evaluated the performance of our models using the publicly available sentence-level rewrite benchmark EditEval[2] (Dwivedi-Yu et al., 2022). This benchmark comprises various datasets that cover different language tasks. Specifically, JFL (Napoles et al., 2017) focuses on language fluency; TRK (Xu et al., 2016) and AST (Alva-Manchego et al., 2020) target at sentence simplification; WNC (Pryzant et al., 2020) addresses text neutralization; FRU (Iv et al., 2022)) and WFI (Petroni et al., 2022) involve updating information that requires external references. More data statistics for each dataset can be found in Table 7.

We only report the results on EditEval datasets that containing more than 100 test examples (see Table 3). The results of LLM baselines and the Copy baseline (which treats the source text as the prediction) are taken directly from the EditEval paper (Dwivedi-Yu et al., 2022). We can observe that the zero-shot performance of Rewrite-PaLM 2 and Rewrite-RL$_{r/w}$-PaLM 2 is mostly on par with or better than the best baselines (*i.e.*, PEER-11 and InsGPT). While our model is specifically designed for long-form text rewriting, it does not sacrifice its capability to handle sentence-level rewriting tasks.

[1]https://sharegpt.com/

[2]https://github.com/facebookresearch/EditEval

| | | JFL | | TRK | AST | WNC | FRU | | WFI | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | SARI | GLEU | SARI | SARI | SARI | SARI | Update-R | SARI | Updated-R |
| Copy | - | 26.7 | 40.5 | 26.3 | 20.7 | 31.9 | 29.8 | 0 | 33.6 | - |
| Tk (Wang et al., 2022b) | 3B | 31.8 | 39 | 32.8 | 29.9 | 31.3 | 12.6 | 3.6 | 1.3 | 4.5 |
| T0 (Sanh et al., 2022) | 3B | 42 | 38.8 | 34.4 | 32.3 | 22.3 | 14.2 | 9.6 | 5.1 | 16.3 |
| T0++ (Sanh et al., 2022) | 11B | 34.7 | 43.2 | 32.9 | 28.2 | 29.3 | 12.6 | 3.7 | 4.4 | 8.1 |
| PEER-3 (Schick et al., 2022) | 3B | 55.5 | 54.3 | 32.5 | 30.5 | 53.3 | 39.1 | 30.9 | 34.4 | 18.7 |
| PEER-11 (Schick et al., 2022) | 11B | 55.8 | 54.3 | 32.1 | 29.5 | **54.5** | **39.6** | 31.4 | **34.9** | 20.4 |
| OPT (Zhang et al., 2022b) | 175B | 47.3 | 47.5 | 32.6 | 31.8 | 31.2 | 35.9 | 27.3 | 26.7 | 11.2 |
| GPT-3 (Brown et al., 2020) | 175B | 50.3 | 51.8 | 33 | 30.5 | 31.7 | 36 | 21.5 | 27.2 | 10.6 |
| InsGPT (Ouyang et al., 2022) | 175B | **61.8** | **59.3** | 38.8 | 38 | 35.4 | 36.3 | 24.7 | 23.6 | 16.1 |
| PaLM 2 (Passos et al., 2023) | M | 36.07 | 2.18 | 34.32 | 35.92 | 25.2 | 24.28 | 26.39 | 11.41 | 20.42 |
| Rewrite-PaLM 2 (Ours) | M | 56.95 | 40.38 | 40.81 | **42.11** | 37.11 | 37.51 | 53.54 | 26.55 | 47.06 |
| Rewrite-RL$_{r/w}$-PaLM 2 (Ours) | M | 55 | 22.89 | **40.87** | 41.71 | 37.81 | 38.56 | **53.93** | 29.25 | **49.53** |

*Table 3.* Model Performance on EditEval (Dwivedi-Yu et al., 2022).

### 5.4. Results on Side-by-Side Evaluation

In this section, we presents the results from AutoSxS (see Section 4.2) analysis. More details are provided in Appendix A.5. Table 4 displays the average win rate of each model against the others and Figure 4 illustrates the detailed pairwise win rate between each pair. As the comparisons can be treated as a series of battles between models, we also employ the Elo rating system, inspired by Bai et al. (2022), to measure and rank the response quality of the models.

| | Mode Size | Win Rate | Elo Rating |
|---|---|---|---|
| PaLM 2 | M | 5.63% | 611 |
| Alpaca | 13B | 57.62% | 1076 |
| Rewrite-PaLM 2 | M | 63.11% | 1116 |
| Rewrite-RL$_{r/w}$-PaLM 2 | M | **66.85%** | **1141** |
| Gold | | 54.58% | 1056 |

*Table 4.* The average win rate and Elo rating from side-by-side evaluation.

Our Rewrite-PaLM 2 outperforms its foundation model PaLM 2 as well as the general-purpose instruction-tuned Alpaca. Applying reinforcement learning further improves its performance.

The side-by-side evaluation also provides us some additional insights. Firstly, the ranking is consistent with the results on automatic metrics. Additionally, the side-by-sides demonstrate the gap between instruction-tuned models and the foundation model in a more direct way. This may indicate that zero-shot prompting on pretrained LLM is not capable of solving the open-ended rewriting task, possibly due to the task's complexity. In addition, from a preliminary revisit of these judgments, we find that the AI judge PaLM 2-L has a strong preference of model responses over human written ones (gold). The fact that AI favors AI outputs while humans favor human outputs is interesting and worthy of future investigation.

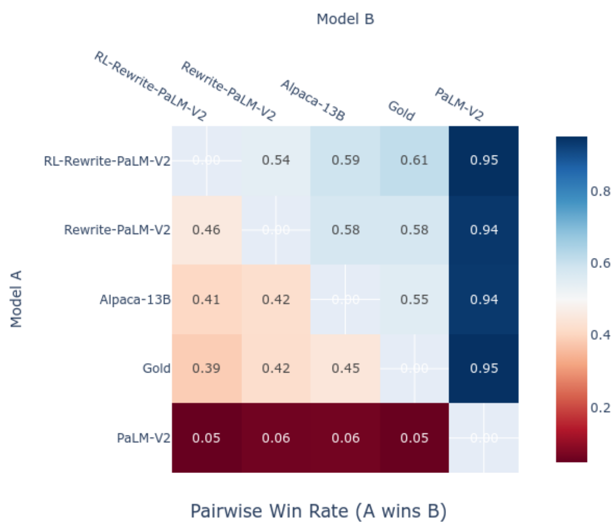



*Figure 4.* Pairwise comparison results. Each cell illustrates the winning rate of Model A over Model B, *i.e.*, the frequency at which PaLM 2-L suggests that Model A outperforms Model B.

## 6. Conclusion

We introduce a novel benchmark for text rewriting with a focus on long form text, covering a wide variety of rewriting types expressed through natural language instructions. We further propose RewriteLM, an instruction-tuned large language model for text rewriting. We present new data generation strategies to reduce human intervention, and verify their effectiveness in our experiments. The automatic metrics and LLM side-by-side results demonstrate that RewriteLM achieves better performance than other pretrained LLMs and instruction finetuned LLMs.

## References


Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., and Specia, L. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4668–4679, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.424.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022, 2003.

Bowman, S., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, 2015.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of NAACL-HLT*, pp. 615–621, 2018.

Dwivedi-Yu, J., Schick, T., Jiang, Z., Lomeli, M., Lewis, P., Izacard, G., Grave, E., Riedel, S., and Petroni, F. Editeval: An instruction-based benchmark for text improvements. arXiv, 2022. doi: 10.48550/ARXIV.2209.13331. URL https://arxiv.org/abs/2209.13331.

Fabbri, A. R., Li, I., She, T., Li, S., and Radev, D. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1074–1084, 2019.

Faltings, F., Galley, M., Hintz, G., Brockett, C., Quirk, C., Gao, J., and Dolan, B. Text editing by command. *arXiv preprint arXiv:2010.12826*, 2020.

Guo, M., Dai, Z., Vrandečić, D., and Al-Rfou, R. Wiki-40b: Multilingual language model dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 2440–2452, 2020.

Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

He, J., Gu, J., Shen, J., and Ranzato, M. Revisiting self-training for neural sequence generation. *arXiv preprint arXiv:1909.13788*, 2019.

Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A., and Matias, Y. True: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pp. 161–175, 2022.

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., and Han, J. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

Huang, L., Cao, S., Parulian, N., Ji, H., and Wang, L. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1419–1436, 2021.

Iv, R., Passos, A., Singh, S., and Chang, M.-W. Fruit: Faithfully reflecting updated information in text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3670–3686, 2022.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Kornilova, A. and Eidelman, V. Billsum: A corpus for automatic summarization of us legislation. *EMNLP-IJCNLP 2019*, pp. 48, 2019.

Lin, C.-Y. and Hovy, E. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pp. 150–157, 2003.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Mallinson, J., Adamek, J., Malmi, E., and Severyn, A. Edit5: Semi-autoregressive text-editing with t5 warm-start. *arXiv preprint arXiv:2205.12209*, 2022.

May, P. Machine translated multilingual sts benchmark dataset. 2021. URL https://github.com/PhilipMay/stsb-multi-mt.

Napoles, C., Sakaguchi, K., Post, M., and Tetreault, J. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 588–593, 2015.

Napoles, C., Sakaguchi, K., and Tetreault, J. Jfleg: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 229–234, Valencia, Spain, April 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E17-2037.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Passos, A., Dai, A., Richter, B., Choquette, C., Sohn, D., So, D., Lepikhin, D. D., Taropa, E., Ni, E., Moreira, E., Mishra, G., Yu, J., Clark, J., Meier-Hellstern, K., Robinson, K., Vodrahalli, K., Omernick, M., Krikun, M., Moussalem, M., Johnson, M., Du, N., Firat, O., Bailey, P., Anil, R., Ruder, S., Shakeri, S., Qiao, S., Petrov, S., Garcia, X., Huang, Y., Tay, Y., Cheng, Y., Wu, Y., Xu, Y., Zhang, Y., and Nado, Z. Palm 2 technical report. Technical report, Google Research, 2023.

Petroni, F., Broscheit, S., Piktus, A., Lewis, P., Izacard, G., Hosseini, L., Dwivedi-Yu, J., Lomeli, M., Schick, T., Mazaré, P., et al. Improving wikipedia verifiability with ai. 2022.

Pryzant, R., Martinez, R. D., Dass, N., Kurohashi, S., Jurafsky, D., and Yang, D. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pp. 480–489, 2020.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., and Yang, D. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023.

Rae, J. W., Potapenko, A., Jayakumar, S. M., Hillier, C., and Lillicrap, T. P. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Rao, S. and Tetreault, J. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 129–140, 2018.

Reif, E., Ippolito, D., Yuan, A., Coenen, A., Callison-Burch, C., and Wei, J. A recipe for arbitrary text style transfer with large language models. *arXiv preprint arXiv:2109.03910*, 2021.

Riley, P., Constant, N., Guo, M., Kumar, G., Uthus, D., and Parekh, Z. Textsettr: Few-shot text style extraction and tunable targeted restyling. *arXiv preprint arXiv:2010.03802*, 2020.

Ristad, E. S. and Yianilos, P. N. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532, 1998.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Le Scao, T., Raja, A., et al. Multitask prompted training enables zero-shot task generalization. In *ICLR 2022-Tenth International Conference on Learning Representations*, 2022.

Schick, T., Dwivedi-Yu, J., Jiang, Z., Petroni, F., Lewis, P., Izacard, G., You, Q., Nalmpantis, C., Grave, E., and Riedel, S. Peer: A collaborative language model. *arXiv preprint arXiv:2208.11663*, 2022.

Sharma, E., Li, C., and Wang, L. Bigpatent: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2204–2213, 2019.

Shazeer, N. and Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604. PMLR, 2018.

Siddique, A., Oymak, S., and Hristidis, V. Unsupervised paraphrasing via deep reinforcement learning. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1800–1809, 2020.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Tikhonov, A., Shibaev, V., Nagaev, A., Nugmanova, A., and Yamshchikov, I. P. Style transfer for texts: Retrain, report errors, compare with rewrites. 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)*, pp. 3936–3945, 2019.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022a.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., et al. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2022b.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698, 2020.

Xu, W., Ritter, A., Dolan, W. B., Grishman, R., and Cherry, C. Paraphrasing for style. In *Proceedings of COLING 2012*, pp. 2899–2914, 2012.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016.

Zhang, H., Song, H., Li, S., Zhou, M., and Song, D. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*, 2022a.

Zhang, R. and Tetreault, J. This email could save your life: Introducing the task of email subject line generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 446–456, 2019.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022b.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Zhang, Y., Ge, T., and Sun, X. Parallel data augmentation for formality style transfer. *arXiv preprint arXiv:2005.07522*, 2020.

# A. Appendix

## A.1. OPENREWRITEEVAL Data

Table 5 provides information on the size of each task and the average word-level lengths of instructions, source texts, and target texts. Our source texts are long-form, and the length ratio (Len Ratio) represents the average and standard deviation of the target's length compared to the source text's length. The formality, paraphrase, and open-ended tasks have average length ratios around 1, indicating similar lengths between the source and target texts. The elaborate task aims to expand the source text, resulting in a higher length ratio. Conversely, the shorten task aims to make the source text more concise, resulting in a length ratio smaller than 1. The edit distance metric shows the word-level edits between the source and target texts. We report the fraction of edit distance over the source text's length (edits over source). Most tasks exhibit edit ratios higher than 50%, indicating that our benchmark dataset involves substantial edits.

| | | # of words | | | | | | NLI | |
|---|---|---|---|---|---|---|---|---|---|
| | Size | Inst | Src | Tar | Len Ratio | Edit Dist | Edit Ratio | src-tar | tar-src |
| All | 1629 | 6.39 | 129.02 | 141.35 | 1.14 | 90.34 | 0.72 | 0.94 | 0.95 |
| $D_{\text{Formality}}$ | 200 | 5.04 | 84.62 | 101.47 | 1.28 | 58.83 | 0.70 | 0.87 | 0.98 |
| $D_{\text{Paraphrase}}$ | 102 | 3.00 | 211.02 | 195.97 | 1.00 | 121.20 | 0.54 | 1.00 | 1.00 |
| $D_{\text{Shorten}}$ | 102 | 4.49 | 211.02 | 165.68 | 0.80 | 72.20 | 0.37 | 1.00 | 1.00 |
| $D_{\text{Elaborate}}$ | 102 | 8.64 | 211.02 | 378.47 | 2.07 | 234.33 | 1.34 | 0.92 | 1.00 |
| $D_{\text{MixedWiki}}$ | 606 | 7.54 | 103.30 | 97.57 | 0.98 | 65.36 | 0.64 | 0.94 | 0.92 |
| $D_{\text{MixedOthers}}$ | 517 | 6.17 | 127.80 | 145.74 | 1.18 | 100.89 | 0.82 | 0.95 | 0.95 |

*Table 5.* Statistics of OPENREWRITEEVAL include the number of examples (Size); average number of words (# of words) of instructions (Inst), source texts (Src), and target texts (Tar); the average length fraction (Len Ratio) between the target and source text; average edit distance (Edit Dist) between source and target; and the fraction between the edit distance and source text (Edit Ratio) for the full set and the subtasks(formality, paraphrase, shorten, elaborate and open). All are measured at the word-level. NLI (src-tar, tar-src) are the NLI scores between the source text and the gold reference.

**Frequent Words**. To illustrate the open-ended instruction statistics, we cluster the open-ended instruction by LDA topic model (Blei et al., 2003)[3] and report top-10 frequent words in the 10 topics in the Table 6. Our open-ended instructions cover a wide range of rewrite requirements. Besides the top words, the rewrite benchmark includes surprise instructions like "haiku", "Shakespeare", Etc.

**Data Sources**. The source texts for the $D_{\text{Formality}}$, $D_{\text{Paraphrase}}$, $D_{\text{Shorten}}$, and $D_{\text{Elaborate}}$ categories are from various datasets, including Multi-News (Fabbri et al., 2019), Wikipedia (Guo et al., 2020), PG-19 book (Rae et al.), BIGPATENT (Sharma et al., 2019), BillSum (Kornilova & Eidelman, 2019), government reports (Huang et al., 2021), scientific papers (Cohan et al., 2018), Enron email (Zhang & Tetreault, 2019), Reddit (Hamilton et al., 2017), IMDB, and Yelp reviews (Maas et al., 2011; Zhang et al., 2015). The $D_{\text{MixedWiki}}$ have the source texts from Wikipedia (Guo et al., 2020) and $D_{\text{MixedOthers}}$ contains C4 (Raffel et al., 2020) and human written ones.

## A.2. Human Rewrite Guideline

- Raters align source text to the instruction, and then rewrite. If the source text is already met the instruction, for example, "make it more formal", then treat the source text as target text and rewrite less formal (put at source side).
- Ensure (1) the content preservation between source and rewrite; (2) maximum word change; and (3) source and target texts are well aligned with instruction. For example, if the instruction is to "make it more polite", then ensure that the target text is much more polite than the source text."
- Elaborate: the rewrite matches source text's tone and format. Add more relevant information and ideas, but do not make up facts.
- Rephrase: the rewrite matches source text's tone, verbosity, format and max changes to existing words.
- Shorten: the rewrite matches source text's tone and format, trims unnecessary words, simplifies sentences, makes them more concise.

[3] https://radimrehurek.com/gensim/

| Topic | Top-frequent Words |
|---|---|
| 1 | expand, easy, text, make, clear, sure, understand, idea, post, reader |
| 2 | use, make, concise, active, voice, copy, edit, points, write, polite |
| 3 | technical, elaborate, make, job, accessible, add, details, expand, idea, audience |
| 4 | less, add, formal, tone, table, change, contents, detail, make, sound |
| 5 | make, concise, personal, persuasive, positive, friendly, text, person, list, tone |
| 6 | make, engaging, rewrite, polite, accessible, general, audience, sound, objective, text |
| 7 | add, details, conclusion, action, call, product, headline, job, person, make |
| 8 | write, prose, language, create, points, tone, polite, use, objective, formal |
| 9 | change, add, tense, past, examples, present, statistics, tone, formal, table |
| 10 | write, style, add, section, formal, list, engaging, personal, job, product |

*Table 6.* Open-ended Instruction Top-10 words in 10 topics.

- Informal-to-Formal: Rewrite the given paragraph so that it is more formal in style. To make the text more formal, try to: (1) Replace informal words associated with chatty spoken styles (such as slang and contractions) with more formal vocabulary. (2) Make the text impersonal: avoid referring directly to the author(s) or reader(s), or expressing subjective opinions. (3) Use strictly standard grammatical forms.

- Formal-to-Informal: Rewrite the given paragraph so that it is less formal in style. To make your writing less formal, try to: (1) Replace long or uncommon words with relaxed, everyday terms. You may include contractions (such as changing “cannot” to “can’t” if it helps the text flow better. (2) Where appropriate, identify the author and the reader to make the text more relatable. (For example, you might be able to change “It is believed that...” to “I think tha...”) (3) If a sentence is very long or stiffly phrased, try breaking it up or rearranging it, even if this doesn’t fit the strictest rules of standard grammar.

### A.3. Hyper-parameter Setting

We use $64$ Tensor Processing Units (TPU) V3 chips for fine-tuning. The batch size is 32, and the maximum training step is $5000$. We use the Adafactor optimizer (Shazeer & Stern, 2018) with a learning rate of $0.003$. Both the input and output sequence lengths are set to $1024$ tokens. The training dropout rate is $0.1$. During inference, the temperature is set to $0.5$, and the top-K value is $40$.

### A.4. EditEval Data

Table 7 shows the EditEval (Dwivedi-Yu et al., 2022) Data statistics.

| Task | Dataset | Abbrev. | Size |
|---|---|---|---|
| Fluency | JFLEG | JFL | 747 |
| Simplification | ASSET | AST | 359 |
| Simplification | TurkCorpus | TRK | 359 |
| Neutralization | WNC | WNC | 1000 |
| Updating | FRUIT | FRU | 914 |
| Updating | WAFER-INSERT | WFI | 4565 |

*Table 7.* EditEval Dataset Statistics

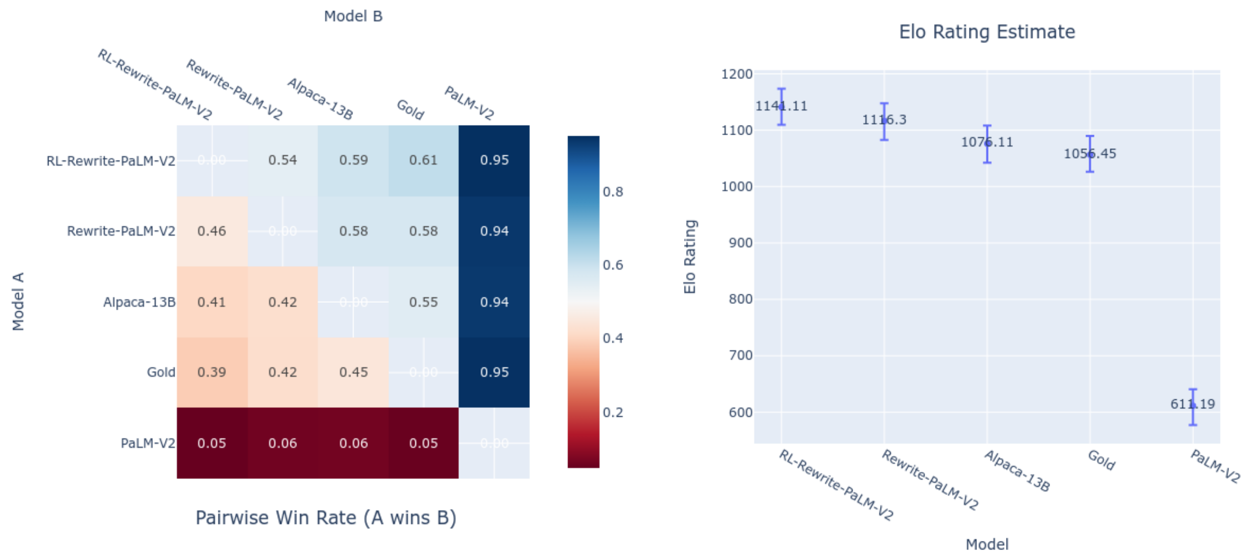
### A.5. Side-by-Side Evaluation (AutoSxS)

To obtain the judgments, for each evaluation example, we prompt PaLM 2-L, the largest one of PaLM 2 family, with few-shot examples as the format showing in Table 8, presenting it with a concatenated input comprising the prompt, the evaluation example, and responses from two different models (A and B). PaLM 2-L would then identify the differences between the responses and determine a winner (A, B, or the same) and provide the explanation. We sample 8 judgments for each input, and 8 more by swapping the positions of response A and B to eliminate the potential of ordering biases.

| **AutoSxS prompts** | |
|---|---|
| task | Judge responses, choose among three choices: (A), (B), or same. |
| instruction | Make this text sound more formal while preserving its meaning. Under the New START Treaty, every deployed heavy bomber with nuclear arms counts as 1 nuclear warhead. This is true if the bomber has cruise missiles or gravity bombs. Neither the US or Russia deploys nuclear arms on their bombers daily. Because the treaty has to count, and reduce, actual warheads carried by deployed delivery cars, the bomber weapons not used daily aren't included. Also, because the parties will use on-site inspections to confirm the number of deployed arms on deployed delivery cars, and the bombers will have no warheads on them during inspections, the parties needed to come up w/ an arbitrary number to assign to the bombers. The number is 1. |
| response (A) | Under the New START Treaty, each deployed heavy bomber armed with nuclear weapons counts as one nuclear warhead, whether the bomber has cruise missiles or gravity bombs. Neither the United States nor Russia deploys nuclear weapons on their bombers on a day-to-day basis. Because the treaty must count, and reduce, actual warheads carried by deployed delivery vehicles, the bomber weapons not used on a day-to-day basis are not counted. Also, because the parties will use on-site inspections to confirm the number of deployed warheads on deployed delivery vehicles, and the bombers will have no warheads on them during inspections, the parties needed to come up with an arbitrary number to assign to the bombers. That number is one. |
| response (B) | In accordance with the New START Treaty, each heavy bomber equipped with nuclear weapons is considered as representing one nuclear warhead, regardless of the presence of cruise missiles or gravity bombs. Neither the United States nor Russia deploy nuclear arms on their bombers on a daily basis. To accurately account for and reduce the actual number of deployed warheads, the treaty only considers weapons that are carried by deployed delivery vehicles. To verify the number of deployed arms, the parties will conduct on-site inspections, during which the bombers will not have any warheads on them. As a result, the parties have agreed on an arbitrary number of one to represent the bombers for treaty purposes. |
| explanation | Response (B) rewrites the context in a formal tone. Response (B) preserves the meaning of context. Response (B) is readable. Response (B) matches the tone of the context. Response (A) rewrites the context in a formal tone. Response (A) preserves the meaning of the context. Response(A) is entirely readable. Response (A) matches the tone of the context well. |
| choose (A), (B) or same | same |

*Table 8.* The prompt format used in AutoSxS.

The results of these pairwise comparisons are illustrated in Figure 5(a). The system measures the relative strength of a "player" compared to others in certain "matches", which is widely used in competitive games like chess and Go. We randomly shuffle all the judgments and treat them as a sequence of "matches" between models. All the models have an initial rating of 1000 and it would increase or decrease when winning or losing a match. Figure 5(b) present the resulting Elo ratings.




(a) Pairwise Comparison Results.

(b) Elo Rating Visualization.

*Figure 5.* The evaluation results judged by PaLM 2-L. (a) Pairwise Comparison Results. Each cell illustrates the winning rate of Model A over Model B, *i.e.*, the frequency at which PaLM 2-L suggests that Model A outperforms Model B. (b) Elo Rating Visualization. The graph exhibits the Elo rating of each model with 97.5% confidence intervals, computed with 1000 rounds of bootstrap iterations on randomly shuffled judgements.

### A.6. Additional Experimental Results

We present comprehensive results for automatic metrics on the full set and each subtask. Table 9 presents the models' performance on the full set of OPENREWRITEEVAL. Tables 10, 11, 12, 13, 14, and 15 show the performance on the formality, paraphrase, shorten, elaborate, mixed Wiki, and mixed others tasks, respectively.

| All | | | NLI | | | | | ROUGE-L | |
|---|---|---|---|---|---|---|---|---|---|
| | Edit Ratio | Len Ratio | s-p | p-s | SARI | BLEU | GLEU | All | Updated |
| Pretrained LLMs | | | | | | | | | |
| PaLM-8B | 0.27 | 0.97 | 0.30 | 0.12 | 26.13 | 2.46 | 0.62 | 9.78 | 8.62 |
| PaLM-62B | 0.31 | 1.36 | 0.25 | 0.11 | 28.24 | 2.87 | 0.74 | 13.35 | 11.99 |
| PaLM 2-M | 1.22 | 5.87 | 0.63 | 0.37 | 28.62 | 2.07 | 0.48 | 8.43 | 8.14 |
| LLaMA-65B | 0.71 | 4.28 | 0.83 | 0.83 | 27.98 | 11.66 | 2.10 | 25.72 | 21.35 |
| Instruction-Tuned | | | | | | | | | |
| Alpaca-7B | 0.11 | 0.90 | 0.90 | 0.85 | 35.37 | 22.80 | 5.97 | 43.40 | 34.14 |
| Alpaca-13B | 0.11 | 0.92 | 0.90 | 0.85 | 36.12 | 23.45 | 6.81 | 43.95 | 34.88 |
| Vicuna-7B | 0.22 | 1.43 | 0.87 | 0.75 | 38.48 | 15.72 | 6.44 | 34.93 | 32.58 |
| Vicuna-13B | 0.23 | 1.50 | 0.89 | 0.77 | 39.05 | 16.39 | 6.84 | 35.79 | 33.31 |
| Flan-PaLM-62B | 0.12 | 0.68 | 0.58 | 0.42 | 24.52 | 13.45 | 1.87 | 28.87 | 6.23 |
| RewriteLMs | | | | | | | | | |
| Rewrite-PaLM-62B | 0.14 | 1.19 | 0.88 | 0.76 | 37.02 | 25.63 | 7.40 | 46.46 | 36.68 |
| Rewrite-Flan-PaLM-62B | 0.15 | 1.15 | 0.88 | 0.72 | 37.74 | 24.54 | 7.58 | 45.20 | 37.06 |
| Rewrite-PaLM 2-M | 0.25 | 1.61 | 0.93 | 0.79 | 40.92 | 23.56 | 9.64 | 44.06 | 39.36 |
| Rewrite-RL-PaLM 2-M | 0.27 | 1.72 | 0.94 | 0.81 | 40.97 | 23.29 | 9.43 | 43.60 | 39.36 |
| Rewrite-RL$_{r/w}$-PaLM 2-M | 0.29 | 1.91 | 0.96 | 0.87 | 40.66 | 24.55 | 9.64 | 44.85 | 40.10 |

*Table 9.* Model Performance on OPENREWRITEEVAL full set.

| $D_{\text{Formality}}$ | | | NLI | | | | | ROUGE-L | |
|---|---|---|---|---|---|---|---|---|---|
| | Edit Ratio | Len Ratio | s-p | p-s | SARI | BLEU | GLEU | All | Updated |
| Pretrained LLMs | | | | | | | | | |
| PaLM-8B | 0.30 | 0.99 | 0.29 | 0.12 | 23.60 | 2.74 | 0.40 | 8.32 | 6.97 |
| PaLM-62B | 0.41 | 1.75 | 0.24 | 0.14 | 27.50 | 3.50 | 0.81 | 14.06 | 12.11 |
| PaLM 2-M | 1.62 | 7.56 | 0.65 | 0.42 | 27.40 | 2.92 | 0.80 | 8.85 | 7.78 |
| LLaMA-65B | 0.97 | 5.43 | 0.83 | 0.84 | 28.88 | 11.34 | 2.57 | 25.30 | 22.42 |
| Instruction-Tuned | | | | | | | | | |
| Alpaca-7B | 0.09 | 0.92 | 0.98 | 0.90 | 39.69 | 23.13 | 8.75 | 48.22 | 42.94 |
| Alpaca-13B | 0.11 | 0.99 | 0.98 | 0.92 | 41.94 | 23.52 | 10.43 | 48.09 | 44.70 |
| Vicuna-7B | 0.16 | 1.27 | 0.93 | 0.87 | 41.34 | 17.79 | 9.42 | 40.59 | 39.65 |
| Vicuna-13B | 0.19 | 1.47 | 0.95 | 0.89 | 42.04 | 17.41 | 9.24 | 39.57 | 38.61 |
| Flan-PaLM-62B | 0.04 | 0.84 | 0.87 | 0.81 | 23.32 | 30.33 | 6.34 | 52.94 | 5.84 |
| RewriteLMs | | | | | | | | | |
| Rewrite-PaLM-62B | 0.06 | 1.00 | 0.99 | 0.98 | 44.80 | 33.48 | 14.59 | 59.19 | 55.07 |
| Rewrite-Flan-PaLM-62B | 0.05 | 1.00 | 1.00 | 0.98 | 45.63 | 35.91 | 15.06 | 61.50 | 55.81 |
| Rewrite-PaLM 2-M | 0.07 | 1.02 | 0.99 | 0.99 | 52.39 | 37.83 | 23.08 | 62.64 | 60.17 |
| Rewrite-RL-PaLM 2-M | 0.07 | 1.02 | 1.00 | 0.99 | 53.05 | 38.22 | 23.61 | 62.64 | 60.19 |
| Rewrite-RL$_{r/w}$-PaLM 2-M | 0.07 | 1.04 | 1.00 | 0.99 | 52.42 | 38.40 | 23.17 | 62.94 | 60.46 |

*Table 10.* Model Performance on OPENREWRITEEVAL formality category.

| $D_{\text{Paraphrase}}$ | | | NLI | | | | | ROUGE-L | |
|---|---|---|---|---|---|---|---|---|---|
| | Edit Ratio | Len Ratio | s-p | p-s | SARI | BLEU | GLEU | All | Updated |
| Pretrained LLMs | | | | | | | | | |
| PaLM-8B | 0.21 | 0.35 | 0.30 | 0.12 | 25.86 | 1.29 | 0.34 | 5.85 | 4.66 |
| PaLM-62B | 0.27 | 1.07 | 0.28 | 0.18 | 28.18 | 3.84 | 0.31 | 14.24 | 11.09 |
| PaLM 2-M | 0.73 | 3.14 | 0.49 | 0.28 | 28.34 | 1.91 | 0.19 | 8.69 | 8.02 |
| LLaMA-65B | 0.84 | 4.87 | 0.84 | 0.83 | 27.19 | 9.88 | 1.29 | 23.46 | 17.92 |
| Instruction-Tuned | | | | | | | | | |
| Alpaca-7B | 0.10 | 0.77 | 0.98 | 0.93 | 37.38 | 18.76 | 4.25 | 41.28 | 36.27 |
| Alpaca-13B | 0.10 | 0.83 | 0.98 | 0.95 | 39.18 | 21.82 | 6.07 | 44.74 | 39.77 |
| Vicuna-7B | 0.15 | 0.89 | 0.97 | 0.92 | 39.77 | 13.38 | 4.71 | 34.81 | 34.38 |
| Vicuna-13B | 0.16 | 0.99 | 0.97 | 0.91 | 39.63 | 13.15 | 4.75 | 35.12 | 34.46 |
| Flan-PaLM-62B | 0.07 | 0.67 | 0.98 | 0.74 | 25.32 | 24.31 | 3.09 | 44.46 | 6.12 |
| RewriteLMs | | | | | | | | | |
| Rewrite-PaLM-62B | 0.10 | 1.02 | 0.96 | 0.90 | 33.98 | 23.99 | 3.16 | 46.95 | 35.99 |
| Rewrite-Flan-PaLM-62B | 0.09 | 0.90 | 0.96 | 0.87 | 36.16 | 24.56 | 4.98 | 47.31 | 38.53 |
| Rewrite-PaLM 2-M | 0.10 | 1.00 | 0.97 | 0.93 | 39.53 | 23.87 | 5.51 | 47.04 | 43.26 |
| Rewrite-RL-PaLM 2-M | 0.11 | 0.99 | 0.98 | 0.92 | 40.29 | 22.66 | 5.36 | 45.67 | 42.70 |
| Rewrite-RL$_{r/w}$-PaLM 2-M | 0.17 | 1.37 | 0.98 | 0.94 | 40.55 | 22.52 | 5.35 | 45.39 | 41.62 |

*Table 11.* Model Performance on OPENREWRITEEVAL paraphrase category.

| $D_{\text{Shorten}}$ | | | NLI | | | | | ROUGE-L | |
|---|---|---|---|---|---|---|---|---|---|
| | Edit Ratio | Len Ratio | s-p | p-s | SARI | BLEU | GLEU | All | Updated |
| Pretrained LLMs | | | | | | | | | |
| PaLM-8B | 0.22 | 0.34 | 0.29 | 0.08 | 22.51 | 1.21 | 0.60 | 5.14 | 4.37 |
| PaLM-62B | 0.32 | 1.29 | 0.28 | 0.12 | 26.28 | 2.25 | 0.87 | 12.31 | 11.13 |
| PaLM 2-M | 1.12 | 5.21 | 0.63 | 0.35 | 26.21 | 2.49 | 0.33 | 8.92 | 6.92 |
| LLaMA-65B | 0.76 | 4.55 | 0.85 | 0.82 | 27.87 | 14.14 | 3.51 | 28.03 | 21.55 |
| Instruction-Tuned | | | | | | | | | |
| Alpaca-7B | 0.12 | 0.58 | 0.97 | 0.87 | 36.41 | 22.88 | 8.47 | 46.67 | 42.15 |
| Alpaca-13B | 0.12 | 0.65 | 0.97 | 0.95 | 37.38 | 24.32 | 11.14 | 48.26 | 43.42 |
| Vicuna-7B | 0.18 | 0.86 | 0.94 | 0.81 | 34.48 | 13.55 | 7.47 | 35.36 | 34.10 |
| Vicuna-13B | 0.16 | 0.77 | 0.97 | 0.87 | 35.70 | 16.51 | 8.94 | 39.41 | 37.57 |
| Flan-PaLM-62B | 0.09 | 0.57 | 0.93 | 0.59 | 25.98 | 28.72 | 4.84 | 48.27 | 5.67 |
| RewriteLMs | | | | | | | | | |
| Rewrite-PaLM-62B | 0.10 | 0.73 | 0.97 | 0.85 | 37.46 | 32.03 | 11.75 | 54.97 | 44.30 |
| Rewrite-Flan-PaLM-62B | 0.11 | 0.60 | 0.95 | 0.79 | 38.09 | 27.61 | 11.55 | 51.49 | 42.30 |
| Rewrite-PaLM 2-M | 0.12 | 0.65 | 0.97 | 0.82 | 38.55 | 27.11 | 10.61 | 51.75 | 44.84 |
| Rewrite-RL-PaLM 2-M | 0.12 | 0.69 | 0.98 | 0.84 | 38.40 | 26.92 | 10.39 | 51.39 | 44.64 |
| Rewrite-RL$_{r/w}$-PaLM 2-M | 0.16 | 0.94 | 1.00 | 0.92 | 39.50 | 28.99 | 11.84 | 53.11 | 46.75 |

*Table 12.* Model Performance on OPENREWRITEEVAL shorten category.

| $D_{\text{Elaborate}}$ | | | NLI | | | | | ROUGE-L | |
|---|---|---|---|---|---|---|---|---|---|
| | Edit Ratio | Len Ratio | s-p | p-s | SARI | BLEU | GLEU | All | Updated |
| Pretrained LLMs | | | | | | | | | |
| PaLM-8B | 0.21 | 0.33 | 0.30 | 0.15 | 20.88 | 0.85 | 0.30 | 6.16 | 3.79 |
| PaLM-62B | 0.29 | 1.03 | 0.33 | 0.08 | 23.32 | 1.23 | 0.38 | 10.61 | 9.25 |
| PaLM 2-M | 1.24 | 5.85 | 0.72 | 0.40 | 26.58 | 2.32 | 1.09 | 10.79 | 9.46 |
| LLaMA-65B | 0.61 | 3.78 | 0.83 | 0.86 | 28.80 | 11.56 | 3.90 | 29.51 | 17.94 |
| Instruction-Tuned | | | | | | | | | |
| Alpaca-7B | 0.18 | 1.04 | 0.73 | 0.57 | 30.63 | 6.19 | 3.06 | 23.74 | 18.63 |
| Alpaca-13B | 0.18 | 1.09 | 0.72 | 0.62 | 31.67 | 7.81 | 4.72 | 26.01 | 18.73 |
| Vicuna-7B | 0.46 | 2.73 | 0.88 | 0.56 | 31.74 | 5.01 | 2.53 | 24.21 | 18.41 |
| Vicuna-13B | 0.46 | 2.69 | 0.89 | 0.50 | 31.71 | 4.80 | 2.57 | 24.03 | 18.20 |
| Flan-PaLM-62B | 0.16 | 0.24 | 0.73 | 0.26 | 23.00 | 2.31 | 0.54 | 13.42 | 3.29 |
| RewriteLMs | | | | | | | | | |
| Rewrite-PaLM-62B | 0.36 | 2.02 | 0.67 | 0.38 | 29.43 | 6.30 | 3.06 | 26.63 | 16.44 |
| Rewrite-Flan-PaLM-62B | 0.36 | 2.04 | 0.68 | 0.35 | 29.01 | 5.07 | 1.84 | 24.77 | 17.92 |
| Rewrite-PaLM 2-M | 0.70 | 3.84 | 0.93 | 0.53 | 31.55 | 5.66 | 3.11 | 26.17 | 17.58 |
| Rewrite-RL-PaLM 2-M | 0.79 | 4.23 | 0.97 | 0.55 | 32.39 | 5.83 | 3.11 | 26.32 | 18.13 |
| Rewrite-$\text{RL}_{\text{r/w}}$-PaLM 2-M | 0.74 | 4.22 | 0.99 | 0.77 | 33.25 | 8.67 | 3.75 | 30.05 | 20.15 |

*Table 13.* Model Performance on OPENREWRITEEVAL elaborate category.

| $D_{\text{MixedWiki}}$ | | | NLI | | | | | ROUGE-L | |
|---|---|---|---|---|---|---|---|---|---|
| | Edit Ratio | Len Ratio | s-p | p-s | SARI | BLEU | GLEU | All | Updated |
| Pretrained LLMs | | | | | | | | | |
| PaLM-8B | 0.33 | 1.63 | 0.31 | 0.12 | 28.00 | 3.47 | 1.15 | 14.41 | 13.48 |
| PaLM-62B | 0.33 | 1.56 | 0.21 | 0.11 | 27.55 | 3.39 | 1.10 | 13.60 | 12.81 |
| PaLM 2-M | 1.41 | 7.08 | 0.67 | 0.41 | 28.21 | 2.39 | 0.56 | 8.47 | 8.83 |
| LLaMA-65B | 0.68 | 4.18 | 0.84 | 0.87 | 29.04 | 14.44 | 2.69 | 27.54 | 24.44 |
| Instruction-Tuned | | | | | | | | | |
| Alpaca-7B | 0.09 | 0.94 | 0.93 | 0.91 | 35.73 | 31.86 | 8.76 | 50.17 | 38.02 |
| Alpaca-13B | 0.08 | 0.95 | 0.92 | 0.91 | 35.87 | 32.45 | 9.00 | 50.49 | 38.41 |
| Vicuna-7B | 0.20 | 1.43 | 0.90 | 0.86 | 39.34 | 23.11 | 9.51 | 41.00 | 38.09 |
| Vicuna-13B | 0.21 | 1.50 | 0.92 | 0.86 | 39.75 | 23.72 | 10.03 | 41.79 | 39.06 |
| Flan-PaLM-62B | 0.20 | 0.83 | 0.11 | 0.09 | 24.73 | 4.57 | 0.72 | 14.26 | 9.20 |
| RewriteLMs | | | | | | | | | |
| Rewrite-PaLM-62B | 0.09 | 0.98 | 0.95 | 0.84 | 38.54 | 35.43 | 10.20 | 53.95 | 42.27 |
| Rewrite-Flan-PaLM-62B | 0.10 | 0.93 | 0.93 | 0.78 | 39.71 | 33.89 | 10.71 | 52.65 | 42.88 |
| Rewrite-PaLM 2-M | 0.13 | 1.05 | 0.93 | 0.83 | 42.81 | 32.52 | 13.35 | 51.28 | 46.47 |
| Rewrite-RL-PaLM 2-M | 0.14 | 1.08 | 0.93 | 0.84 | 42.50 | 32.41 | 12.80 | 50.79 | 46.38 |
| Rewrite-$\text{RL}_{\text{r/w}}$-PaLM 2-M | 0.15 | 1.21 | 0.94 | 0.87 | 42.93 | 34.42 | 13.48 | 52.25 | 47.12 |

*Table 14.* Model Performance on OPENREWRITEEVAL mixed Wiki category.

| $D_{\text{MixedOthers}}$ | | | NLI | | | | | ROUGE-L | |
|---|---|---|---|---|---|---|---|---|---|
| | Edit Ratio | Len Ratio | s-p | p-s | SARI | BLEU | GLEU | All | Updated |
| Pretrained LLMs | | | | | | | | | |
| PaLM-8B | 0.22 | 0.57 | 0.31 | 0.12 | 26.73 | 1.96 | 0.20 | 7.31 | 6.15 |
| PaLM-62B | 0.27 | 1.10 | 0.28 | 0.10 | 30.71 | 2.28 | 0.41 | 13.34 | 11.86 |
| PaLM 2-M | 0.97 | 4.47 | 0.60 | 0.30 | 30.49 | 1.25 | 0.23 | 7.62 | 7.48 |
| LLaMA-65B | 0.56 | 3.54 | 0.79 | 0.76 | 26.85 | 8.86 | 0.96 | 23.83 | 19.40 |
| Instruction-Tuned | | | | | | | | | |
| Alpaca-7B | 0.11 | 0.88 | 0.85 | 0.75 | 34.03 | 15.82 | 1.78 | 37.27 | 28.52 |
| Alpaca-13B | 0.11 | 0.88 | 0.86 | 0.74 | 33.42 | 16.53 | 2.29 | 38.00 | 27.24 |
| Vicuna-7B | 0.23 | 1.30 | 0.75 | 0.56 | 38.51 | 10.20 | 3.07 | 28.47 | 26.35 |
| Vicuna-13B | 0.23 | 1.36 | 0.78 | 0.59 | 39.05 | 11.35 | 3.21 | 30.11 | 27.42 |
| Flan-PaLM-62B | 0.09 | 0.57 | 0.83 | 0.59 | 24.58 | 14.37 | 0.94 | 32.83 | 3.60 |
| RewriteLMs | | | | | | | | | |
| Rewrite-PaLM-62B | 0.21 | 1.47 | 0.76 | 0.63 | 34.25 | 13.99 | 2.16 | 34.89 | 25.63 |
| Rewrite-Flan-PaLM-62B | 0.22 | 1.46 | 0.78 | 0.59 | 34.35 | 12.40 | 1.90 | 32.54 | 25.44 |
| Rewrite-PaLM 2-M | 0.41 | 2.36 | 0.88 | 0.68 | 36.85 | 10.30 | 2.00 | 29.82 | 25.42 |
| Rewrite-RL-PaLM 2-M | 0.45 | 2.59 | 0.89 | 0.71 | 36.85 | 9.66 | 1.85 | 29.26 | 25.55 |
| Rewrite-$\text{RL}_{\text{r/w}}$-PaLM 2-M | 0.50 | 2.92 | 0.94 | 0.83 | 35.16 | 10.28 | 1.47 | 30.36 | 26.32 |

*Table 15.* Model Performance on OPENREWRITEEVAL mixed others category.