"Can make mistakes": Prompting ChatGPT to Enhance Literary MT output

Gys-Walt van Egdom
Utrecht University
Trans 10, 3512 JK Utrecht, the Netherlands
g.m.w.vanegdom@uu.nl

Onno Kosters
Utrecht University
Trans 10, 3512 JK Utrecht, the Netherlands
o.r.kosters@uu.nl

Christophe Declercq
Utrecht University
Trans 10, 3512 JK Utrecht, the Netherlands
c.j.m.declercq@uu.nl

Abstract

Operating at the intersection of generative artificial intelligence, machine translation, and literary translation, this paper examines to what extent prompt-driven post-editing can enhance the quality of machine-translated literary texts. We assess how different types of instruction influence post-editing performance, particularly focusing on literary nuances and author-specific styles. Situated within posthumanist translation theory, which often challenges traditional notions of human intervention in translation processes, the study explores the practical implementation of generative artificial intelligence in multilingual workflows. While the findings suggest that prompted post-editing can improve translation output to some extent, its effectiveness varies, especially in literary contexts. This highlights the need for a critical review of prompt engineering approaches and emphasizes the importance of further research to navigate the complexities of integrating AI into creative translation workflows effectively.

1 Introduction

Ever since ChatGPT was released in November 2022, the world of language automation for translation purposes – up to that point dominated by neural machine translation (NMT) (Ranathunga 2023) – has entered a new era, the paradigm shifts of which are not yet overly clear. Amid evolving roles of humans and technological processes, the lines between human and non-human translation become increasingly blurred (O'Thomas, 2017). As the need for new theoretical concepts grows, the Huxley family re-emerge (Aldous Huxley's 1932 posthuman *Brave New World* society as well as Julian Huxley's 1957 essay on posthumanism)[1]. Posthumanist theory addresses the expanding human-technology interaction and challenges traditional translation theory by reducing human intervention and pushing human expertise to the periphery of translatorial efforts. Recent advancements in NMT and generative artificial intelligence (GenAI) do indeed offer new methodologies for automating and enhancing multilingual tasks (see Lee 2023 and He 2024). Yet, within that increasing aspiration of translation automation the accurate conveyance of literary works still poses a unique challenge, one that traditional machine translation (MT) systems typically struggle to even remotely approximate (Guerberof-Arenas and Toral 2022; Macken et al. 2022). However, the integration of GenAI tools in the partly, largely or fully automated translation workflow may present a promising avenue for enhancing the quality of MT output in this domain.

At the same time, several questions remain: By combining the precision of machine algorithms with the supposed creativity of Large Language Models (LLMs), can GenAI tools offer a transformative approach to post-editing (PE) neural output? Can

[1] Posthumanist here refers to a collective concept that encompasses various critical theories, all with a shared aim of envisioning a future world that transcends the current material realities defining human existence (see also O'Thomas 2017).

Proceedings of the 1st Workshop on Creative-text Translation and Technology, pages 10–20
June 27, 2024 ©2024 European Association for Machine Translation

prompting mechanisms for GenAI learn from earlier endeavors in automatic post-editing (APE), and vice versa? To what extent can these designs provide for an increase in quality of literary translations that have gone through an automated pipeline? This paper therefore explores the posthumanist intersection of GenAI, MT and literary translation, highlighting both limitations and potential. It aims to reveal how insights from APE can aid GenAI in enhancing already machine translated text and how prompt templates in GPT-4 are an effective means to improve the quality of MT output of literary texts.

2 Related work

APE is utilized to correct MT errors automatically, to enhance the outcomes of MT system, and to reduce human editing work (Vu and Haffari, 2018; Shterionov et al., 2020; Chollampatt et al., 2020). Moreover, APE has become an invaluable methodology when addressing decoder limitations and enabling advanced text analysis beyond typical decoding capabilities (see Bojar et al. 2017).

The practice of adjusting MT output to make sense of nonsensical results has existed since the early days, when MT was also called "mechanical translation" (Bar-Hillel, 1951; Reiffer, 1952). The idea of automating PE tasks, however, remained mostly theoretical for a long period. It remained an idea awaiting the advancement of computing models capable of actualizing the concept (see, for example, Povlsen et al., 1998). This does not exclude ongoing attempts to kickstart an automated pipeline at the back end of MT output. Such a pipeline was needed in situations where initial automated quality estimation would lead to a decision mechanism determining if output should be rejected, accepted for human revision, or assigned to APE in cases of medium MT quality. It should therefore come as no surprise that for years a mature and robust APE application was sought after. Initially grounded in late and hybrid rule-based systems (e.g. Knight and Chander 1994), APE methodologies were designed to fix common mistakes in rule-based MT by capitalizing on the potential of Statistical MT (SMT) techniques. This method proved somewhat effective in addressing consistent errors (see Do Carmo et al. 2021). The proliferation of extensive datasets and the increase in computational power quickly led to a gradual shift towards statistical approaches as state-of-the-art approaches to MT. These SMT models were able to leverage bilingual corpora to identify error patterns and their corrections, signifying a pivotal move toward automation and scalability (ibid.). Within this context, APE was explored to refine SMT output through a two-stage process involving a monolingual translation phase to correct initial translation errors. A wide range of techniques was applied: from maintaining source text (ST) connections for better lexical accuracy to focusing on fluency and correcting data in case of sparsity issues (i.e. a lack of sufficient training examples). Strategies were primarily designed with a view to improving word choice and sentence structure, but they were employed with varying success; SMT models continued to falter in grasping the subtle intricacies of textual and contextual nuances (ibid.).

The concept of APE was central to two EU-funded projects of Belgium-based Crosslang. The Bologna Translation Service (ICT-PSP 270915, March 2011 – February 2013) integrated rule-based and statistical MT with translation memory and automatic and human PE (Depraetere et al. 2011), their APE-Quest (Connecting Europe Facility, project 2017-EU-IA-0151) provided a quality gate by sequencing quality estimation (QE) and APE for medium quality output into the translation workflow of the eTranslation MT system (Depraetere et al. 2020).[2] Towards the end of the 2010s, the concept of APE as a phase in iterative solutions gained attention. APE approaches witnessed a remarkable surge in popularity following their application to so-called "black-box MT systems", such as Neural MT (NMT; see Shterionov et al., 2020; Do Carmo et al. 2021).

NMT ushered in a revolutionary era for APE methodologies. Within this context, techniques have shifted towards leveraging the strengths of neural processing, such as synthetic data training, multiple-source training, and fine-tuning with advanced models like BERT. While the application of APE has diversified, addressing issues like domain adaptation and reduction in retraining needs, the core aim remains to enhance MT output (Vu and Haffari, 2018; Chatterjee 2019; Shterionov et al., 2020; Chollampatt et al., 2020). Neural techniques, particularly those using a transformer architecture, have made significant improvements in grasping context, expressing idiomatic expressions, and identifying more delicate stylistic features. Moreover, employing deep learning techniques, neural APE models seem to offer more coherent and precise corrections, addressing a broader spectrum of errors beyond simple lexical or grammatical errors.

Despite recent advancements in machine translation (MT), challenges persist, largely because neural processing, while significantly improving MT output quality, has introduced greater opacity within the

[2] The APE component for the neural MT output was based on neural copycat networks, itself based on Ive 2019.

processing systems. This "black box" phenomenon makes it difficult to pinpoint the exact locations where errors occur, complicating debugging strategies (Huang et al. 2019; Zhang & Wan 2022). Moreover, there remains an ongoing need for better evaluation metrics that can accurately reflect human judgments of accuracy, fluency and style (Van Egdom et al. 2023; Lyu et al. 2024). These developments and challenges highlight the gap that remains in place between current technological capabilities and the complex requirements of translation.

Amidst these developments in computational linguistics, it is pertinent to highlight that APE, in the traditional sense of the words, is rooted in clear-cut programming paradigms, and no linguists are directly involved in this process – a rather 'posthumanist' endeavor. However, as Generative AI continues to revolutionize the language (technology) industry (see Lyu et al. 2024), radically new forms of APE can be conceived. Novel approaches can seek active engagement of language service providers in supervised editing processes (e.g. 'interactive MT'), and take into account highly context-specific requirements of specific projects (e.g. "stylized translation" and "translation memory based MT"; see Lyu et al. 2024). For instance, this paradigm shift heralds the introduction of 'prompt engineering' within the translation profession (see Raunak et al. 2023). This phenomenon, which gives rise to 'prompted PE', can be considered as a semi-automatic approach to enhancing translation quality. Recent research underscores that the performance of Generative AI can be notably improved through directed instruction, also within the context of translation: in line with the principles of temporary in-context learning, clear and specific prompts are believed to increase the likelihood of obtaining the intended translation output (see Longpre et al. 2023).

3 Materials and Methods

To address the research question about the effectiveness of prompt templates in improving the quality of MT output, a detailed methodological strategy was developed. This strategy aims to assess systematically how structured prompts influence the performance and accuracy of prompted PE results produced by GPT-4.

3.1 Materials

The paper engaged with outputs from a source text previously leveraged in research focusing on MT quality, specifically the work of Van Egdom et al. (2023). In their research project, the output of four MT systems were examined: DeepL, Google Translate, Systran and Sig3Big (the latter being a custom Literary MT engine (CLMT) developed by Toral et al. (2020, 2021)). Over a three-year span, an annual quality evaluation was conducted to assess the development of MT engines, evaluating whether enhancements in self-learning capabilities, data volume, and algorithmic sophistication would yield improved performance over time. In this project, the Sig3Big system was excluded from periodic evaluations. As a result, ten versions derived from the same source text, "I wrote a letter…" by Donald Barthelme (524 words), were analyzed (for a detailed discussion of results, see Van Egdom et al., 2023).

For this present exploratory study, which can be considered an associated spin-off project, new outputs from the systems mentioned above were utilized, along with additional translations from LLMs powered by GPT-3.5 (in ChatGPT, free license), GPT-4 (in ChatGPT, paid license) and Gemini (Bard, free license), all generated in the fall of 2023. The prompt used to generate MT outputs with Generative AI was: "Translate into Dutch". This resulted in a set of seven unedited MT outputs that provided a baseline for the study.

This first step was followed by the generation of three different PE versions of these MT outputs under the following conditions. To generate these versions, ChatGPT was used (GPT-4, paid license). For the first set of revisions ChatGPT was prompted to follow the simple directive "As an expert translation post editor, your task is to post-edit this machine-translated Dutch translation" (condition 1). For the second set (condition 2), the instruction was further refined to draw attention to the original's literary features: "As an expert translation post editor, your task is to post-edit this machine-translated Dutch translation. Pay attention to the literary features in the ST." Under condition 3, target texts were crafted following a more detailed approach: the program was instructed to focus on the unique narrative voice of Donald Barthelme via a scaffolded prompt: Step 1: "Collect information about Donald Barthelme's unique literary style online." Step 2: "Analyze the ST and identify Barthelme's stylistic features in "I wrote a letter…". Use results of online search as a frame of reference." Step 3: "As an expert translation post editor, your task is to post-edit this machine-translated Dutch translation. Pay attention to the literary features in the source text described under step 2". In response to complex assignments (conditions 2 and 3), ChatGPT

was asked to explicitly name the steps undertaken, in order to gain insight into (issues with) reasoning. In total, 21 variations were compiled, incorporating both the original text and the unmodified MT outputs within the prompts, to ensure comprehension for the PE tasks at hand. It should be noted that iterations with identical prompts could have led to different outcomes, as each output is considered 'unique'. This variability in output under identical conditions could be said to limit the generalizability of results (Chen et al. 2023).

3.2 Methods

In translation quality assessment, methodologies typically oscillate between holistic and analytical approaches (for an overview of approaches, see Van Egdom et al. 2018). Holistic evaluation tends to view the text as a whole, focusing on the general impression the translated materials leave on the assessor. Analytical methods, on the other hand, tend to dissect the translation minutely, focusing on specific text characteristics, but this goes at the expense of the overall cohesiveness and impact of the text. In our research, an item-specific analytical method, known as the "rich point method", was adopted (for a discussion of the rich point method, see Van Egdom et al. 2018). This approach was designed to pinpoint challenges within the translation task, considering intricacies of the source text (ST), the linguistic gap separating the source and target languages involved, and the explicit information contained in the translation brief.

The selected items, or "rich points", were assessed under three main criteria reflecting critical dimensions of translation quality: accuracy, fluency, and style. These criteria were deemed instrumental in evaluating the general as well as the literary qualities of the outputs, with accuracy and fluency addressing the fundamental correctness and readability of the translation. Over the years, various frameworks for categorizing MT errors have emerged, spanning from broad classifications to more intricate systems like Multidimensional Quality Metrics (MQM) or the SCATE taxonomy. Core to the last are the broad categories of fluency and accuracy, each further divided into separate subcategories (see for instance Fonteyne, Tezcan and Macken 2020). The criterion 'style' addressed the rendering of the literary features of the ST.

The qualitative evaluation of the original MT outputs used to establish a baseline incorporated a meticulously structured analysis based on 28 ST items (see Appendix 1). These elements were deemed crucial for ensuring high-quality output: the list of items consisted of 5 items for accuracy, 7 items for fluency, and 12 items for style. The selection was conducted by two assessors with extensive literary knowledge and near-native proficiency in English and native proficiency in Dutch, in addition to a deep understanding of the relevant cultural contexts. The assessment was conducted by the same assessors, in alignment with the criteria established in the model contract for literary translations in the Netherlands, as outlined by Auteursbond & GAU (2023). Their evaluations classified the solutions into three categories: correct solutions, questionable solutions, and incorrect solutions. Solutions deemed questionable were discussed among the assessors and subsequently reclassified as either correct or incorrect. This classification laid the foundation for a nuanced qualitative analysis of the MT outputs. To ensure robustness and objectivity in the evaluation process, a third assessor, matching the first two in language proficiency and cultural knowledge, was engaged to validate the assessments made by the first assessors (i.e. to ensure inter-rater agreement). This multilayered evaluation methodology aimed to cultivate a comprehensive understanding of target text (TT) quality, grounded in a systematic analysis of text items that reflected key translation challenges in this specific context.

The second phase of the analysis involved a manual assessment conducted by our two assessors. During this phase, the assessors scrutinized the solutions found for the 28 ST items, marked them as either correct, questionable or incorrect, discussed questionable items to ensure dichotomous scoring and then employed a polytomous rating scale, 'neutral' indicating no change in quality; 'positive' denoting improvements; and 'negative' signifying deterioration with regard to the raw output. This evaluation method was designed to capture the nuances of how different instructions influenced the quality of the translated text.

By structuring the analysis in this way, our study aimed to provide a clear overview of how different levels of prompt specificity and instruction can influence the quality of prompted PE outputs. The comparative assessment of raw and PE versions, informed by detailed human evaluation, seeks to offer insights into the practical benefits and limitations of employing advanced AI-driven strategies for enhancing MT output.

4 Results

In the first stage, a detailed qualitative assessment was undertaken to set a standard for translation quality. This encompassed a systematic evaluation of outputs from 7 distinct MT systems. The assessors could award a maximum of 28 points to each text, aligning with the 28 specific items scrutinized during the

assessment process. As can be inferred from the results presented in Table 1, the quality of the unedited 'raw MT' outputs appears to be suboptimal, indicating a significant need for thorough post-editing to achieve a level of quality suitable for publication. The aggregate analysis reveals that, on average, the seven systems attained a score of 7, signifying that approximately 25% of the selected source items were accurately translated. In the dataset, two outliers can be identified. The CLMT engine achieves a fairly decent score: nearly 40% of the selected items (11/28) are correctly represented in the TT. In contrast, Systran exhibited the poorest performance, correctly translating only two items, which equates to a mere 7% of the total items. This paper also introduced evaluations of newer systems, including GPT-3.5, GPT-4, Gemini (Bard). Intriguingly, the former two displayed marginally superior performancse compared to established systems like DeepL and Google Translate, while the latter fell behind. Still, it should be noted that, despite optimism vis-à-vis LLM's potential as an MT proxy (Open AI, 2023; Raunak et al. 2023), differences were minimal.

	Accuracy (/5)	**Fluency (/7)**	**Style (/16)**	**Total (/28)**
CLMT 22	2	2	7	**11**
DeepL 23	1	3	3	**7**
Google 23	0	0	7	**7**
Systran 23	0	0	2	**2**
GPT-3.5	0	3	5	**8**
GPT-4	0	1	7	**8**
Gemini	0	1	5	**6**
Sum total raw MT	*3 (/25)*	*10 (/49)*	*36 (/112)*	*49 (/196)*

Table 1. Baseline quality of MT outputs

A positive aspect of itemized evaluation is that it provides insight into the average quality of output, but also reveals that there are various textual aspects where improvements can be observed. For example, when analyzing items concerning 'accuracy' (corresponding to 5 items in total in the ST), only the CLMT (2/5) and DeepL (1/5) systems were noted for correctly rendering items pertinent to this criterion. This shows that the qualitative analysis can be said to serve as a guidepost for targeted improvements, particularly in facets of the translations that directly impact textual accuracy.

Having established a baseline quality for unedited MT output, the study analyzed the impacts of three differentiated editing instructions. Under the first condition, ChatGPT was tasked with comprehensive PE (Full PE) of the MT outputs while considering the source content. Analysis of the data shows a general improvement in translation quality: on average, each text now correctly represents 8.29 items, marking an increase compared to the original MT output (1.29 items more than with the raw MT). Roughly 30% of source items were more accurately rendered, indicating a modest enhancement in overall quality. These results suggest that PE prompting appears to be reasonably effective, and that further specification of prompts could indeed provide additional improvements.

	Accuracy (/5)	**Fluency (/7)**	**Style (/16)**	**Total (/28)**
CLMT 22	1	3	6	**10**
DeepL 23	1	2	5	**7**
Google 23	0	4	3	**7**
Systran 23	1	3	3	**7**
GPT-3.5	0	5	3	**8**
GPT-4	1	3	6	**10**
Gemini	1	3	5	**8**
Sum Total FPE	*5 (/25)*	*23 (/49)*	*31 (/112)*	*57 (/196)*

Table 2. Output quality under condition 1 (FPE)

However, this improvement could also be said to present a complex picture. Notably, the quality enhancement is not uniform across texts: almost half of the FPE texts show quality levels similar to those of the original MT outputs (DeepL, Google, GPT-3.5). Substantial improvements can be primarily attributed to gains in performance observed in the Systran version, which jumped from two to seven correctly resolved items. Gemini and GPT-4 also showed some improvement, enhancing its score by three and two additional items. Conversely, there was one instance of a decrease in quality: after FPE, a correctly resolved item is lost in the CLMT output (score: 10).

The detailed breakdown into subcategories reveals even more nuanced results. While the 'accuracy' category demonstrates room for significant improvement, FPE versions outperform the raw MT results slightly in this respect, increasing from three to five correctly interpreted items in total (5/30). What seems noteworthy is that the CLMT output slightly regressed in terms of accuracy. In contrast, 'fluency' showed a rather marked improvement after FPE, with the number of instances in which fluency-related problems were satisfactorily resolved more than doubling (from ten to twenty-three correct instances after full PE). Despite these gains, a trade-off is observed in the 'style' category, which experienced a serious decline post-FPE. Whereas the raw MT outputs had

initially provided satisfactory solutions for style-related items 36 times, this number suddenly dropped to 31 following comprehensive PE. This trend is hardly unexpected, as research on PE guidelines shows that style improvement is rarely explicitly addressed (see Hu & Cadwell 2016). The shifts in output quality for our three subcategories underscore the inherent challenges and compromises involved in balancing the intricate elements of accuracy, fluency, and style in the process of enhancing MT texts with the aid of GenAI technology.

The second condition of the experiment focused specifically on stylistic aspects of the ST, as ChatGPT was tasked with full PE of the outputs while remaining mindful of the literary nuances of the ST. This directive was expected to enhance TT quality by making the instructions more explicit, and, more importantly, tailored to the literary purpose of the text. In theory at least, this instruction would enhance the system's in-context learning performance (Longpre et al. 2023). However, it is not superfluous to add that no specific guidelines were provided regarding the unique literary attributes that were to be preserved or highlighted, thus, leaving ChatGPT to interpret these stylistic nuances autonomously.

	Accuracy (/5)	Fluency (/7)	Style (/16)	Total (/28)
CLMT 22	0	2	5	**7**
DeepL 23	0	3	4	**7**
Google 23	1	1	2	**4**
Systran 23	0	2	2	**4**
GPT-3.5	0	4	6	**10**
GPT-4	0	3	4	**7**
Gemini	1	3	4	**8**
Sum Total FLPE	*2 (/25)*	*18 (/49)*	*27 (/112)*	*47 (/196)*

Table 3. Output quality under condition 2 (Full Literary PE)

As can be inferred from Table 3, preliminary data indicate that the overall quality of the translations does not exhibit the anticipated improvement under these tailored instructions. After implementing a focus on literary features of the original, no fewer than four out of the seven texts experience a decline in performance. On average, under this condition, the translations accurately represent approximately 6.7 out of 28 items, resulting in a meager success rate of 24%. Still, there were a number of exceptions to the rule. Systran, being the odd one out, displays marginal improvement from the base MT output: in the Full LPE version, two additional fluency-related items were rendered successfully (from two to four correct items). Similarly, the GPT-3.5 and Gemini versions show a slight uptick when it comes to performance, with enhancement observed under both fluency and style for GPT-3.5 and accuracy and fluency for Gemini.

Still, the overarching trend points to a diminution in quality. This decrease becomes even more pronounced when analyzing the remaining versions. An already limited success in conveying accuracy seen in previous conditions further regresses, with almost all items (2 in total) being misrepresented under condition 2. The odd exceptions are observed in the Google version and the Gemini version: each version managed to capture one single item for accuracy. Moreover, contrary to expectations, the 'style' category, the primary focus of this condition, witnesses a substantial downturn: initially, the raw outputs collectively presented 36 correct solutions, yet, under condition 2, this tally decreases to 27. It can be safely assumed that this reduction stems from GPT's unique interpretation of 'literariness', which seems to stray from the traditional (highly intricate) balance between form and content found in literary style, instead veering towards a more embellished, often overwrought rendition. This interpretation tends to produce what can be considered a 'pastiche' version of the ST rather than a faithful literary rendition. GPT, rather than representing the literary style specific to the ST, applies lexical choices it presumably understands as 'literary'. In doing so, it shows its inability to source beyond the overwhelming amount of stylistically unremarkable (clichéd, hackneyed) non-literary understandings of literature it can find. Nevertheless, a rather interesting observation emerges in the category 'fluency', where the literary tone of voice appears to foster fluency: this is evidenced by an increase from ten to eighteen correct translation solutions. This suggests that while attempts to infuse a literary style clearly compromises accuracy and literary authenticity, the unintentional result is an improvement in the overall fluency displayed in the texts.

In an attempt to refine the approach to stylistic fidelity, the third condition of the experiment was construed around an even more structured and detailed prompt. The prompt was divided into three stages, providing a scaffolded approach. The task involved: 1) collecting online information about Donald Barthelme's unique literary style; 2) analyzing the ST to identify Barthelme's stylistic elements in "I wrote a letter…"; and 3) utilizing this understanding during the PE process to maintain the original literary qualities (typical of Barthelme's writing) in the subsequent versions. This third instruction was aimed at guiding ChatGPT towards a deeper engagement with the literary characteristics of the ST, moving beyond a highly superficial interpretation of 'literariness'.

Surprisingly, the results presented in Table 4 show that this intensified focus led to a mere 15.3% of items being accurately resolved across the board. The Systran and the Gemini versions were the sole versions demonstrating any improvement under these author-specific directives. Gemini showed a rise to six correctly represented items (raw MT score: 2). With seven accurately rendered items, Gemini performed marginally better under the author-specific condition (raw MT score: 6). The remaining systems failed to solve more than four items correctly, suggesting a broad decline in performance.

	Accuracy (/5)	**Fluency (/7)**	**Style (/16)**	**Total (/28)**
CLMT 22	0	3	1	**4**
DeepL 23	1	1	2	**4**
Google 23	0	1	2	**3**
Systran 23	0	3	5	**8**
GPT-3.5	0	1	1	**2**
GPT-4	0	1	1	**2**
Bard (Gemini)	1	3	3	**7**
Sum Total Tailored LPE	2 (/25)	13 (/49)	15 (/112)	**30 (/196)**

Table 4. Output quality under condition 3 (Tailored Literary PE)

The breakdown of results further underscores the challenges introduced by the author-tailored instruction. Unlike the previous conditions, where some degree of improvement was noted in at least one category (accuracy under FPE, fluency under FLPE), precise and clear instruction with a focus on Barthelme's literary style had a detrimental effect on performance in all categories. Again, this decline can be attributed to several factors. Firstly, a noticeable increase in omissions can be found in the target output, with ChatGPT tending to exclude significant portions of the text (mostly toward the end of the text), resulting in a blatant loss of content, as well as a distortion and simplification of Barthelme's short story. Similar issues are observed in other studies focusing on LLMs, particularly in chain-of-thought settings (e.g. Raunak et al. 2023). LLM's are prone to not only omitting key elements but also inventing non-existent off-target content or twisting the existing information in incomprehensible ways. This phenomenon, referred to as 'edit hallucinations', compounds the distortion and simplification observed in Barthelme's short story. Furthermore, our tailored approach seemed to encourage an over-the-top form of pastiche – a kind of pastiche of the pastiche – transitioning from a general literary imitation to an unsatisfactory mimicry of Barthelme's literary style. Particularly, the nuanced balance between the mundane and the absurd that is characteristic of Barthelme's story is completely lost on GPT. In the latest iterations, this stark imbalance manifested in versions that simply veer towards the grotesque, stripping away the subtlety and nuanced banality, the hallmarks of Barthelme's narrative style. This misinterpretation, particularly evident in hyperbolic renditions of the texts, highlights the difficulties in capturing the intricate interplay of tones and themes inherent to Barthelme's oeuvre using LLMs.

5 Discussion

The findings from this study clearly reflect the challenges of prompt engineering as a means to optimize MT output through PE instructions. Reflecting on the improvement brought about by FPE, it becomes evident that while prompted PE can indeed enhance translation output – a finding that is consistent with observations made in Raunak 2023 et al. – its effectiveness seems limited and is markedly inconsistent. The experiment's venture into more tailored instructions, under the condition 'Full Literary PE' , brought to light the complexities of encoding stylistic nuances in language models. The decline observed in output quality under this condition prompts a critical reassessment of approaches to 'literariness' in Transformer architectures. The third condition's attempt to incorporate author-specific nuances into PE widened the divide separating algorithmic interpretation from literary sensibility even further.

The nuanced implications of these findings beckon a reevaluation of our expectations from prompt engineering and language models, particularly within the context of MT output optimization. Both within and beyond the academic realm, there is significant emphasis on the importance of prompt engineering and the refinement of prompts and prompt templates. While it is acknowledged that LLMs display unpredictable responses to similar prompts, there seems to be a need for precise and refined prompts and templates (see Longpre et al. 2023; Lyu et al. 2024). However, it appears that refinement, particularly in the form of instructions tailored to a literary context, currently leads to weaker output. This issue is primarily due to the tendency to beautify texts, a tendency associated within translation theoy with Berman's "ennoblement" or "popularization" (1985). The question now is whether this tendency can be suppressed through radically different or more refined instruction or specific settings (e.g. system instructions as provided through custom GPT's).

6 Conclusion

The project detailed in this paper is situated at the intersection of on the one hand posthumanist translation theory, which in itself reconsiders notions of human intervention in translation, and the practical application of GenAI in multilingual workflows on the other. With this project, we have sought to explore the potential of prompted PE, a form of semi-automatic PE, as a substitute for human PE or an intermediary step to refine MT outputs and add an additional step to translation automation in workflows. Our exploratory study scrutinized seven MT versions of a literary short story through the PE process, revealing that prompted PE, under specific conditions, yields marginal improvements. It was striking that more specific instructions, targeted toward literary translations, led to weaker performances. This outcome was quite intriguing as the view is widely held that prompt specificity is a driver of performance in AI-driven tasks, such as language translation (Longpre et al. 2023).

Still, it is crucial to acknowledge the preliminary nature of these findings. As with much research in the nascent field of Generative AI and translation, our study faces limitations that underscore a great need for further exploration. From a fundamental point of view, different takes on 'literariness' and 'style' can be applied to measure the creative prowess of GenAI (see Boase-Beier 2020). For a more comprehensive understanding of the 'literariness' of PE outputs, future research should also include a greater variety of literary genres and styles. Additionally, there is a great need to expand research on the effects of prompted PE across a broader spectrum of languages (as in Lyu et al. 2024). Finally, adverse effects of prompted PE might be mitigated when using different prompting strategies than the ones used in this study. To counteract observed 'pastiche effect', example-based prompts, laying down clear criteria for the tone and the expected levels of faithfulness to the original, can be explored. Another avenue for future research in the domain of literary translation is investigating the effects of customizing GPT's using domain-specific language resources such as translation memories (see Zhang and Wan 2022).

Recent advancements in language automation have illuminated the potential of AI integration into linguistic workflows, not in the least in creative text domains. However, amidst the hype surrounding GenAI, the intrinsic complexity of creative tasks (e.g. literary translation) often gets overlooked or oversimplified in research in computational linguistics and translation studies. Despite the critical acclaim for AI's creativity and the benefits of human-language prompting, our research has shown that it is and will always remain crucial to ensure a tight alignment between creativity and fidelity in the context of creative translation.

References

Bar-Hillel, Y. 1951. The Present State Of Research On Mechanical Translation. In *American Documentation*, 2(4): 229-237.

Barthelme, Donald. 1992. I wrote a letter… In *The Teachings of Don B.* (pp. 11-12). Berkeley : Counterpoint.

Berman, A. 1985. La traduction et la lettre ou l'auberge du lointain. In *Les Tours de Babel* (pp. 31-85). Mauvezin: Trans-Europ-Repress.

Boase-Beier, J. 2020. *Translation and Style* (2nd edition). London : Routledge.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., & Turchi, M. 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.

Chen, B., Zhang, Z., Langrené, N, & Zhu, S. 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review. ArXiv: https://arxiv.org/pdf/2310.14735

Chatterjee, R., Federmann, C., Negri, M., & Turchi, M. 2019. Findings of the WMT 2019 shared task on automatic post-editing. In P*roceedings of the Fourth Conference on Machine Translation* (Volume 3: Shared Task Papers, Day 2), pages 11–28. Florence, Italy: Association for Computational Linguistics.

Chollampatt, S., Susanto, R. H., Tan, L., & Szymanska, E. 2020. 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.

Do Carmo, F., Shterionov, D., Moorkens, J., et al. 2021. A review of the state-of-the-art in automatic post-editing. *Machine Translation*, 35: 101–143. https://doi.org/10.1007/s10590-020-09252-y

Depraetere, H., Van den Bogaert, J., & Van de Walle, J. 2011. Bologna translation service: Online translation of course syllabi and study programmes in English. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, pages 29-34. Leuven, Belgium, May.

Depraetere, H., Van Den Bogaert, J., Szoc, S., & Vanallemeersch, T. 2020. APE-QUEST: An MT quality gate. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 473-474.

Do Carmo, F., Shterionov, D., Moorkens, J., et al. 2021. A review of the state-of-the-art in automatic post-editing. *Machine Translation*, 35: 101–143.

Fonteyne, M., Tezcan, A., & Macken, L. 2020. Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In *12th International Conference on Language Resources and Evaluation (LREC)*. *European Language Resources Association (ELRA)*, pages 3783-3791.

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

He, S., 2024. Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts. *arXiv preprint* arXiv:2403.00127.

Hu, K., & Cadwell, P. 2016. A comparative study of post-editing guidelines. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 34206–353.

Huang, X., Liu, Y., Luan, H., Xu, J., & Sun, M. 2019. Learning to copy for automatic post-editing. 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 6122–6132. Hong Kong, China: Association for Computational Linguistics.

Huxley, A. 1932. *Brave New World*. London: Chatto and Windus.

Huxley, J. 1957. Transhumanism. In *New Bottles in New Wine*, London: Chatto and Windus, pages 13-18. https://archive.org/details/NewBottlesForNewWine/page/n7/mode/2up

Ive, J., Madhyastha, P. S., & Specia, L. 2019. Deep copycat networks for text-to-text generation. 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 3227-3236.

Knight, K., & Chander, I. 1994. Automated postediting of documents. In Proceedings of the 12th National Conference on Artificial Intelligence (AAAI 1994), Vol. 1: 779–784. Seattle, Washington, USA.

Lee, T.K., 2023. Artificial intelligence and posthumanist translation: ChatGPT versus the translator. *Applied Linguistics Review*, https://doi.org/10.1515/applirev-2023-0122.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., & Roberts, A. 2023. The Flan Collection: Designing data and methods for effective instruction tuning. arXiv. https://arxiv.org/abs/2301.13688

Lyu, C., Du, Z., Xu, J., Duan, Y., Wu, M., Lynn, T., Aji, A. F., Wong, D. F., Liu, S., & Wang, L. 2024. A paradigm shift: The future of machine translation lies with large language models. arXiv. https://arxiv.org/abs/2305.01181

Macken, L., Vanroy, B., Desmet, L., & Tezcan, A. 2022. Literary translation as a three-stage process: Machine translation, post-editing and revision. In *23rd Annual Conference of the European Association for Machine Translation. European Association for Machine Translation*, pages 101-110.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., & Mian, A. S. 2023. A Comprehensive Overview of Large Language Models. ArXiv, abs/2307.06435. Retrieved from https://api.semanticscholar.org/CorpusID:259847443

O'Thomas, M. (2017). Humanum ex machina: Translation in the post-global, posthuman world. *Target* 29(2): 284-300.

Povlsen, C., Underwood, N. L., Music, B., & Neville, J. 1998. Evaluating text-types suitability for Machine Translation: a case study on an english-danish MT System. In *LREC*, pages *27-34.*

Ranathunga, S., Lee, E. S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., & Kaur, R. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys, 55(11): 1-37.*

Raunak, V., Sharaf, A., Wang, Y., Awadalla, H., & Menezes, A. 2023. Leveraging GPT-4 for Automatic Translation Post-Editing. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 12009–12024. Singapore: Association for Computational Linguistics.

Reifler, E. 1952. Mechanical translation with a pre-editor, and writing for MT. In *Proceedings of the Conference on Mechanical Translation.*

Shterionov, D., Do Carmo, F., Wagner, J., Hossari, M., Paquin, E., & Moorkens, J. 2020. A roadmap to neural automatic post-editing - an empirical approach. *Machine Translation*, 34: 67–96.

Shterionov, D., Wagner, J., & Do Carmo, F. 2019. APE through neural and statistical MT with augmented data. ADAPT/DCU submission to the WMT 2019 APE shared task. In *Proceedings of the Fourth Conference on Machine Translation (WMT2019), Volume 3: Shared Task Papers*, pages 132–138. Florence, Italy.

Van Egdom, G.W., Verplaetse, H., Schrijver, I., Kockaert, H., Segers, W., Pauwels, J., Bloemen, H. & Wylin, B. (2018). How to put the translation test to the test? On preselected items evaluation and perturbation. In *Quality Assurance and Assessment Practices in Translation and Interpreting* (pp. 26-56). Hershey [MA]: IGI Global.

Van Egdom, G.W., Kosters, O., & Declercq, C. 2023. The Riddle of (Literary) Machine Translation Quality: Assessing Automated Quality Evaluation Metrics in a Literary Context. *Revista Tradumàtica*, 21: 129–159.

Vu, T.-T., & Haffari, G. 2018. Automatic post-editing of machine translation: A neural programmer-interpreter approach. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,

pages 3048–3053. Brussels, Belgium: Association for Computational Linguistics.

Zhang, X., & Wan, X. 2022. An empirical study of automatic post-editing. arXiv. https://arxiv.org/abs/2209.07759

Appendix 1. Source text items and corresponding analytical criteria

	Item	Criterion
1	, asked him	Style - colloquialism
2	towaway zones	Accuracy
3	and I didn't like it	Fluency
4	Cost me …, plus	Style - colloquialism
5	tiny little cars	Style - colloquialism
6	You ever notice …? You ever seen…? No you haven't.	Style - colloquialism [Fluency]
7	, and to keep some mental health warm …,	Style - colloquialism [Fluency]
8	a bucket of ribs	Accuracy
9	Which I would gladly carry up there…	Fluency - idiom
10	I cabled him	Style - absurdism [Accuracy]
11	and, by the way, what was the apartment situation up there?	Style - colloquialism [Fluency]
12	It was bad,	Fluency - idiom
13	he replied by platitudinum plate	Style - absurdism [Accuracy]
14	but what could he do?	Fluency - idiom
15	root cellar	Accuracy
16	'cause of me being a friend of the moon.	Style - colloquialism
17	pretty nice place	Fluency
18	the Space Shuttle Hurry-Up Fund	Accuracy
19	Drumming fiercely on a hollow log with a longitudinal slit tuned to moon frequencies	Style - absurdism [Accuracy]
20	employment, medical coverage, retirement benefits, tax shelterage, convenience cards, and Christmas Club accounts	Accuracy
21	That's a roger,	Fluency - idiom
22	he moonbeamed back	Style- absurdism [Accuracy]
23	by means of curly little ALGOL circuits I had knitted myself on my Apple computer	Style(absurdism), [Accuracy]
24	that ticktacktoe was about as far as they'd got in that direction	Style - absurdism [Accuracy]
25	via flights of angels with special instructions	Style – absurdism [Accuracy]
26	it looked to me like he had things pretty well in hand up there	Fluency
27	Part-time if need be?	Style – colloquialism
28	a shower of used-car asteroids with blue-and-green bumper stickers	Style – absurdism [Accuracy]