Extending CREAMT: Leveraging Large Language Models for Literary Translation Post-Editing

Antonio Castaldo
University of Naples "L'Orientale"
University of Pisa
antonio.castaldo@phd.unipi.it

Sheila Castilho
Dublin City University
sheila.castilho@adaptcentre.ie

Joss Moorkens
Dublin City University
joss.moorkens@dcu.ie

Johanna Monti
University of Naples "L'Orientale"
jmonti@unior.it

Abstract

Post-editing machine translation (MT) for creative texts, such as literature, requires balancing efficiency with the preservation of creativity and style. While neural MT systems struggle with these challenges, large language models (LLMs) offer improved capabilities for context-aware and creative translation. This study evaluates the feasibility of post-editing literary translations generated by LLMs. Using a custom research tool, we collaborated with professional literary translators to analyze editing time, quality, and creativity. Our results indicate that post-editing LLM-generated translations significantly reduces editing time compared to human translation while maintaining a similar level of creativity. The minimal difference in creativity between PE and MT, combined with substantial productivity gains, suggests that LLMs may effectively support literary translators working with high-resource languages.

1 Introduction

Post-editing of MT has become an increasingly common service, given the cost-efficiency and good quality compromise that this practice offers. However, while several studies have confirmed that post-editing MT boosts productivity in terms of translation speed (Terribile, 2023), the benefits diminish significantly when dealing with poor-quality MT outputs (Guerberof Arenas, 2014; Sanchez-Torron and Koehn, 2016). This challenge is particularly pronounced for literary texts, where the final quality often suffers not only in terms of translation accuracy but also in the preservation of creativity, as discussed by Guerberof-Arenas and Toral (2020).

Recent LLM advancements have demonstrated significant improvements in handling context issues and figurative language to generate highly accurate and fluent translations. Unlike NMT systems that often tend towards generating translations that are either too literal or inaccurate, LLMs leverage large training data to generate context-aware translations less literally. Nevertheless, the extent to which they may support literary translators, without sacrificing creativity, remains underexplored.

In this study, we collaborated with four professional translators to evaluate the feasibility of post-editing literary translations generated by LLMs, focusing on three key aspects: editing time, translation quality, and creativity. We compare the performance of GPT-4, GPT-3.5, and a literary-adapted Mistral-7B model. We also developed a custom research tool called UniOr-PET (Castaldo et al., 2025) to collect detailed statistics on the editing process of a literary sci-fi novel.

Our findings reveal that post-editing LLM-generated translations between well-supported languages significantly reduces editing time compared to human translation while maintaining a similar level of creativity. As the difference in creativity scores between human and post-edited LLM translations appears to be minimal, our findings suggest that LLMs can serve as valuable tools for literary translators.

2 Related Work

Research on post-editing has traditionally centered on technical and commercial texts, where terminological consistency and turnaround time are often prioritized (Moorkens et al., 2018). However, trans-

Proceedings of Machine Translation Summit XX Volume 1, pages 506–515, June 23-27, 2025

lating creative works such as literature introduces unique challenges. NMT models have been shown to struggle with creative phraseological challenges, such as translating idiomatic expressions, where they often produce overly literal outputs.

Corpas Pastor and Noriega-Santiáñez (2024) highlighted these limitations, particularly in the context of literary texts. In contrast, Raunak et al. (2023) demonstrated that LLMs are capable of generating less literal and more contextually appropriate translations, especially when translating idiomatic expressions that tend to be generated with a higher level of abstraction, defined by the authors as "figurative compositionality". Further studies on idiomatic expression translation, particularly for the English-Italian language pair, have confirmed the high-quality results achieved by general-purpose LLMs (Castaldo and Monti, 2024). Their findings suggest that these models could address some of the shortcomings observed in NMT systems when translating literature, making them a promising tool for literary translation.

A study conducted by Guerberof-Arenas and Toral (2022) concluded that NMT was unable to handle the complex demands of translating literature or supporting literary translators effectively, resulting in low-quality outputs and diminished creativity. Their findings revealed the limitations of such models in preserving creativity during translation, becoming a constraint for the translator's creativity when used. Human translation (HT) consistently outperformed MT and PE in creativity, as evidenced by the annotation of units of creative potential. These findings align with the study by Castilho and Resende (2022), that showed how the features found in post-edited translations align more closely with the ones found in the MT output than in the HT. However, more recent advances in LLMs may shift this paradigm.

As demonstrated by Karpinska and Iyyer (2023) and Castilho et al. (2023), LLMs excel at leveraging training data to deal with context-related issues, which is critical for translating creative works that require discourse-level coherence and contextual understanding. Techniques such as in-context learning (Brown et al., 2020) and prompt engineering allow LLMs to maintain higher degrees of fluency, consistency, and stylistic fidelity compared to NMT systems. Finally, their ability to adapt to specific linguistic patterns and translation memories in real time, as shown by Moslem et al. (2023), further enhances their applicability in the creative translation domain, suggesting that LLMs could potentially overcome the creativity gap identified in NMT outputs, supporting professional translators in producing high-quality creative translations with context-aware terminology and accurate lexicon.

Drawing on Guerberof-Arenas and Toral (2022), in this study we consider creativity as a process that requires both originality and effectiveness (Runco and Jaeger, 2012). This implies that in order for a product to be creative, it needs not only to be novel but also of value, and therefore acceptable, for the context in which it is created. In Section 5, we will use the annotations of units of creative potential to reflect the original units introduced by the translators (novelty), and translation quality metrics as a proxy for the translation acceptability.

3 Methodology

We collaborated with four professional translators who specialize in literary and editorial translations to translate and post-edit excerpts from the novel "Oryx and Crake" by Canadian author Margaret Atwood (Atwood, 2004) from English into Italian. The novel was selected for its extensive use of playful and thought-provoking neologisms, vivid imagery, and richly detailed language, which present significant challenges in the translation process (Miller, 2019; Gurov, 2022; Noriega-Santiáñez and Corpas Pastor, 2023)

3.1 Participants

Each translator post-edited outputs of comparable length (roughly 2200 words), generated by three LLMs (see §3.2). We designed our study so that each translator contributes equally to the evaluation of the four models, rotating the chunks so that each translator works on three unique chunks, each generated by a different model. In this way, we minimize biases introduced by translator-specific behavior. We demonstrate our approach in Figure 1.

In addition, each translator produced a segment of the same excerpt translated from scratch. This experimental setup enabled us to collect fully post-edited translations for each model and a complete HT of the text for comparative analysis.

3.2 Models and Training

We employed three LLMs for generating the initial translations: `GPT-4`, `GPT-3.5`, and a literary-adapted `Mistral-7B` model, ordered by parameter

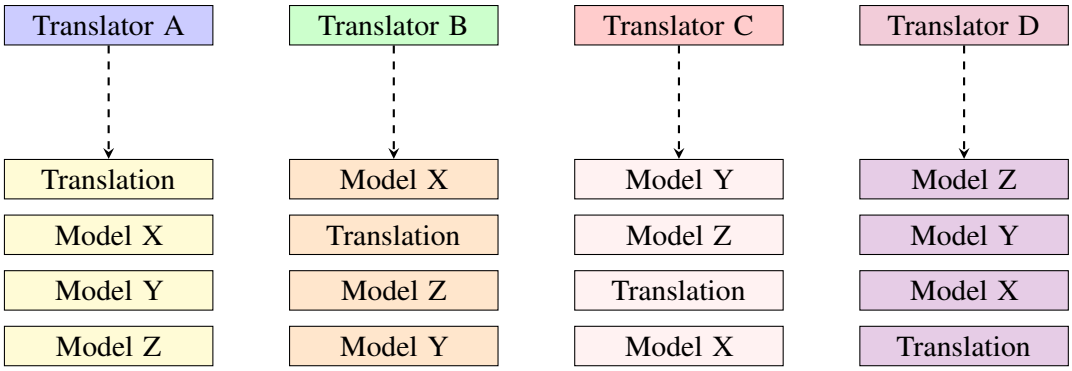

Figure 1: Each translator translates from scratch one chunk of original text (Translation) and post-edits a different chunk of each model's output (Model X, Y, Z), minimizing the translator's effect.

size. Access to the GPT models (OpenAI et al., 2024) was obtained through the OpenAI API,[1] as they both operate under closed-source licenses. In contrast, Mistral-7B (Jiang et al., 2023) was obtained as an open-source checkpoint, allowing us to fine-tune it locally for literary translation. Mistral-7B was fine-tuned on a curated corpus of modern literary works obtained from Opus Corpus (Tiedemann and Thottingal, 2020), for a total of 30,000 parallel segments. The model was fine-tuned for three epochs using Low-Rank Adaptation (Hu et al., 2021), a fine-tuning technique which injects small trainable matrices in the model's weights. The training corpus encompassed contemporary novels, short stories, and excerpts from science fiction and fantasy genres. The corpus was selected for its stylistic resemblance to the target text.

After fine-tuning, translation quality metrics and human inspection confirmed that Mistral-7B displayed improved handling of figurative language, idiomatic expressions, and higher accuracy. In terms of quality metrics, it achieved +4 points of corpus-level BLEU and +7 points of COMET as compared to its off-the-shelf counterpart.

3.3 Tools and Workflow

To facilitate the translation and post-editing process and collect meaningful data, we used two tools: our custom-built UniOr-PET and the established PET tool (Aziz et al., 2012).

UniOr-PET was designed specifically for this study, offering a browser-based platform that eliminates the need for software installation (see Figure 2). This feature addresses concerns often raised by translators regarding the inconvenience of downloading external applications, as is the case with

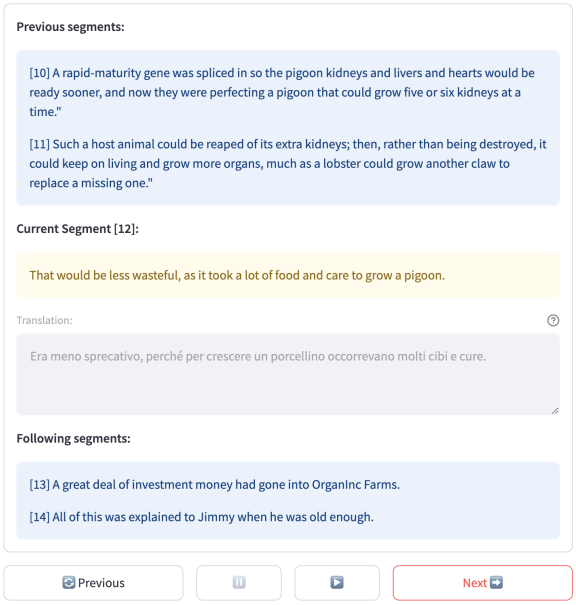

Figure 2: UniOr PET user interface

the PET tool. The tool records key metrics such as editing time, the number and types of edits, keeping track of insertions, and deletions. Similarly to the PET tool, UniOr PET gives the ability to read the texts, before recording editing time, making the results from both tools equally comparable. Translators could also save their work and revisit previously edited segments. The interface was configured to present the ST, LLM output, and an editable field, with a horizontal or vertical layout.

Recognizing the importance of context in literary translation (Nelson Jr., 1989; House, 2006), UniOr-PET also allowed translators to view a configurable number of preceding and following segments alongside the current one. This feature ensured that they could maintain consistency in tone, style, and narrative flow, an essential consideration when translating richly detailed texts, such as literature.

[1] https://openai.com

In addition to UniOr-PET, translators could opt to use the PET tool, which remains a popular choice for post-editing research due to its robust functionality and familiarity among professional translators, and researchers alike. Like its browser-based relative, PET captures data such as editing times and the types of edits made, providing a rich dataset for analysis. These tools provided translators with the flexibility to choose the interface that best suited their workflow preferences while allowing us to capture detailed post-editing data.

4 Results and Analysis

Thanks to the use of UniOr-PET and PET, we were able to collect significant data on each translation version providing foundation for a comparative analysis of the different models. More specifically, we have calculated quality metrics with BLEU (Papineni et al., 2002), ChrF (Popović, 2015) and COMET (Rei et al., 2020), which we average and normalize by time, as well as aggregated editing times. Finally, we compute Human-targeted Translation Edit Rate (Snover et al., 2006).

4.1 Editing Times

Source	Total
GPT-4	64.33
Mistral-60k	87.12
HT	115.68
GPT-3.5	119.74

Table 1: Editing Times (in Minutes)

Table 1 presents the aggregated total editing times (in minutes) for all translators and each part of the dataset. We find that editing time is shorter when post-editing outputs of the larger and best performing model used in our experiment, GPT-4. Interestingly, the literary-adapted Mistral model, despite its smaller size, demonstrated editing times significantly shorter than those for GPT-3.5. This suggests that domain adaptation, even in smaller models, can have a measurable impact on post-editing efficiency. These findings align with previous research indicating that better translation quality leads to reduced post-editing effort (Sanchez-Torron and Koehn, 2016; Zouhar et al., 2021).

The longest editing times were recorded when translating from scratch, which is expected since it requires significantly more technical (typing) effort than post-editing pre-generated MT outputs.

4.2 Human Translation Edit Rate (HTER)

Table 2 presents the HTER scores for the post-editing outputs from different MT systems. HTER is a widely used metric that quantifies the minimum number of edits required to improve an MT output when post-editing, where lower values indicate fewer required minimum edits. Therefore, HTER does not necessarily correspond to the actual number of edits, but rather represents an estimate of post-editing effort.

Source	T1	T2	T3	T4	Doc
GPT-3.5	44.4	41.9	62.2	31.8	**52**
GPT-4	50.4	66.5	52.2	29.9	54
Mistral-60k	66.1	66.0	71.5	54.5	**71**
HT	81.5	71.2	61.0	56.2	66
Total	242.4	245.6	247.0	172.4	226.85

Table 2: Human Translation Edit Rate. Lowest and highest HTER values are displayed in **bold**.

The results indicate varying levels of post-editing effort across the systems and across the four translators, with Translator 4 (T4) standing as an outlier when working with GPT models. This may be due to the adoption of a lighter form of post-editing, or an inclination to accept MT outputs considered sufficiently fluent and accurate.

We find that outputs from GPT-3.5 generally required the fewest edits, as reflected in the lowest HTER values among the systems. However, despite requiring fewer edits, post-editing outputs from GPT-3.5 took more time compared to the other models, as shown in Table 1. As both tools offer the possibility to read the texts, before performing translation, the results suggest that while the initial quality of GPT-3.5 translations was relatively higher, the type of edits required may have been more complex or time-consuming.

Interestingly, GPT-4 translations required more edits than GPT-3.5 but less overall editing time, indicating that its errors were likely easier to correct. Mistral-60k, while requiring more edits than GPT-3.5 and GPT-4, had comparable or shorter editing times, possibly due to simpler or more predictable error patterns. Translations from the ST show a significant difference from the reference translation, consistent with the lack of post-editing constraints.

As expected, we confirm a strong inverse correlation between HTER and quality metrics of the

original MT outputs, displayed in Figure 3, indicating that lower quality MT outputs require more post-editing efforts.

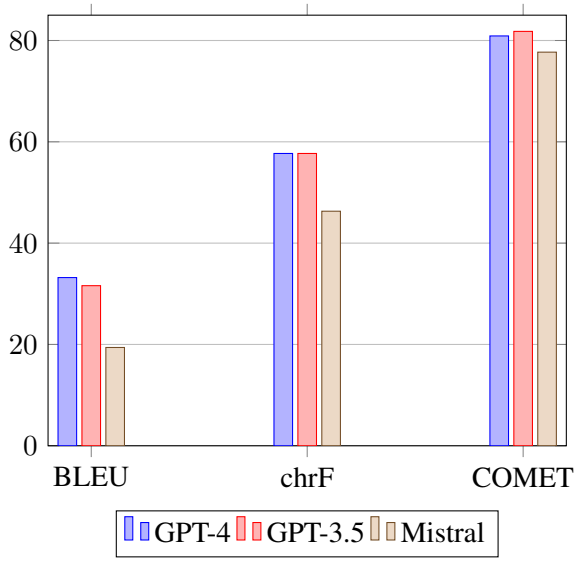

Figure 3: Quality metrics scores (BLEU, chrF, COMET) for different MT systems.

4.3 Quality-to-Time Ratio

Table 3 shows the normalized quality-to-time ratio for each MT system, calculated as the average of all quality metrics (BLEU, ChrF, and COMET) divided by the total editing time (Table 1). This ratio provides a measure of efficiency, combining the quality of the post-edited output with the time required to achieve it. Higher values indicate more efficient systems where higher-quality translations are achieved in less time.

Source	Ratio	BLEU	chrF	COMET
GPT-4	**0.38**	31.8	58.2	83.1
Mistral-60k	0.29	27.6	55.0	83.6
GPT-3.5	0.28	30.8	58.7	84.0
HT	0.23	27.1	54.4	80.5

Table 3: Quality-to-Time Ratio, calculated as the average of all quality metrics divided by the total editing time, along with BLEU, chrF, and COMET scores.

The results reveal that GPT-4 achieves the highest quality-to-time ratio (0.86), demonstrating the initial quality of the translation and the reduced post-editing effort, leading to good-quality post-edited translations in the shortest time.

Interestingly, Mistral-60k achieves the lowest ratio across the three models, despite requiring less editing time compared to GPT-3.5. This suggests that while Mistral translations may be quicker to edit, their initial quality presents challenges that limit their effectiveness in producing high-quality outputs efficiently, possibly resulting in the translator's decision to perform a lighter form of post-editing (Nitzke and Hansen-Schirra, 2021).

5 Creativity Annotation

To evaluate creativity in the post-edited translations and conduct a model-wise comparison, we annotated units of creative potential in the ST and creative shifts in the target texts (TT), that were originally generated by the three LLMs, and then post-edited by four translators.

Annotation Process. Our annotation framework follows the methodology proposed by Guerberof-Arenas and Toral (2022), where units of creative potential (UCPs) are defined as units that could invite creative deviations during post-editing, aimed at preserving or enhancing the creativity found in the ST, and creative shifts reflect the actual creative units introduced by translators during post-editing. Annotations were performed by two linguists with expertise in translation studies, who are native speakers of the target language and proficient in English. After annotating 10% of the dataset, inter-annotator agreement (IAA) was calculated to ensure the reliability of the annotations. The initial agreement, measured with Cohen's Kappa, was equal to $K = 0.35$ for Type Agreement and $K = 0.85$ for Span Agreement, due to disagreements primarily on the type of creative shift to assign, rather than the identification of the creative shifts themselves. Following a collaborative resolution process, we refined the annotation guidelines and calculated agreement on the final annotations, reaching a Type Agreement equal to $K = 0.57$ and a similarly high Span Agreement, equal to $K = 0.86$.

Creativity Score. A creative work must be both novel and acceptable, thereby achieving a balance between creativity and quality. In order to account for both novelty, as indicated by the number of creative shifts, and acceptability, as reflected by translation quality, we used `WMT22-COMET-DA` (Rei et al., 2022) for an automatic reference-based quality evaluation, and calculated the creativity score across the four translations.

In this study, we employ COMET as our primary metric for assessing translation quality, recogniz-

$$\text{Creativity Score} = \left(\frac{\#\text{CSs}}{\#\text{UCPs}} - \frac{\#\text{error points} - \#\text{Kudos}}{\#\text{words in ST}}\right) \times 100.$$

Figure 4: The original creativity score formula, that we started from to create our score.

ing that MQM would provide a more fine-grained evaluation of translation errors. Our decision to use COMET is motivated by its strong correlation with human judgments, as demonstrated in previous research (Rei et al., 2020; Kocmi et al., 2024), and by its practical advantage in automatic evaluation, in light of constraints related to time and resources. Having been trained on MQM-annotated datasets, COMET should effectively reflect the types of errors found in the outputs. Therefore, we integrate COMET in our creativity evaluation formula, as a proxy for translation acceptability.

Compared to the formula used in the original study, presented in Figure 4, we adapt the acceptability equation to accommodate the use of a quality metric, where higher means better, in place of the original error metric. Therefore, we multiply the creative shifts ratio by COMET scores, and then multiply by 100 to express it as a percentage. This allows us to reward creativity in proportion to quality, similarly to the original study. We present the new creativity score formula below.

$$\text{CS} = \left(\frac{\textit{Creative Shifts}}{\textit{UCPs}} \times \text{COMET}\right) \times 100 \quad (1)$$

5.1 Annotation Results

Table 4 summarizes the annotation results for each translation variant. For each system, we present the number of the creative shifts introduced by the translators, the COMET score, and the resulting creativity score, calculated with our new formula. A higher creativity score suggests a better balance between the introduced creative elements and the final translation quality.

System	CS Ratio	COMET	Creativity
HT	0.30	**0.85**	25.5%
GPT-3.5	0.24	0.84	20.1%
Mistral	0.30	0.83	24.9%
GPT-4	**0.32**	0.83	**26.5%**

Table 4: Creativity annotation results, where we display Creative Shifts ratio, COMET Score, and Creativity Score for each system.

6 Discussion

Taken together, our results show that a larger and more advanced model (GPT-4) generated translations that required fewer edits and resulted in a higher-quality post-edited translation, as resulted from the lower editing time and the higher quality-to-time ratio. The creativity score is also the highest, suggesting an interesting correlation between original MT quality and creativity in post-editing.

The domain-adapted Mistral-7B model also displayed promising performance, obtaining a quality-to-time ratio higher than the one obtained by the larger GPT-3.5, requiring more edits but a significantly lower editing time, while obtaining a similar creativity score. In this case, we find that Mistral's creativity comes at the cost of increased post-editing effort. HT, despite requiring a significantly higher editing time, is the most accurate translation variant according to COMET scores and it presents a high creativity score that is very similar to the post-edited texts.

In Table 5 we present two segments for each translation version with the highest and lowest post-editing effort, as measured by HTER. In displaying the segments, we ignore cases where the HTER is equal to zero due to translators not making any changes to the MT output. The examples reveal several interesting patterns. In some cases, the translators decided to merge or split certain sentences. Extensive edits were made in segments containing UCPs, as in the second example for GPT-3.5. Similarly, we find several edits where the original MT quality was particularly low, as seen in the second segment from the Mistral model. Interestingly, we find that where the MT systems failed to render neologisms effectively, translators were forced to produce a creative alternative, effectively improving the creativity of the translation.

Overall, we find that the creativity score does not differ significantly between the four models, as both the number of identified creative shifts and the quality metrics are similar across all translation variants. These findings are in contrast with what was found in the original study, where the difference between the two modalities (HT and PE) was substantial and HT was found to be notably more

creative than their post-edited variant. We speculate that the higher and more fluent MT quality given by LLMs may be of less constraint to the translator in the post-editing process, leading to equally creative translations.

7 Conclusion

In this study, we investigated the potential of LLM-based post-editing in the literary domain, comparing a literary-adapted Mistral model with GPT-4 and GPT-3.5. By collaborating with four professional literary translators, we collected detailed data on editing times, error rates, and post-editing efficiency, using our custom-built tool UniOr-PET. We demonstrate the contributions that LLMs can make in literary post-editing workflows, bridging the gap between productivity and creativity.

Our findings highlight two important benefits granted by the adoption of LLMs. First, we demonstrate that, in the context of our study, creativity does not present a significant difference between human translation and post-edited LLM translations. The marginal difference in creativity between the four translation variants suggests that the post-edited outputs may preserve creativity effectively. This may be due to the more fluent and higher-quality outputs given by the original MT versions, that represent less of a constraint to the translators, compared to NMT outputs.

Second, we observe a clear productivity gain in post-editing compared to human translation, even when post-editing translations generated by a smaller model. Given that the creativity gap is relatively small across translation variants, the productivity gains may offset the minor differences in creativity, achieving similarly creative translations with significantly less effort and time.

Finally, we reinforce the potential of fine-tuning techniques for literary MT workflows, demonstrating that even by adopting a small literary-adapted model, it is possible to achieve a good balance between translation quality and efficiency.

8 Limitations

One of the main limitations of this study is that our data collection process involved only four translators working in a single relatively well-resourced language pair and a relatively short literary excerpt. Further studies, on a larger scale, are required to investigate the possible correlations between creativity and other metrics. It is also worth mentioning that although our study follows established proxies for measuring creativity, these should be verified with a reception study, as suggested by Guerberof-Arenas and Toral (2020).

For the acceptability score, meant to balance creativity by translation quality in the post-edited texts, we used COMET scores in place of human evaluation. While COMET has shown strong correlations with human judgment, it remains an automated metric and may not fully capture the extent of literary translation quality.

Finally, while our literary-adapted Mistral model showed promising performance, its fine-tuning was performed using a modest-sized corpus, leaving open the way for further experimentation.

8.1 CO2 Emission Related to Experiments

Experiments were conducted using Amazon Web Services in region eu-west-1, which has a carbon efficiency of 0.62 $kgCO_2eq/kWh$. A cumulative of 3 hours of computation was performed on hardware of type RTX A6000 (TDP of 300W).

Total emissions are estimated to be 0.56 $kgCO_2eq$ of which 100 percents were directly offset by the cloud provider.

Estimations were conducted using the Machine Learning Impact calculator presented in Lacoste et al. (2019).

Acknowledgments

We thank the two annotators who took part in this study. Part of this work has been funded by the Italian National PhD programme in Artificial Intelligence, partnered by University of Pisa and University of Naples “L’Orientale", through a doctoral grant (ID 39-411-24-DOT23A27WJ-6603) established by Ex DM 318, of type 4.1, co-financed by the National Recovery and Resilience Plan. The second and third authors benefit from being members of the ADAPT SFI Research Centre at Dublin City University, funded by the Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2.

References

Margaret Atwood. 2004. *Oryx and Crake*. Number v.1 in The MaddAddam Trilogy Ser. Knopf Doubleday Publishing Group, New York.

Wilker Aziz, Sheila Castilho, and Lucia Specia. 2012. PET: a Tool for Post-editing and Assessing Machine

Translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3982–3987, Istanbul, Turkey. European Language Resources Association (ELRA).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.

Antonio Castaldo, Sheila Castilho, Joss Moorkens, and Johanna Monti. 2025. Unior PET: An Online Platform for Translation Post-Editing. In *20th Machine Translation Summit: Products and Projects track*, Geneva, Switzerland. European Association for Machine Translation.

Antonio Castaldo and Johanna Monti. 2024. Prompting Large Language Models for Idiomatic Translation. In *Proceedings of the First Workshop on Creative-text Translation and Technology*, pages 37–44, Sheffield, UK. Accepted: 2024-06-19T21:00:05Z.

Sheila Castilho, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. Do online Machine Translation Systems Care for Context? What About a GPT Model? In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 393–417, Tampere, Finland. European Association for Machine Translation.

Sheila Castilho and Natália Resende. 2022. Post-Editese in Literary Translations. *Information*, 13(2):66. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Gloria Corpas Pastor and Laura Noriega-Santiáñez. 2024. Human versus Neural Machine Translation Creativity: A Study on Manipulated MWEs in Literature. *Information*, 15(9):530.

Ana Guerberof Arenas. 2014. Correlations between productivity and quality when post-editing in a professional context. *Machine Translation*, 28(3):165–186.

Ana Guerberof-Arenas and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, 11(2):184–212.

Andrey Gurov. 2022. Literary Translation as An Insurmountable Obstacle for Neural Networks. *SSRN Electronic Journal*.

Juliane House. 2006. Text and context in translation. *Journal of Pragmatics*, 38(3):338–358.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint*. ArXiv:2106.09685 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint*. Issue: arXiv:2310.06825 arXiv:2310.06825 [cs].

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. Issue: arXiv:2304.03245 arXiv:2304.03245 [cs].

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies. Issue: arXiv:2401.06760 arXiv:2401.06760 [cs].

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Tristan Miller. 2019. The Punster's Amanuensis: The Proper Place of Humans and Machines in the Translation of Wordplay. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 57–65, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Joss Moorkens, Antonio Toral, Sheila Castilho, and Andy Way. 2018. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2):240–262. Publisher: John Benjamins Publishing Company.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive Machine Translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Lowry Nelson Jr. 1989. Literary Translation. *Translation Review*, 29(1):17–30. Publisher: Routledge.

Jean Nitzke and Silvia Hansen-Schirra. 2021. *A short guide to post-editing (Volume 16)*. Language Science Press.

Laura Noriega-Santiáñez and Gloria Corpas Pastor. 2023. Machine vs Human Translation of Formal Neologisms in Literature: Exploring E-tools and Creativity in Students. *Tradumàtica tecnologies de la traducció*, (21):233–264.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, and Greg Brockman. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan Awadalla. 2023. Do GPTs Produce Less Literal Translations? ArXiv: 2305.16806.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Mark A. Runco and Garrett J. Jaeger. 2012. The Standard Definition of Creativity. *Creativity Research Journal*, 24(1):92–96.

Marina Sanchez-Torron and Philipp Koehn. 2016. Machine Translation Quality and Post-Editor Productivity. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, pages 16–26, Austin, TX, USA. The Association for Machine Translation in the Americas.

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.

Silvia Terribile. 2023. Is post-editing really faster than human translation? *Translation Spaces*, 13(2):171–199. Publisher: John Benjamins Publishing Company.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Vilém Zouhar, Martin Popel, Ondřej Bojar, and Aleš Tamchyna. 2021. Neural Machine Translation Quality and Post-Editing Performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10204–10214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Model	Type	Text
GPT-3.5 *(Lowest HTER)*	ST	*But he hadn't wet his bed for a long time...*
	HT	*Eppure era un pezzo che non bagnava il letto...* yet was a while that not wet the bed...
	MT	*Ma non aveva bagnato il letto da molto tempo...* but not had wet the bed since much time...
	PE	*Eppure era da un pezzo che non bagnava il letto...* yet was since a while that not wet. the bed...
GPT-3.5 *(Highest HTER)*	ST	*Some cheap do-it-yourself enlightenment handbook*
	HT	*Uno scadente manuale di auto rivelazione* a poor manual of self revelation
	MT	*Una specie di manuale economico per l'illuminazione* a kind of manual cheap for the.enlightenment
	PE	*Una specie di manuale a poco prezzo per raggiungere l'illuminazione* a kind of manual at little price to reach the.enlightenment
GPT-4o *(Lowest HTER)*	ST	*All of this was explained to Jimmy when he was old enough.*
	HT	*Tutto questo fu spiegato a Jimmy quando fu abbastanza grande.* all this was explained to Jimmy when was sufficiently big.
	MT	*Tutto questo fu spiegato a Jimmy quando era abbastanza grande.* all this was explained to Jimmy when he.was sufficiently big.
	PE	*Tutto questo venne spiegato a Jimmy quando fu abbastanza grande.* all this came explained to Jimmy when was sufficiently big.
GPT-4o *(Highest HTER)*	ST	*She's got her own ideas.*
	HT	*Ha le sue idee.* has the her ideas.
	MT	*He le sue proprie idee.* he the his own ideas.
	PE	*Abbiamo opinioni diverse sulla cosa.* we.have opinions different on.the thing.
Mistral *(Lowest HTER)*	ST	*Ramona was one of his dad's lab technicians.*
	HT	*Ramona era uno dei tecnici di laboratorio di suo padre.* Ramona was one of.the technicians of laboratory of his father.
	MT	*Ramona era una delle tecniche del laboratorio del padre.* Ramona was one of.the technicians.FEM of.the laboratory of.the father.
	PE	*Ramona era una dei tecnici del laboratorio di suo padre.* Ramona was one.FEM of.the technicians of.the laboratory of his father.
Mistral *(Highest HTER)*	ST	*They called the cities the pleeblands.*
	HT	*Chiamavano le città plebopoli.* they.called the cities plebopolis.
	MT	*Chiamavano le città le plebe.* they.called the cities the plebs.
	PE	*Si riferivano alle città chiamandole terre di plebelandia.* they referred to.the cities calling.them lands of plebelandia.

Table 5: Examples of source text (ST), human translation (HT), machine translation (MT), and post-edited output (PE) for GPT-4o, GPT-3.5, and Mistral, showing segments with glosses and the lowest and highest post-editing effort as measured by HTER.