# LangMark: A Multilingual Dataset for Automatic Post-Editing

**Anonymous ACL submission**

## Abstract

Automatic post-editing (APE) aims to correct errors in machine-translated text, enhancing translation quality, while reducing the need for human intervention. Despite advances in neural machine translation (NMT), the development of effective APE systems has been hindered by the lack of large-scale multilingual datasets specifically tailored to NMT outputs. To address this gap, we present and release **LangMark**, a new human-annotated multilingual APE dataset for English translation to seven languages: Brazilian Portuguese, French, German, Italian, Japanese, Russian, and Spanish. The dataset has 206,983 triplets, with each triplet consisting of a source segment, its NMT output, and a human post-edited translation. Annotated by expert human linguists, our dataset offers both linguistic diversity and scale. Leveraging this dataset, we empirically show that *Large Language Models* (LLMs) with few-shot prompting can effectively perform APE, improving upon leading commercial and even proprietary machine translation systems. We believe that this new resource will facilitate the future development and evaluation of APE systems.

## 1 Introduction

Machine translation has become increasingly efficient and effective thanks to the development of ever-larger transformer models (Vaswani, 2017). Recent advances in *Large Language Models* (LLMs) have significantly influenced the field, enabling more fluent and contextually accurate translations (Zhu et al., 2024; Zhang et al., 2023; Li et al., 2024; Briakou et al., 2024). Studies have shown that LLMs can match or even outperform specialized systems in various Natural Language Processing (NLP) tasks (Radford et al., 2019; Touvron et al., 2023; Wang et al., 2022).

Despite these advances, machine-translated text often still contains errors that require correction to meet the quality standards expected in professional translations. Automatic Post-Editing (APE) aims to automatically correct these errors in MT output, improving translation quality while reducing the need for human intervention (Knight and Chander, 1994). Modern APE models take the source text and machine-translated text as input and produce the post-edited text with the necessary changes as output. We refer to these components as triplets: *source*, *translated*, and *post-edited* segments (see Figure 1).

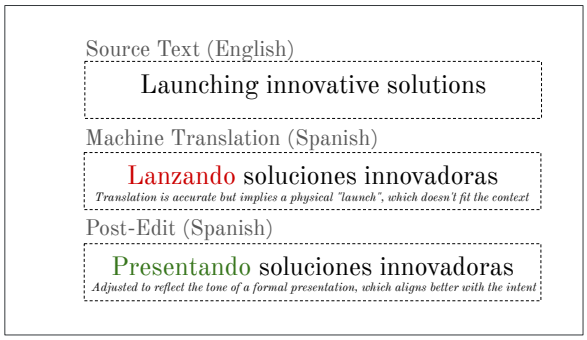


Figure 1: Example of a triplet in an automatic post-editing task.

Recently, automatic post-editing has shown great success on Statistical Machine Translation (SMT) outputs (Junczys-Dowmunt and Grundkiewicz, 2018; Correia and Martins, 2019), even when trained with a limited number of samples. However, even strong APE models face significant challenges (Chatterjee et al., 2019, 2018; Ive et al., 2020) due to the already high quality of modern NMT systems. Junczys-Dowmunt and Grundkiewicz (2018) concluded that the usefulness of "neural-on-neural APE" was minimal, suggesting that the marginal gains may not justify the effort.

However, Chollampatt et al. (2020a) demonstrated that a fine-tuned transformer model has the potential to improve upon the outputs of state-of-the-art NMT systems. Their study introduced the

Table 1: Number of triplets and average *source*, *NMT* and *Post Edited* tokens (tokenized using *tiktoken*[1]) per triplet for all languages in LangMark.

| Locale | Triplets | Tokens Per Triplet (AVG) | | |
|---|---|---|---|---|
| | | Source | NMT | PE |
| EN-DE | 33,308 | 16.12 | 21.73 | 21.72 |
| EN-ES | 32,799 | 16.58 | 20.80 | 21.16 |
| EN-FR | 33,027 | 16.38 | 22.16 | 22.35 |
| EN-IT | 32,512 | 16.42 | 23.47 | 23.71 |
| EN-JP | 28,170 | 15.26 | 26.34 | 27.30 |
| EN-BR | 31,981 | 16.52 | 20.36 | 20.30 |
| EN-RU | 8,648 | 14.90 | 20.40 | 21.23 |

SubEdits dataset, which contains approximately 160,000 triplets but is limited to the English-German language pair. This highlights a gap in the availability of large-scale, multilingual datasets necessary to advance APE research on NMT outputs.

In an effort to address this gap, we introduce **LangMark**; a new multilingual, human-post-edited APE dataset comprising 206,983 triplets from English to seven languages: Brazilian Portuguese (BR), French (FR), German (DE), Italian (IT), Japanese (JP), Russian (RU), and Spanish (ES) (see Table 1). Each triplet consists of a source segment in English, its NMT output, and a human post-edited translation. Labeled by expert linguists, this dataset offers both linguistic diversity and scale, making it, to the best of our knowledge, the largest human-post-edited dataset for APE on NMT outputs.

Leveraging this dataset, we empirically show that LLMs with few-shot prompting can effectively perform APE, improving upon leading commercial and proprietary MT systems. Our experiments highlight the potential of combining large-scale, high-quality datasets with advanced LLMs to enhance translation quality across multiple languages. Moreover, this work examines a critical aspect of APE: the model's capability to discern whether a segment requires editing, which is often overlooked in prior research.

The contributions of this work can be summarized as follows:

1. We present and release **LangMark**, a new, human-annotated, multilingual dataset with over 200,000 triplets across seven languages, that serves as a strong benchmark for APE tasks.

2. Leveraging this dataset, we show that LLMs with few-shot prompting can effectively perform APE to improve upon NMT outputs even from proprietary MT systems.

3. We provide a comprehensive analysis of the dataset and the performance of LLMs on APE tasks, offering insights for future research.

[1] https://github.com/openai/tiktoken

## 2 Related Work

This section reviews previous research on automatic post-editing, focusing on recent advancements involving Large Language Models. We also examine retrieval methods for few-shot in-context learning and discuss relevant datasets used for post-editing tasks.

### 2.1 Automatic Post-Editing

Automatic post-editing aims to automatically correct errors in machine-translated text, improving translation quality without human intervention. A great amount of prior research has focused on the development of neural models for the APE task (Vu and Haffari, 2018; Shterionov et al., 2020; Chatterjee, 2019; Góis et al., 2020; Correia and Martins, 2019; Voita et al., 2019; Chollampatt et al., 2020b; do Carmo et al., 2021). Shterionov et al. (2020) presented a comprehensive roadmap for APE, highlighting challenges and potential directions for future research. Chatterjee (2019) explored the use of deep learning techniques for APE while Góis et al. (2020) investigated the use of automatic ordering techniques to refine translations. Correia and Martins (2019) proposed a simple yet effective neural model for APE using transfer learning, demonstrating promising results.

Voita et al. (2019) introduced a context-aware approach to APE, incorporating source context information into the neural model to generate more accurate post-edits. Chollampatt et al. (2020b) examined the use of LLMs for APE to improve overall translation quality for NMT models, investigating the effects of various factors in the APE task. do Carmo et al. (2021) provided an overview of various techniques and approaches in the field of APE, covering both traditional and neural-based methods. Overall, these studies (and many references therein) have explored different architectures,

learning strategies, and contextual information integration in neural models to improve the quality of post-edited translations.

### 2.2 Leveraging Large Language Models for Post-Editing

There has been growing interest in leveraging LLMs for post-editing. For example, Vidal et al. (2022) explored the use of GPT-3 for post-editing using glossaries, while Raunak et al. (2023) investigated the use of GPT-4 for automatic post-editing of neural machine translation outputs. Their work focuses on rectifying errors in NMT outputs without preliminary quality assessment, aiming to enhance translation quality directly.

Ki and Carpuat (2024) further enhances machine translation by guiding large language models to post-edit MT outputs using fine-grained feedback from error annotations. Their experiments across multiple language pairs demonstrate that both zero-shot prompted and fine-tuned LLMs benefit from this approach, effectively addressing specific translation errors and improving translation metrics.

While these works make significant contributions to the exploration of LLMs for post-editing, they do not constitute a benchmark for evaluating the multilingual post-editing capabilities of LLMs. In contrast, we believe that **LangMark**, coupled with the experiments presented in this paper, can serve as a robust benchmark for this purpose, enabling a more comprehensive assessment of LLM performance across multiple languages.

| Dataset | Lang. | Size | Domain |
|---|---|---|---|
| WMT'18 APE (Chatterjee et al., 2018) | EN-DE | 15K | IT |
| WMT'19 APE (Chatterjee et al., 2019) | EN-RU | 17K | IT |
| WMT'23 APE (Bhattacharyya et al., 2023) | EN-MR | 18K | Mixed |
| QT21 (Specia et al., 2017) | EN-LV | 21K | Life Sciences |
| APE-QUEST (Ive et al., 2020) | EN-NL<br>EN-FR<br>EN-PT | 11K<br>10K<br>10K | Legal |
| SubEdits (Chollampatt et al., 2020a) | EN-DE | 161K | Subtitles |
| eSCAPE (Artificial) (Negri et al., 2018) | EN-DE<br>EN-IT<br>EN-RU | 7.2M<br>3.3M<br>7.7M | Mixed |
| LangMark (this work) | EN-DE<br>EN-ES<br>EN-FR<br>EN-IT<br>EN-JP<br>EN-BR<br>EN-RU | 33.3K<br>32.7K<br>33.1K<br>32.5K<br>28.1K<br>31.9K<br>8.6K | Marketing |

Table 2: Datasets for automatic post-editing on NMT outputs. All but eSCAPE offer human labels.

### 2.3 Datasets for Automatic Post-Editing

Several datasets have been introduced to support the development and evaluation of post-editing methods. Early efforts in APE focused on statistical machine translation (SMT) outputs (Bojar et al., 2015, 2016, 2017). These tasks provided post-edited data on the order of 10,000 to 25,000 triplets. The largest collection of human post-edits on SMT outputs was released by Zhechev (2012), consisting of 30,000 to 410,000 triplets across 12 language pairs. While APE showed impressive gains on SMT datasets (Junczys-Dowmunt, 2017; Tebbifakhr et al., 2018), its performance on neural machine translation (NMT) outputs showed less promising results, with only marginal improvements (Chatterjee et al., 2019).

To improve APE performance on NMT outputs, several studies proposed generating artificial APE data (Junczys-Dowmunt and Grundkiewicz, 2016; Freitag et al., 2019; Specia et al., 2017; Negri et al., 2018; Li et al., 2024) with moderate success. As Neural Machine Translation (NMT) systems get better the required post-edits become more nuanced and thus harder to mimic using artificial data, making human-annotated datasets more valuable.

The WMT APE shared tasks have provided human-annotated datasets (Chatterjee et al., 2018, 2019), but these are relatively small, each comprising less than 20,000 instances. Chollampatt et al. (2020a) introduced the SubEdits dataset, which significantly increased the number of instances to approximately 161,000. However, SubEdits is limited to a single language pair, English to German, lacking multilingual diversity. On the other hand, Negri et al. (2018) proposed a dataset with a much larger volume, but the edits are artificially generated and there are no human annotations involved. Table 2 summarizes previous datasets and their sizes.

These datasets contribute valuable resources for studying post-editing but are limited in linguistic diversity or scale when providing human annotations. In contrast, the dataset featured in this work is a multilingual, human-annotated corpus consisting of translations from English to seven languages, with over 200,000 triplets. To the best of our knowledge, **LangMark** is the largest multilingual, human-annotated dataset for APE on NMT

Table 3: Machine translation performance across languages for different NMT engines on all triplets of the **LangMark** dataset.

| **MT Engine** | **EN-DE** | | **EN-ES** | | **EN-FR** | | **EN-IT** | | **EN-JP** | | **EN-PT** | | **EN-RU** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Metric** | CHRF | TER↓ | CHRF | TER↓ | CHRF | TER↓ | CHRF | TER↓ | CHRF | TER↓ | CHRF | TER↓ | CHRF | TER↓ |
| Google Translate | 73.95 | 42.16 | 79.79 | 27.54 | 76.57 | 33.14 | 79.80 | 28.98 | 62.11 | 78.64 | 83.70 | 21.12 | 64.34 | 53.46 |
| DeepL | 73.03 | 43.15 | 75.01 | 33.70 | 74.74 | 36.27 | 76.96 | 33.05 | 55.26 | 91.52 | 83.93 | 22.68 | 67.74 | 47.41 |
| Microsoft Translator | 75.74 | 40.35 | 80.32 | 27.55 | 76.07 | 34.29 | 82.57 | 25.29 | 62.82 | 84.06 | 84.97 | 20.35 | 64.38 | 54.39 |
| Amazon Translate | 73.70 | 43.13 | 79.01 | 29.78 | 76.27 | 34.42 | 81.66 | 26.52 | 60.93 | 86.62 | 84.27 | 21.96 | 62.65 | 56.00 |
| Proprietary MT (this dataset) | **81.09** | **31.35** | **86.04** | **19.39** | **81.54** | **26.99** | **89.73** | **14.58** | **69.77** | **74.66** | **89.13** | **14.64** | **68.45** | **45.54** |

outputs.

## 3 LangMark Dataset

The absence of large-scale, multilingual, human-annotated corpora for post-editing NMT outputs presents a gap in the resources available for advancing APE research. To address this limitation, we introduce **LangMark**, a new dataset comprising over 200,000 triplets across seven language pairs: English to Japanese (JA), Russian (RU), Brazilian Portuguese (BR), Spanish (ES), French (FR), Italian (IT), and German (DE).

The **LangMark** dataset contains a large number of segments that require models to make nuanced edits, which makes it challenging as a benchmark. Neural Machine Translation (NMT) outputs in the dataset are often technically correct but fail to align with the intended context (see Figure 3). To successfully post-edit these samples the model has to demonstrate contextual understanding.

### 3.1 Dataset Source

The **LangMark** dataset is sourced from various Smartsheet[2] documents, a platform designed for collaborative work management. These documents, which are marketing-related, were first segmented by a translation management system (TMS) into intuitive units (often sentences or short phrases) before translation. This standard industry practice ensures efficient processing, storage, and translation workflows. The triplets were then randomly selected from 967 unique files.

To protect sensitive information, we used Google's dlp[3] tool, specifically designed to identify and remove personally identifiable information (PII) and other sensitive data. We also removed duplicate triplets for each language pair; apart from this preprocessing step, the segments are presented in their original form, reflecting the nature of real-world industry data. We consider this characteristic a positive feature, as it allows the evaluation of model performance on authentic, unaltered data, closely mirroring practical use cases in the industry.

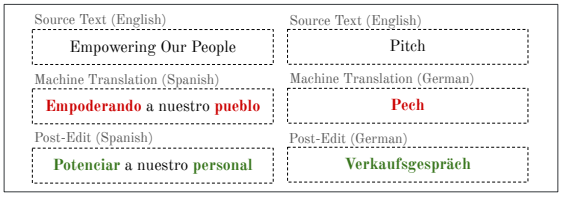


Figure 3: Two triplets from the **LangMark** dataset. These examples illustrate the nuanced nature of the required corrections. While the translations provided by the NMT engine are not inherently incorrect, they are inappropriate given the context of the source material (official marketing documents). For example, "our people" was misinterpreted as "our nation/community" in Spanish, and "pitch" was translated based on the meaning of "tar" in German instead of its intended meaning in a business context.

[2] https://www.smartsheet.com

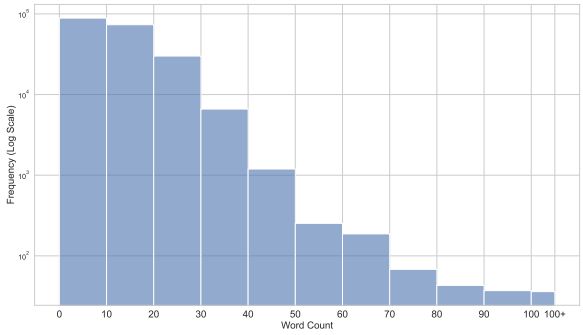


Figure 2: Distribution of word counts for the *source* segments across languages.

### 3.2 Neural Machine Translation

The dataset features neural machine translation (NMT) outputs generated by a proprietary MT system tailored to Smartsheet, along with post-edited translations produced by expert linguists. Because

[3] https://cloud.google.com/dlp

these proprietary machine translation engines are trained on in-domain data, they can be particularly strong in narrow areas, providing high-quality outputs that set a rigorous baseline. This ensures that automatic post-editing (APE) systems are evaluated against a robust benchmark, making any improvements reflective of real-world challenges. Table 3 shows the difference in performance between the NMT comprised in **LangMark** and commercial MT systems.

### 3.3 Dataset Statistics

The dataset comprises a total of 206,983 triplets, from English to seven languages. Each triplet includes a source segment, its corresponding NMT output, and a human post-edited translation.

Figure 2 illustrates the distribution of word counts in the source segments. The frequency distribution shows a natural balance in segment lengths, with most segments being neither excessively short nor too long. This ensures that the dataset captures a realistic range of text complexities.

### 3.4 Linguist Qualifications

We source and deploy linguists with credentials such as degrees in linguistics or translation, native-level fluency in the target language, and strong cultural knowledge—preferably as in-country professionals. All linguists are required to have over five years of industry experience, advanced proficiency in translation tools, and a proactive approach to continuous improvement. Additionally, they must specialize in translating and post-editing content within specific subject matter domains, often with more than three years of expertise in these areas. Following onboarding, linguists receive ongoing support and training to maintain quality, monitored through structured Language Quality Assessments (LQAs). Based on these evaluations, further training or reassignment ensures alignment with project needs. For information on linguist compensations see A.1.

### 3.5 Post-Editing Process

In constructing the dataset, our human post-editors (see Section 3.4), refined the raw NMT output within a Translation Management System (TMS). They made the necessary edits to ensure accuracy, adherence to stylistic and terminology standards, and overall readability, rather than rewriting the translation. The editors have access to glossaries, do-not-translate lists, and any necessary domain-specific materials. Common corrections addressed capitalization, punctuation, spacing, omissions, word order, morphological agreement, locale conventions, and terminology consistency. This process ensures that the final post-edited translations are aligned with client and domain expectations.

## 4 Experimental Setup

To evaluate the performance of the models, we split the dataset into "training" and testing sets, with 90% of the triplets used as potential examples to be retrieved and the remaining 10% reserved for experiments. The split is performed randomly for each language pair, ensuring a proportional representation of all languages.

We adopt this split and retrieval approach because even top-performing LLMs struggle to surpass the proprietary neural machine translation (NMT) engines in this dataset when presented with no context. The nuanced nature of the required edits makes zero-shot approaches insufficient, which motivates the inclusion of in-context examples to guide the model's post-editing decisions. Furthermore, by limiting results to the test set, we make benchmarking on this dataset more affordable for future users. We evaluate all models with 20-shot prompts. For completeness, zero-shot results are provided in the Appendix A.2.

### 4.1 Retrieval

We constructed the retrieval database by embedding the source segments using OpenAI's "text-embedding-3-small" model.[4] Each source segment is stored alongside its corresponding post-edited translation. For retrieval during experiments, the source segment to be post-edited is embedded, and cosine similarity is used to identify the twenty most similar source-human post-edit pairs from the database. Retrieval is conducted within the same language pair, ensuring that no cross-lingual retrieval occurs.

[4] https://platform.openai.com/docs/models/

**System Prompt**

Your input fields are:
1. **source**: The source segment.
2. **pre_translation**: The translation to be edited.
3. **language**: The language to translate to.
4. **translation_pairs**: Similar translation pairs reviewed by experts that MIGHT be relevant. If they are relevant, use them as a reference to guide your translation.

Your output fields are:
1. **reasoning**
2. **answer**: The post-edited translation in JSON format.

All interactions will be structured in the following way:

`[[ ## source ## ]]`
{*source*}

`[[ ## pre_translation ## ]]`
{*pre_translation*}

`[[ ## language ## ]]`
{*language*}

`[[ ## translation_pairs ## ]]`
{*translation_pairs*}

`[[ ## reasoning ## ]]`
{*reasoning*}

`[[ ## answer ## ]]`
{*answer*}

`[[ ## completed ## ]]`

**In adhering to this structure, your objective is:**
You are an expert linguist and translator. You receive both the source text and a translation. Make the necessary changes to the translation. It is possible that the translation doesn't need any changes at all.

Do not translate:
- Variable names (typically camelCase or snake_case)
- Standard technical terms (e.g., "URL", "API", "HTML")
- URLs
- Email addresses

Make sure to preserve the casing (lower, upper case) of the pre-translation.

Return your translation (or the original segment if no translation is required) as a JSON string as follows:
{''translation'': ''translation''}.

**User Prompt**

`[[ ## source ## ]]`
Get clarity

`[[ ## pre_translation ## ]]`
Verschaffen Sie sich Klarheit

`[[ ## language ## ]]`
de-DE

`[[ ## translation_pairs ## ]]`
Clear contents→ Inhalt löschen
⋮
Get the big picture → So behalten Sie den Überblick

Respond with the corresponding output fields, starting with the field 'reasoning', then 'answer', and ending with the marker for 'completed'.

Figure 4: Structure of the few-shot prompting format used for LLMs. If the model's API does not support a system prompt we simply prepend it to the user prompt.

### 4.2 Models and Prompting

We evaluate the performance of both open-source and closed-source models in our experiments. To facilitate this, we leverage the `dspy` library (Khattab et al., 2024, 2022), which integrates with `LiteLLM`[5] to manage API requests to the various models. For open-source models, we utilize HuggingFace endpoints[6] to set up and manage the necessary infrastructure to process requests.

All models are evaluated using the same 20-shot prompting setup. Specifically, for each segment to be post-edited, we include 20 pairs of source segments and their human post-edited version in the prompt. This ensures a uniform evaluation framework across all models. The prompt format used in our experiments is illustrated in Figure 4.

## 5 Results and Discussion

We benchmark the performance of various models on the **LangMark** test set and discuss broader challenges when evaluating performance on automatic post-editing (APE) tasks. While we have chosen CHRF (Popović, 2015) to show performance in the main text, we report other metrics in the Appendix (A.3).

### 5.1 Model Performance

Table 4 presents the CHRF scores of various closed- and open-source models performing automatic post-editing on the **LangMark** test set using $n$-shot prompting ($n = 20$). The results indicate that *GPT-4o* consistently achieves the highest CHRF scores, being the only closed-source model that consistently improves the NMT output (except for Portuguese), especially in languages where more edits are required (i.e., Japanese and Russian). We also benchmark two open-source models of the *Qwen* and *Llama* family. We found that the performance of the *Qwen* model is impressive for its size, rivaling the best closed-source models and even performing best in Russian.

The strong performance of certain models should not overshadow the broader challenge presented by this dataset. Note that all of the models (except *GPT-4o*) are unable to improve on the NMT baseline, which emphasizes the strength of this dataset as a benchmark for APE.

### 5.2 To Edit or Not to Edit

A critical aspect of automatic post-editing (APE) lies in determining when edits are necessary: some segments require changes while others are best left unchanged. This introduces a classification problem that the model must solve. As NMT systems continue to improve, the challenge shifts. High-performing NMT systems produce outputs that are closer to human translations. In this context, a

[5] https://www.litellm.ai/

[6] https://endpoints.huggingface.co/

Table 4: CHRF scores for different models and languages when performing APE on the test set. Scores are compared across models, with the proprietary MT serving as the baseline.

| | Languages | | | | | | |
|---|---|---|---|---|---|---|---|
| **Model** | **EN-RU** | **EN-BR** | **EN-JP** | **EN-IT** | **EN-FR** | **EN-ES** | **EN-DE** |
| **Baseline** | 68.90 | **89.44** | 70.22 | 89.58 | 81.96 | 86.07 | 81.29 |
| **Gemini-1.5 Flash** | 68.92 | 89.18 | 71.69 | 89.40 | 82.20 | 86.24 | 81.01 |
| **Gemini-1.5 Pro** | 67.73 | 87.65 | 68.92 | 85.68 | 80.46 | 85.01 | 77.88 |
| **Claude 3.5-Sonnet** | 68.63 | 86.47 | 67.14 | 85.10 | 80.31 | 82.73 | 78.44 |
| **Claude 3.5-Haiku** | 69.08 | 88.81 | 71.64 | 88.76 | 82.21 | 86.08 | 80.66 |
| **GPT-4o mini** | 68.55 | 87.73 | 68.47 | 87.47 | 81.45 | 84.94 | 79.81 |
| **GPT-4o** | 69.68 | 89.21 | **73.94** | **89.79** | **82.75** | **86.62** | **81.41** |
| **Open Source** | | | | | | | |
| **Llama 3.1-70B** | 69.55 | 86.82 | 68.37 | 86.80 | 80.97 | 83.75 | 79.12 |
| **Qwen2.5-72B** | **70.13** | 89.03 | 72.93 | 89.10 | 82.34 | 86.44 | 81.16 |

language model that makes only a few highly accurate edits can achieve better evaluation scores than one that identifies more issues but fails to correct them in the exact manner a human would. This raises a crucial question for evaluating APE systems: *"How conservative should models be when deciding that an edit is required?"*

Figure 5 illustrates the correlation between the edits (i.e., deletion, addition, modification) made by the models and those made by human linguists. We observe that *Gemini-1.5 Flash* makes the fewest edits, while *Gemini-1.5 Pro* and *Claude 3.5-Sonnet* show editing behavior more closely aligned with human linguists. Interestingly, even models with the highest number of edits still make fewer changes than the human baseline, highlighting the complexity of this task in **LangMark**.

In the same fashion, Figure 6 shows the recall and precision on the triplets that need correction for all models averaged across languages. Note that we do not explicitly prompt the model to classify each triplet. Thus, in this context:

$$\text{Recall} = \frac{|\{i \in \mathcal{D} \,|\, MT_i \neq H_i \wedge MT_i \neq PE_i\}|}{|\{i \in \mathcal{D} \,|\, MT_i \neq H_i\}|} \quad (1)$$

$$\text{Precision} = \frac{|\{i \in \mathcal{D} \mid MT_i \neq H_i \wedge MT_i \neq PE_i\}|}{|\{i \in \mathcal{D} \mid MT_i \neq PE_i\}|} \quad (2)$$

Where:

- $\mathcal{D}$ is the set of triplets in the dataset.
- $\text{MT}_i$ is the machine translation output for segment $i$.
- $\text{H}_i$ is the human post-edit (ground truth) for segment $i$.
- $\text{PE}_i$ is the model post-edit for segment $i$.

Using this formulation, we can quantify both the frequency with which models detect segments that need edits and their accuracy in determining when a segment needs to be edited. Models with higher precision, such as *GPT-4o*, tend to achieve better overall performance on machine translation evaluation metrics despite having lower recall. We refer to these as "conservative" models. In contrast, "aggressive" models like *Claude 3.5 Sonnet*, perform worse, despite having higher recall.

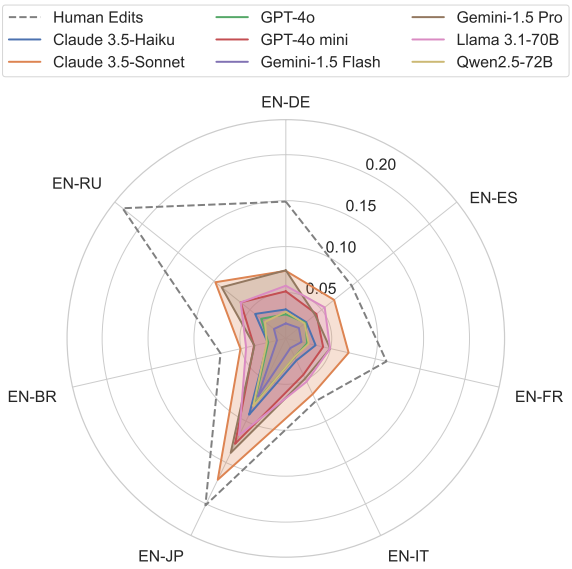


Figure 5: Normalized number of edits made by each model on the NMT output. Note that all models made significantly fewer edits than the human baseline. This indicates that there is still considerable room for improvement

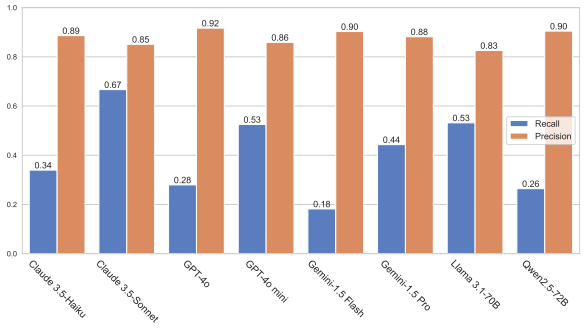


Figure 6: Precision and recall of models when determining that a segment needs to be edited. We see that the models with high recall are not the best performing on machine translation metrics (see Table 4). Instead, the more "conservative" models (low recall, high precision) perform best.

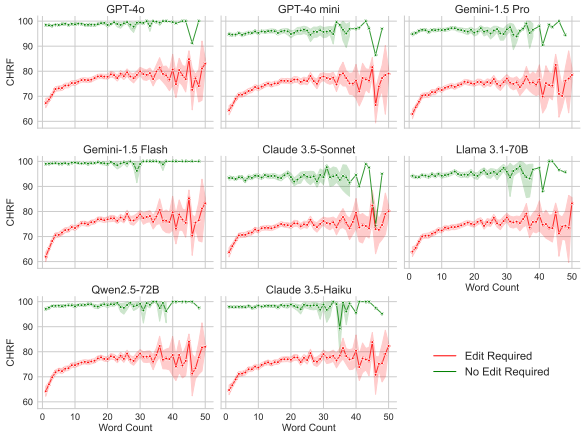


Figure 7: Average performance of each model across segments of varying lengths, separated into those that require edits (*red*) and those that do not (*green*). Models perform substantially worse on shorter segments that need editing, due to limited context. More "aggressive" models (*e.g., Claude 3.5 Sonnet*, *GPT-4-mini*) often modify segments that do not require edits. Only segments of up to 50 words are shown for visualization purposes.

Figure 7 reports the CHRF scores for each model, averaged across all test-set segments and grouped by segment length. For segments requiring no modifications, most models maintain high CHRF scores. However, performance is consistently lower on segments that need correction, hinting at the nuanced nature of the required edits. Editing shorter segments proves especially challenging, likely due to their limited context, which makes it more difficult for APE systems to accurately apply the necessary modifications.

Figures 6 and 7 show that models with a higher recall often over-detect necessary edits. For instance, *Claude 3.5-Sonnet* identifies more segments that require changes but frequently introduces edits where none are needed, affecting performance. This shows that the task of determining whether a segment requires editing is a key challenge in APE settings, especially when nuanced edits are required.

### 5.3 Towards Better Evaluation Metrics

These findings suggest that relying solely on machine translation evaluation metrics is insufficient to fully evaluate APE systems. An ideal evaluation metric should consider both the quality of the final output and the number of edits performed, accounting for the balance between unnecessary conservatism and excessive intervention. Although this work does not propose such a metric, we hope that the dataset introduced here fosters further research into the development of comprehensive evaluation frameworks and promotes the design of APE systems that better align with human post-editing strategies.

## 6 Conclusions

This work introduces **LangMark**, a human-annotated multilingual dataset for automatic post-editing (APE) on neural machine translation (NMT) outputs. The translation is performed *from* English to seven languages, and the data is composed of over 200,000 triplets. The dataset and the results presented in this work constitute a valuable benchmark for evaluating APE systems and advancing research in the field.

Our experiments demonstrate that large language models (LLMs) with few-shot prompting can improve translation quality, outperforming proprietary NMT systems. The fact that most state-of-the-art language models fail to improve on the NMT output that comprises our dataset highlights the strength of **LangMark** as a benchmark for APE systems. Further, we emphasize that machine translation evaluation metrics, while essential to measure performance, fail to account for the classification part of any APE tasks (i.e., determining whether the NMT output needs to be edited). This highlights the need for metrics that better reflect human editing behavior.

We hope that this dataset and the accompanying analysis provide a foundation for further research and benchmarking of Automatic Post-Editing (APE) systems.

## Limitations

Although **LangMark** offers a large-scale, multilingual dataset for automatic post-editing (APE), it also comes with some limitations. First, **LangMark** is derived from a single domain—marketing content—which may constrain the generalizability of APE models trained on it. The dataset's linguistic style and error types may not accurately capture challenges in other domains such as medical, legal, or literary texts.

Second, the dataset is unidirectional, covering only translations *from* English *into* seven target languages. This scope excludes the reverse direction (or translations among non-English languages).

Lastly, despite efforts to remove sensitive or personally identifiable information, the original content—drawn from real marketing documents—may still carry domain-specific biases or cultural nuances. Researchers and practitioners should carefully consider these factors when extending or applying **LangMark** to other use cases or domains.

## Acknowledgments

We would like to express our gratitude to Smartsheet for providing the resources and data that made this research possible. Their support and collaboration were instrumental in the development of the multilingual automatic post-editing dataset presented in this paper. This work would not have been possible without their commitment to advancing research in the field of natural language processing and machine translation.

## References

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2023. Findings of the wmt 2023 shared task on automatic post-editing. In *Proceedings of the Eighth Conference on Machine Translation*, pages 672–681.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. *arXiv preprint arXiv:2409.06790*.

Rajen Chatterjee. 2019. Automatic post-editing for machine translation. *arXiv preprint arXiv:1910.08592*.

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the WMT 2019 shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28, Florence, Italy. Association for Computational Linguistics.

Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018. Findings of the WMT 2018 shared task on automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 710–725, Belgium, Brussels. Association for Computational Linguistics.

Shamil Chollampatt, Raymond Susanto, Liling Tan, and Ewa Szymanska. 2020a. Can automatic post-editing improve nmt? In *Proceedings of EMNLP*.

Shamil Chollampatt, Raymond Hendy Susanto, Liling Tan, and Ewa Szymanska. 2020b. Can automatic post-editing improve NMT? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2736–2746, Online. Association for Computational Linguistics.

Gonçalo M. Correia and André F. T. Martins. 2019. A simple and effective approach to automatic post-editing with transfer learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3050–3056, Florence, Italy. Association for Computational Linguistics.

Félix do Carmo, Dimitar Shterionov, Joss Moorkens, Joachim Wagner, Murhaf Hossari, Eric Paquin, Dag Schmidtke, Declan Groves, and Andy Way. 2021.

A review of the state-of-the-art in automatic post-editing. *Machine Translation*, 35:101–143.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

António Góis, Kyunghyun Cho, and André Martins. 2020. Learning non-monotonic automatic post-editing of translations from human orderings. *arXiv preprint arXiv:2004.14120*.

Julia Ive, Lucia Specia, Sara Szoc, Tom Vanallemeersch, Joachim Van den Bogaert, Eduardo Farah, Christine Maroti, Artur Ventura, and Maxim Khalilov. 2020. A post-editing dataset in the legal domain: Do we underestimate neural machine translation quality? In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3692–3697, Marseille, France. European Language Resources Association.

Marcin Junczys-Dowmunt. 2017. The AMU-UEdin submission to the WMT 2017 shared task on automatic post-editing. In *Proceedings of the Second Conference on Machine Translation*, pages 639–646, Copenhagen, Denmark. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 751–758, Berlin, Germany. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. MS-UEdin submission to the WMT2018 APE shared task: Dual-source transformer for automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 822–826, Belgium, Brussels. Association for Computational Linguistics.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.

Dayeon Ki and Marine Carpuat. 2024. Guiding large language models to post-edit machine translation with error annotations. *arXiv preprint arXiv:2404.07851*.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *AAAI*, volume 94, pages 779–784.

Chen Li, Meishan Zhang, Xuebo Liu, Zhaocong Li, Derek Wong, and Min Zhang. 2024. Towards demonstration-aware large language models for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13868–13881, Bangkok, Thailand. Association for Computational Linguistics.

Matteo Negri, Marco Turchi, Nicola Bertoldi, and Marcello Federico. 2018. Online neural automatic post-editing for neural machine translation. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. *arXiv preprint arXiv:2305.14878*.

Dimitar Shterionov, Félix do Carmo, Joss Moorkens, Murhaf Hossari, Joachim Wagner, Eric Paquin, Dag Schmidtke, Declan Groves, and Andy Way. 2020. A roadmap to neural automatic post-editing: an empirical approach. *Machine Translation*, 34:67–96.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Burchardt, Viviven Macketanz, Inguna Skadin, Matteo Negri, and Marco Turchi. 2017. Translation quality and productivity: A study on rich morphology languages. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 55–71, Nagoya Japan.

Amirhossein Tebbifakhr, Ruchit Agrawal, Matteo Negri, and Marco Turchi. 2018. Multi-source transformer with combined losses for automatic post editing. In

*Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 846–852, Belgium, Brussels. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Blanca Vidal, Albert Llorens, and Juan Alonso. 2022. Automatic post-editing of MT output using large language models. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 84–106, Orlando, USA. Association for Machine Translation in the Americas.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

Thuy-Trang Vu and Reza Haffari. 2018. Automatic post-editing of machine translation: A neural programmer-interpreter approach. In *Empirical Methods in Natural Language Processing 2018*, pages 3048–3053. Association for Computational Linguistics (ACL).

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. 2022. What language model architecture and pretraining objective works best for zero-shot generalization? In *International Conference on Machine Learning*, pages 22964–22984. PMLR.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Ventsislav Zhechev. 2012. Machine translation infrastructure and post-editing performance at autodesk. In *Workshop on Post-Editing Technology and Practice*.

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

# A Appendix

## A.1 Linguist Compensation

In terms of our freelance supplier pool, we prioritize fair compensation for our linguists based on the complexity of their tasks and prevailing market rates. We ensure that our pay rates reflect the market value for each language combination and required skill set, guaranteeing equitable remuneration for all services provided.

Beyond fair pay, we are dedicated to supporting local rural communities in India and Africa through our impactful sourcing program. This initiative creates valuable opportunities for individuals in marginalized communities who might not otherwise have access to such work. Currently, we are running three successful programs in collaboration with companies in these regions.

Additionally, we place great emphasis on engaging with our linguist community. We regularly conduct surveys to gather feedback and continuously refine our work practices, ensuring we meet the needs and expectations of our talented linguists.

## A.2 Zero-Shot Results

Table 5: Zero-shot CHRF scores for different models and languages when performing APE on the test set. Scores are compared across models, with the proprietary MT serving as the baseline.

| | Languages | | | | | | |
|---|---|---|---|---|---|---|---|
| **Model** | **EN-RU** | **EN-PT** | **EN-JP** | **EN-IT** | **EN-FR** | **EN-ES** | **EN-DE** |
| **Baseline** | 68.90 | **89.44** | 70.22 | **89.58** | 81.96 | 86.07 | **81.29** |
| **Gemini-1.5 Flash** | 68.80 | 88.97 | 71.59 | 88.95 | 82.26 | 86.14 | 80.85 |
| **Gemini-1.5 Pro** | 65.95 | 86.65 | 68.01 | 84.42 | 79.74 | 84.45 | 77.67 |
| **Claude 3.5-Sonnet** | 67.83 | 87.68 | 68.00 | 86.78 | 80.73 | 83.43 | 79.18 |
| **Claude 3.5-Haiku** | 68.62 | 88.86 | 71.90 | 88.99 | 82.24 | 86.01 | 80.57 |
| **GPT-4o mini** | 67.78 | 87.84 | 69.73 | 87.99 | 81.40 | 84.91 | 80.10 |
| **GPT-4o** | **68.99** | 89.21 | **73.46** | 89.29 | **82.24** | **86.34** | 81.06 |
| **Open Source** | | | | | | | |
| **Llama 3.1-70B** | 66.84 | 85.41 | 68.80 | 85.30 | 79.88 | 81.54 | 77.07 |
| **Qwen2.5-72B** | 68.62 | 89.21 | 72.86 | 89.23 | 82.27 | 86.07 | 81.08 |

Table 6: Zero-shot TER↓ (Snover et al., 2006) scores for different models and languages when performing APE on the test set. Scores are compared across models, with the proprietary MT serving as the baseline.

| | Languages | | | | | | |
|---|---|---|---|---|---|---|---|
| **Model** | **EN-RU** | **EN-PT** | **EN-JP** | **EN-IT** | **EN-FR** | **EN-ES** | **EN-DE** |
| **Baseline** | 45.40 | 14.27 | 74.15 | **14.61** | 26.67 | 19.28 | **31.26** |
| **Gemini-1.5 Flash** | 45.71 | 14.67 | 72.87 | 15.40 | 25.60 | 19.28 | 31.61 |
| **Gemini-1.5 Pro** | 49.51 | 17.65 | 74.52 | 20.94 | 28.76 | 21.42 | 35.77 |
| **Claude 3.5-Sonnet** | 47.16 | 16.18 | 79.14 | 18.24 | 27.70 | 22.75 | 33.74 |
| **Claude 3.5-Haiku** | 45.70 | 14.66 | 74.75 | 15.28 | **25.56** | 19.41 | 31.76 |
| **GPT-4o mini** | 46.66 | 15.63 | 76.08 | 16.52 | 26.52 | 20.58 | 32.47 |
| **GPT-4o** | **45.35** | 14.67 | **71.75** | 14.96 | 25.87 | **19.04** | 31.30 |
| **Open Source** | | | | | | | |
| **Llama 3.1-70B** | 47.77 | 18.67 | 76.20 | 19.59 | 28.83 | 27.85 | 41.08 |
| **Qwen2.5-72B** | 45.69 | **14.22** | 71.25 | 15.00 | 25.66 | 19.34 | 31.30 |

Table 7: Zero-shot BLEU (Papineni et al., 2002) scores for different models and languages when performing APE on the test set. Scores are compared across models, with the proprietary MT serving as the baseline.

| | Languages | | | | | | |
|---|---|---|---|---|---|---|---|
| **Model** | **EN-RU** | **EN-PT** | **EN-JP** | **EN-IT** | **EN-FR** | **EN-ES** | **EN-DE** |
| **Baseline** | **49.13** | **80.16** | 14.28 | **79.93** | **64.91** | 73.75 | **64.13** |
| **Gemini-1.5 Flash** | 48.90 | 79.51 | 33.61 | 79.09 | 66.56 | 74.28 | 63.61 |
| **Gemini-1.5 Pro** | 44.31 | 75.31 | 32.80 | 71.28 | 62.68 | 71.44 | 58.34 |
| **Claude 3.5-Sonnet** | 47.44 | 77.12 | 30.93 | 75.34 | 64.44 | 69.76 | 60.82 |
| **Claude 3.5-Haiku** | 48.63 | 79.37 | 33.38 | 79.13 | 66.73 | 74.06 | 63.20 |
| **GPT-4o mini** | 47.62 | 77.69 | 27.51 | 77.47 | 65.37 | 72.30 | 62.40 |
| **GPT-4o** | 48.99 | 79.58 | **34.95** | 79.51 | 66.02 | **74.49** | 63.82 |
| **Open Source** | | | | | | | |
| **Llama 3.1-70B** | 46.03 | 73.87 | 32.31 | 73.17 | 63.03 | 65.58 | 54.83 |
| **Qwen2.5-72B** | 48.45 | 79.79 | 34.24 | 79.46 | 66.62 | 74.20 | 63.90 |

## A.3 Additional Metrics

Table 8: TER↓ scores (Snover et al., 2006) for different models and languages when performing APE on the test set. Scores are compared across models, with the proprietary MT serving as the baseline. Lower is better.

| | Languages | | | | | | |
|---|---|---|---|---|---|---|---|
| **Model** | **EN-RU** | **EN-PT** | **EN-JP** | **EN-IT** | **EN-FR** | **EN-ES** | **EN-DE** |
| **Baseline** | 45.40 | **14.27** | 74.15 | 14.61 | 26.67 | 19.28 | 31.26 |
| **Gemini-1.5 Flash** | 45.62 | 14.42 | 71.59 | 14.81 | 25.83 | 19.14 | 31.43 |
| **Gemini-1.5 Pro** | 47.53 | 16.37 | 70.84 | 19.52 | 27.95 | 20.76 | 35.60 |
| **Claude 3.5-Sonnet** | 46.56 | 17.82 | 75.66 | 20.57 | 28.34 | 23.67 | 34.90 |
| **Claude 3.5-Haiku** | 45.60 | 14.72 | 72.12 | 15.59 | 25.71 | 19.51 | 31.78 |
| **GPT-4o mini** | 46.17 | 16.08 | 74.68 | 17.27 | 26.54 | 20.56 | 32.74 |
| **GPT-4o** | 44.49 | 14.41 | 69.01 | **14.25** | **25.30** | **18.64** | **30.91** |
| **Open Source** | | | | | | | |
| **Llama 3.1-70B** | 45.12 | 17.44 | 73.94 | 18.39 | 27.80 | 22.26 | 33.80 |
| **Qwen2.5-72B** | **43.91** | 14.45 | **68.75** | 15.23 | 25.71 | 18.95 | 30.95 |

Table 9: BLEU (Papineni et al., 2002) scores for different models and languages when performing APE on the test set. Scores are compared across models, with the proprietary MT serving as the baseline.

| | Languages | | | | | | |
|---|---|---|---|---|---|---|---|
| **Model** | **EN-RU** | **EN-PT** | **EN-JP** | **EN-IT** | **EN-FR** | **EN-ES** | **EN-DE** |
| **Baseline** | 49.13 | **80.16** | 14.28 | 79.93 | 64.91 | 73.75 | 64.13 |
| **Gemini-1.5 Flash** | 48.69 | 79.80 | 34.17 | 79.59 | 66.50 | 74.37 | 63.71 |
| **Gemini-1.5 Pro** | 46.35 | 77.04 | 36.27 | 73.23 | 63.74 | 72.47 | 58.16 |
| **Claude 3.5-Sonnet** | 47.53 | 74.83 | 33.94 | 71.92 | 63.61 | 68.20 | 59.08 |
| **Claude 3.5-Haiku** | 48.58 | 79.17 | 35.72 | 78.72 | 66.61 | 74.11 | 63.10 |
| **GPT-4o mini** | 47.92 | 77.30 | 27.81 | 76.17 | 65.21 | 72.27 | 61.89 |
| **GPT-4o** | 49.79 | 79.86 | **37.96** | **80.12** | **66.91** | **74.84** | **64.20** |
| **Open Source** | | | | | | | |
| **Llama 3.1-70B** | 49.28 | 75.76 | 33.01 | 74.97 | 64.22 | 70.27 | 60.70 |
| **Qwen2.5-72B** | **50.31** | 79.59 | 37.43 | 79.16 | 66.60 | 74.79 | 64.01 |