Style and Terminology in Commercial Flows: Mixing APE with Iterative Feedback

Marthe Lamote **Ewoenam Tokpo** **Tom Vanallemeersch**
Sara Szoc **Koen Van Winckel**
CrossLang
{firstname.lastname}@crosslang.com

Abstract

MT systems have demonstrated strong capabilities in natural language translation, especially in the era of multilingual large language models, which exhibit impressive text understanding and generation capabilities across multiple languages. However, meeting specific stylistic and lexical requirements of end users remains challenging. In this paper, we explore the effectiveness of an automated post-editing approach (APE) in a real-world MT production environment using LLMs and iterative human feedback. We further examine current limitations and potential areas for improvement. To this end, we propose an end-to-end methodology for APE and provide analytical and empirical insights into challenges of APE in commercial settings, primarily relying on client and expert feedback and analysis. We analyze the performance of LLM-based APE across five main stylistic dimensions, assess the terminology correction component of the pipeline, and demonstrate how iterative feedback, through prompt refinement and glossary enhancement, contributes to improving both stylistic and terminological consistency.

1 Introduction

Machine Translation (MT) systems have shown considerable adeptness in translating natural language text (Zhu et al., 2024b; Alves et al., 2024), especially in the era of multilingual large language models (MLLMs), which are increasingly exhibiting impressive capabilities in understanding and generating texts in multiple languages. In spite of these advances, it remains challenging to generate translations that meet specific stylistic and lexical requirements of end users (Raunak et al., 2023). This limitation is even more prominent in out-of-the-box pretrained translation systems.

To address the above limitation, MT systems are often finetuned on user-specific data to adapt the model to stylistic and lexical preferences of users (Finkelstein et al., 2026). Finetuning a model, however, imposes significant resource demands that are often not available in run-time environments. Additionally, finetuning, if not properly carried out, could cause catastrophic forgetting of useful knowledge previously learned by the model (Zhai et al., 2024; Luo et al., 2023).

Another potential solution to this issue is in-context learning (ICL). This generally involves selecting a set of translation examples to demonstrate the task (Dong et al., 2024). These task exemplars are appended to the input prompt to form part of the context for text generation. Although ICL does not impose significant resource demands, it is sensitive to the choice of exemplars, so that poor exemplars could have a detrimental effect on the output quality (Zhu et al., 2024a). This makes ICL challenging to deploy in situations where a suitable translation memory is not available.

In practice, some type of post-editing often needs to be carried out to provide fine-grained stylistic and lexical adaptations to suit user preferences. The use of LLMs for automatic post-editing (APE) has recently been adopted in localization frameworks (Uguet et al., 2024).

In this paper, we explore the effectiveness of an

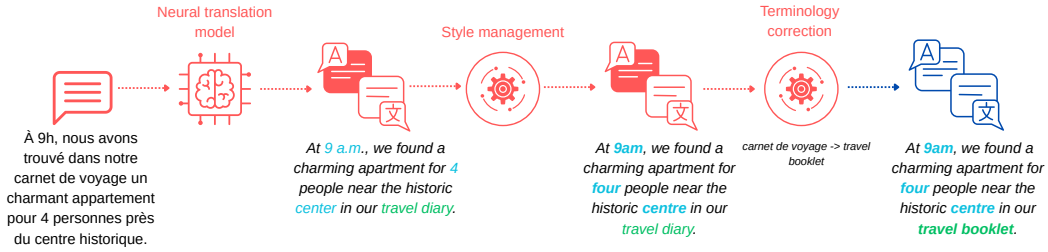

Figure 1: A depiction of our end-to-end APE pipeline with an example translation. Words that undergo style adaptations are shown in blue, whereas words that undergo terminology correction are shown in green. Correctly resolved words according to user preferences are in **bold**.

APE approach using LLMs and iterative human feedback in a real-world production setting, its current limitations, and possible improvements. To this effect, we (1) propose an end-to-end methodology for APE, showing how different phases of post-editing can be effectively automated using LLMs (Figure 1) and (2) provide some analytical and empirical insights on challenges of APE in commercial settings, primarily relying on client and expert feedback in the process. We used GPT-4.1 for all LLM implementations in this project. This choice is motivated by the fact that this model is a highly optimized general-purpose model designed for speed and responsiveness, which was an important consideration for this project.

2 Related works

Recent works have explored the use of LLMs for APE. Raunak et al. (2023) explore the use of GPT-4 to post-edit outputs from an NMT system. Their work indicates the capacity of GPT-4 to produce meaningful edits that align with human judgement. They report an improvement on WMT-22 in four language directions based on COMET scores. In spite of these score improvements, they note a tendency for the GPT-4 model to hallucinate by proposing edits in its chain of thought that it does not incorporate into the final post-edited translation. Uguet et al. (2024) also show in their work that LLMs demonstrate great potential in APE but are not fully mature in their current state for this purpose; instead of improving the raw MT translations, they consistently deteriorated the unedited "raw" translation outputs in evaluation metrics such as BLEU (Papineni et al., 2002).

Ki et al. (2024) investigate LLaMA-2 7B for APE. They show reliable improvements for results in Chinese-English, English-German and English-Russian. However, they show that the LLM struggles to incorporate fine-grained instructions; rather, it focuses on broad and generic instructions for post-editing. Chen et al. (2024) use an iterative approach to refine translations. An LLM, over multiple rounds, refines the translation to improve the naturalness of the translation. Their approach is based on the observation that humans typically modify translations repeatedly to improve them. They report that this approach improves fluency and naturalness, especially for high-resource languages.

The above works primarily explore the use of APE with standard benchmark datasets, which often do not reflect and reveal frequently encountered real-world challenges. Hence, our aim is to investigate the effectiveness of APE with LLMs in an MT production environment, particularly in commercial settings. Nunziatini and Diego (2024) explore post-editing translations in a production environment but limit their scope to gender-inclusive post-edits. They show GPT-4's capability of generating post-edits that mitigate gender stereotypes, but point out the tendency of the model to make unnecessary and inconsistent changes. Closest to our line of work is Vidal et al. (2022), who give a brief presentation on the possibilities of incorporating LLMs in an APE workflow in a real-word production environment. Our aim is to take this further by exploring and proposing a more detailed APE workflow in production environments, focusing on stylistic and terminological adaptations. We also provide a fine-grained and in-depth analysis of APE based on human feedback and analysis.

3 APE Approach

In this section, we describe the end-to-end pipeline that makes up our automated approach to post-editing. Our APE pipeline consists of style management, terminology correction, and iterative human feedback.

3.1 Style Management

The style management component of our APE, as depicted in Figure 2, begins with the conversion of the client style guide containing the language preferences and requirements into concise instructions that can be integrated into our APE architecture. The rules of the style guide, initially contained in a PDF style guide, are extracted and converted to instructions using the LLM.

After this initial extraction stage, a manual validation process is conducted by our internal prompt engineer and linguist. This process is essential in ensuring that all necessary rules are properly captured and that the instructions are clear, concise, non-ambiguous, and understandable. The output is engineered into a prompt for the LLM using various prompt engineering strategies in a sandbox environment, where a sample of the client's data is used for prompt testing. An example of the automatic conversion and manual validation of a style guide section can be found in the appendix.

3.2 Terminology Correction

The terminology correction step begins with the conversion of the glossary into a predefined format, which also allows for the addition of notes or metadata (such as part of speech) to guide the LLM on how to apply specific term translations, especially in instances where the usage of the term can be ambiguous.

The formatted glossary is used to identify segments in the source corpus containing relevant keywords. Segments selected through this glossary lookup are sent to the term correction component. For this step, the source sentence and the translation are added to the prompt with the identified glossary terms and, when available, the terms' metadata. The prompt is set up to direct the term correction model to integrate the terms in the sentence without breaking already correct translations and to ensure final grammatical correctness.

3.3 Iterative human feedback

The APE pipeline supports flexible adaptation of both the prompts and glossary resources throughout the project lifecycle, guided by ongoing linguistic feedback. Recurring issues identified in the APE output are systematically analyzed by the linguist and prompt engineer, and the system configuration is adjusted accordingly. When errors originate from the style management step, prompt refinements can be applied to better steer the model's behavior. Issues arising from the term correction step can be mitigated by enriching the glossary with additional metadata to more effectively guide the model's application of terminology. This iterative feedback loop enables continuous improvement of the system and ensures sustained alignment with evolving client requirements.

4 Experimental setup

As mentioned above, the main objective of this work is to investigate the efficacy of an APE pipeline with LLMs as the core component in a *real-world industry setting*. To this end, we use an existing dataset from a large-scale French-to-English project carried out for a client in the travel industry. The original dataset consists of a total of almost 49,738 segments from 118 different input files, consisting of website content and travel descriptions.

To allow for a thorough and qualitative analysis of how client style rules are applied across different stylistic dimensions, we downscaled the dataset to a sample of 250 instances. This sample was extracted after randomly shuffling the dataset using Excel's rand() function. Upon inspection, it showed that the extracted segments covered most of the input files, thus creating a representative sample of the full project. After a preprocessing step to remove extraneous instances consisting solely of non-translatable content, a sample of 242 lines remained.

The source texts are used as inputs to our proposed APE pipeline, saving the outputs from each intermediate step in the pipeline for qualitative analysis, as depicted in Figure 1.

The qualitative analysis was performed by the linguist assigned to the project, who has extensive knowledge of the client style guide and glossary. In addition, we conducted a quantitative evaluation using five established MT metrics: BLEU, TER (Snover et al., 2006), ChrF (Popović,

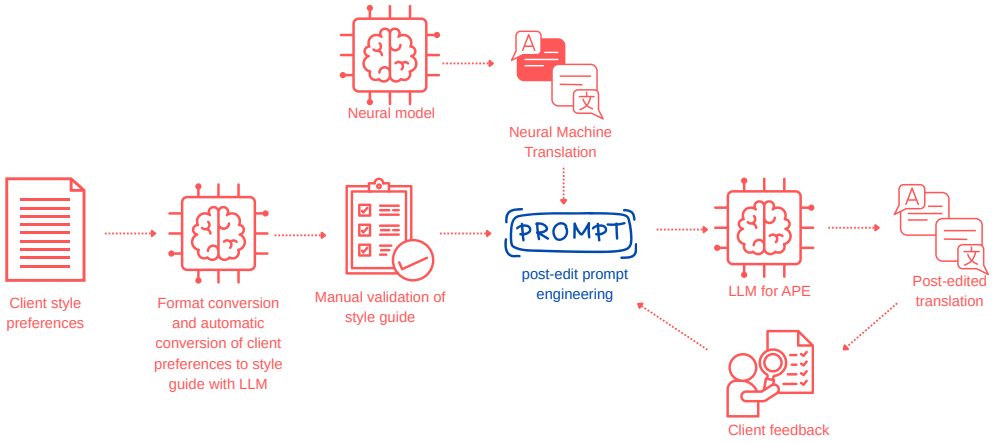

Figure 2: An illustration of the style management pipeline, showing the connections between the various steps.

2015), COMET (Webber et al., 2020), and BertScore (Zhang et al., 2019).

We use *gpt-4.1-2025-04-14* for all LLM implementations in this work, with a temperature value of 0.2 and a maximum output length of 200 tokens. All other configurations are set to OpenAI's default values. We use gpt-4.1-2025-04-14 for the initial "raw" translations as well.

5 APE Analysis

We analyze the quality of APE outputs generated by GPT-4.1 across five main stylistic dimensions. These dimensions, discussed below, are derived from the client style guide and serve as the framework for our analysis. Statistics of the APE outputs relating to these dimensions are provided in Table 1. We will further discuss this in Section 5.2 In addition to these dimensions, we also assess the term correction component of the APE pipeline, which is likewise implemented using GPT-4.1. Lastly, we will examine the benefits of iteratively integrating human feedback.

5.1 Stylistic dimensions

Language and form This stylistic dimension primarily covers the required language variant and general instructions aimed at ensuring structurally sound and grammatically correct translations, with an emphasis on clarity and simplicity. In this use case, the translation is required to be in natural clear British English with accessible wording and well-formed sentence structures, adhering to conventions outlined in the Guardian Style Guide[1].

[1]https://www.theguardian.com/info/series/the-guardian-style-guide

Formatting and punctuation This dimension addresses instructions related to formatting and punctuation, encompassing a wide range of conventions derived from both general language standards and client-specific preferences. These include the use of colons, quotation marks, dashes, and commas[2], as well as formatting rules for numbers, dates, times, units of measurement, and currencies.

Content conventions This dimension covers guidelines governing the translation of foreign place names, currencies, and abbreviations.

Inclusive and sensitive language This dimension comprises guidelines requiring inclusive, non-stereotyping, and bias-aware language, with particular attention to gender-neutral and accessibility-conscious formulations.

Tone and register This last dimension relates to tone and register, defining the desired voice of the translation. This category differs from the previous ones in that its instructions—such as "Use an immersive, authentic, and engaging travel writing style" and "Preserve local color"—are inherently less explicit and therefore more difficult to operationalize and evaluate systematically.

5.2 Stylistic analysis

Language and form In general, the baseline translation exhibits a high level of grammatical accuracy with relatively few issues identified. Nevertheless, we identify eight segments that are modified by the APE system with respect to this stylistic

[2]This also includes the Oxford comma, which is the last commas before the coordinating conjunction (such as *and* or *or*) in lists. Its use is a matter of style preference and is generally not preferred by the client.

Stylistic dimensions	Modified instances	Improvement	Deterioration	Non-functional modification	Unapplied cases
Language & Form	8	7	1	0	0
Formatting & Punctuation	45	45	0	0	1
Content Conventions	4	1	3	0	0
Inclusive & Sensitive Language	1	1	0	0	0
Tone & Register	117	14	3	100	0

Table 1: Statistics of modifications made by the APE pipeline regardind the five stylistic dimensions.

dimension. Six segments feature minor edits that improve grammatical correctness and enhance the fluency of the translation, as is the case in example (1). One of the remaining segments involves the correction of a word to its British spelling. The last segment, however, is a less desirable instance in which a less common and archaic British variant is introduced.

(1) ***Source***: *Vous prenez la direction de l'ouest, avec derrière vous les vagues de l'Atlantique et leur grondement solennel.*
Raw MT: *You head west, leaving behind you the waves of the Atlantic and their solemn rumble.*
APE: *You head west, **leaving behind** the waves of the Atlantic and their solemn rumble.*

Formatting and punctuation The most frequently observed interventions in the analyzed sample include the following:

- Removal of redundant Oxford comma: the Oxford comma was removed when not required for clarity.
- Conversion of units of measurement: as the French source text employs the metric system, the corresponding imperial units were systematically added in parentheses in the translation.
- Parenthetical clauses: en dashes with surrounding spaces were used to replace em dashes without spaces in parenthetical constructions.
- Number formatting: numbers from one to ten were spelled out, while numbers above ten were expressed in numerals.

These guidelines are implemented during the style management step, resulting in the correction of formatting and punctuation in 46 segments, one of which is shown in example (2). One instance is identified in which a formatting guideline is not applied as expected. In this segment, numbers below ten are rendered as numerals instead of in their spelled-out form. This represents the only false negative observed. Notably, the segment contains both HTML tags and template variables, suggesting that the failure to apply the style rule may be an anomaly caused by the presence of such elements.

(2) ***Source***: *Ruelles escarpées, galeries et balcons de bois ouvragé, le vieux centre a été réhabilité : on se croirait au XIIXe siècle.*
Raw MT: *Steep alleyways, galleries, and crafted wooden balconies—the old town centre has been revitalized: it feels like stepping back into the 18th century.*
APE: *Steep alleyways, galleries and crafted wooden balconies – the old town centre has been revitalized: it feels like stepping back into the 18th century.*

Content conventions In the analyzed sample, four segments are identified in which these instructions affect the translation output. Of these, one segment shows an improvement over the baseline translation, whereas three segments reflect a deterioration.

All four instances could be traced back to the instruction "Verify and use official English names for places/companies where they exist; otherwise, preserve the local name". The positive example involves a French brand name that was initially translated into English by the baseline MT system, which the APE step correctly restores to the French name, as no official English equivalent exists for the brand in question. In the three negative examples, proper nouns were first rendered in their correct English forms by the baseline MT, after which the APE step reverts them to their French spellings. An example is shown in (3).

(3) ***Source***: *Excursion sur les pas des pharaons, à Guizeh et Saqqarah.*

***Raw MT**: An excursion following in the footsteps of the pharaohs, at Giza and Saqqara*
***APE**: An excursion following in the footsteps of the pharaohs, at Giza and Saqqarah.*

Inclusive and sensitive analysis In this dimension, there is one instance in the analyzed sample where the inclusive language guidelines prompt a modification; the example is shown in (4). The style guide specifies the use of gender-neutral language where appropriate, recommending singular *they* in cases of unknown gender and the avoidance of unnecessarily gendered nouns (e.g., using *postal workers* instead of *postmen*). In the observed example, the French term *cavalier* is translated as *horseman* by the baseline MT system. This is subsequently revised to the gender-neutral variant *horse rider* during the style management step. In addition, the associated possessive pronouns are adjusted from *his* to *their*, ensuring consistency with the prescribed inclusive language conventions.

(4) ***Source**: Le cavalier mongol avec son aigle au bras est à la fois un symbole national et une figure de l'imaginaire collectif.*
***Raw MT**: The Mongolian horseman with his eagle on his arm is both a national symbol and a figure in the collective imagination.*
***APE**: The Mongolian **horse rider** with **their** eagle on **their** arm is both a national symbol and a figure in the collective imagination.*

Tone and register The analysis of the selected sample identifies 117 segments in which the modifications cannot be clearly attributed to any specific instruction from the defined stylistic dimensions. While some of these changes may stem from tone-related instructions, this cannot be established with certainty.

Among these segments, 17 demonstrates meaningful changes to the baseline, of which 14 constitute improvements and 3 result in a deterioration of translation quality. The remaining 100 segments involve changes that neither correspond to identifiable stylistic instructions nor produce an objective difference in meaning or quality.

This tendency is illustrated by cases involving purely preferential lexical substitutions, such as replacing *prefers* with *favours* and subsequently reverting *favour* back to *prefer*, as is shown in example (5).

(5) ***Source**: Votre guide sait quel acacia le guépard préfère ces jours-ci, qu'une hyène a récemment mis bas à deux petits adorables, que les éléphants favorisent ce point d'eau plutôt qu'un autre ...*
***Raw MT**: Your guide knows which acacia the cheetah prefers these days, that a hyena has recently given birth to two adorable cubs, and that the elephants favour this waterhole over another ...*
***APE**: Your guide knows which acacia the cheetah favours these days, that a hyena has recently given birth to two adorable cubs, and that the elephants prefer this waterhole to another ...*

As also observed by Nunziatini and Diego (2024), the style management step appears prone to introducing unnecessary preferential modifications, despite explicit prompt constraints such as "Do not change phrasing or word choices unless required by these guidelines."

A closer examination of these instances reveals a notable pattern: in 95 out of the 100 cases involving unnecessary and non-functional changes, such modifications occur in segments that do not require any stylistic revision. Only five cases exhibit co-occurrence with genuinely required, guide-line-driven revisions. This distribution suggests that the model predominantly introduces preferential changes in segments where no rule-based intervention is required, indicating a tendency to "over-edit" in the absence of clear stylistic triggers.

5.3 Term correction

Out of the total of 242 segments in the sample, 88 contain at least one term listed in the glossary. Not all of these instances require intervention, however, as the baseline MT output often already aligned with the prescribed terminology. This is particularly common in the travel domain, which is relatively non-technical: in 40 out of the 88 segments, no correction is necessary, and the translations are therefore left unchanged. In the remaining 48 segments, the term correction step modifies the output to ensure compliance with the glossary.

The majority of the term correction interventions are successful, but seven instances are identified in which glossary entries are applied incorrectly. In all cases, the issue involves relatively common terms for which the glossary prescribes

a more specific translation that is not appropriate in every context. In example (6), the French noun *salon* is associated in the glossary with the translations *trade fair* and *show*. While valid in certain contexts, these translations do not cover the full semantic range of the noun in French. As a result, the glossary is applied outside its intended scope, leading to incorrect translations—such as rendering *salon* as *trade fair* or *show* in contexts where it instead refers to a lounge.

(6) ***Source**: Bassin à l'ombre de la cour intérieure, Honesty Bar au **salon**, allées éclairées à la bougie, tout ici évoque l'élégance.*
***Raw MT**: A pool in the shade of the courtyard, an honesty bar in the lounge, candlelit paths—everything here exudes elegance.*
***APE**: A pool in the shade of the courtyard, an honesty bar in the trade fair, candlelit paths – everything here exudes elegance.*

5.4 Iterative human feedback

Section 3.3 conceptually described how recurring issues identified in the APE output can be addressed through targeted modifications to the prompt and/or by enriching the glossary entries with additional metadata. To evaluate the effectiveness of this feedback loop in practice, the sample was reprocessed using an updated prompt and revised glossary designed to mitigate the issues outlined above. This allows for a comparative assessment of whether the proposed adjustments lead to measurable improvements in translation quality and system behavior.

The analysis of the initial APE outputs reveals 13 segments in which translation quality deteriorates following the style management step, as well as 6 instances in which the term correction results in mistranslations.

The update to the style prompt followed involved the addition of explicit guidelines targeting specific stylistic issues, such as clusters of Latinate nouns and zeugmatic constructions in lists. The most significant modification, however, is a shift in overall prioritization: the revised prompt places greater emphasis on natural, idiomatic target-language expression rather than strict adherence to source-text structure.

The translations of all 13 segments that previously exhibited a decline in quality following the style management step were improved with the updated prompt, resulting in more fluent and idiomatic renderings across all cases. One such improvement is visualized in example (7).

(7) ***Source**: Aux alentours, le Piémont multiplie les séductions, de l'élégante capitale Turin au charme provincial d'Alba ou d'Asti, le tout porté par une gastronomie de haute volée.*
***Previous APE**: All around, Piedmont multiplies its allure, from the elegant capital Turin to the provincial charm of Alba or Asti, all carried by high-flying gastronomy.*
***Improved APE**: All around, Piedmont **offers countless attractions**, from the elegant capital Turin to the provincial charm of Alba or Asti – all carried by **outstanding cuisine**.*

With regards to the glossary-related issues, five out of the seven previously identified errors were resolved through the addition of enriched metadata designed to better guide the model's selection and application of terminology. In this way, example (6) is improved, as the terminology step preserves the appropriate translation suggested by the baseline MT system. Two instances remained incorrect, in which the model selected the less idiomatic option out of two available glossary translation options despite the instruction to "choose the best translation according to the context".

Apart from the two remaining errors, all previously identified issues were successfully addressed through prompt refinement and glossary enhancement, demonstrating the effectiveness of the feedback loop. These results further highlight the value of iterative collaboration between linguists and prompt engineers in the development and refinement of robust automatic post-editing pipelines.

6 Quantitative Evaluation

In Table 2, we show the results of the quantitative evaluation of our APE approach[3]. We evaluate the outputs of the style management and terminology correction components to examine how much improvement each corresponding step contributes.

We first observe a strong correlation between the five adopted metrics: COMET, BLEU, TER, ChrF, and BertScore. The latter four particularly showed perfect correlation, further validating the process.

[3]We draw the readers' attention to the fact that the target references used in the quantitative evaluation, while being human-edited, were adapted from outputs of a previous run of the APE pipeline. See discussion in the Limitations section.

system	BLEU	TER	ChrF	COMET	BertScore
Raw MT	60.044	34.410	75.119	0.873	0.775
Style Management	62.489	32.630	76.404	0.873	0.777
Style & Terminology	62.892	32.345	76.786	0.873	0.780

Table 2: Comparison of raw MT and APE outputs.

Generally, the various metrics show a consistent improvement at each step in our pipeline. We see a relatively big improvement in score after the style management step, and a further, although less significant, improvement after the terminology correction step.

We observe no considerable change in COMET scores across the three outputs. This is likely due to the fact that COMET, being a semantic metric, is relatively insensitive to certain stylistic and terminological changes that do not affect the semantics of the text.

7 Conclusion

In this work, we investigate the effectiveness of an APE approach based on GPT-4.1, enhanced through an iterative feedback loop, in a real-world commercial setting. We propose an end-to-end APE approach to this effect and carry out a detailed qualitative analysis across five stylistic dimensions, as well as the terminology correction component.

Our analysis of the five stylistic dimensions indicates that using GPT-4.1 in an APE pipeline considerably improves translations in the language and form, formatting and punctuation, and inclusive and sensitive language dimensions. However, it struggles to improve translations with respect to the content conventions. It is difficult to ascertain whether the improvements with regard to tone and register can be directly attributed to the style guide or whether these modifications are influenced by the intrinsic style of the LLM.

Our analysis of the term correction component shows that it reliably preserves terms that are already correctly translated and effectively corrects instances where the baseline MT output does not comply with the glossary. At the same time, it exhibits a tendency to over-apply glossary entries for relatively common terms whose prescribed translations are context-dependent. In such cases, the system may enforce a specific glossary equivalent even when it is not appropriate for the given context.

The above-mentioned issues can be mitigated by prompt refinement and glossary enhancement based on linguistic feedback. Our findings indicate that the human feedback loop is pivotal in an APE pipeline, allowing the system to be systematically improved to address reported issues and reflect the exact preferences of the client, as well as to adapt to the dynamic business environment of clients.

Finally, we show, through quantitative evaluation using five industry-established metrics, how the various components of our APE architecture consistently improve translations.

Limitations

A real-world test case introduces several challenges, reflecting the operational realities of commercial localization workflows. Chief among them in this work is obtaining target references for quantitative evaluation. As earlier described, in this use case, references were obtained through human post-edits on outputs from a previous run of the APE system, which may introduce a bias in favor of our APE system. Additionally, although the use of standard MT quality metrics such as BLEU are effective at capturing overall improvements and terminological corrections, they may not adequately capture fine-grained stylistic improvements by the setup. Nevertheless, the detailed qualitative analysis carried out in this work gives an outlook that is consistent with the quantitative results, thus further confirming our conclusions.

Secondly, due to the proprietary and client-specific nature of the project, the full prompts used for our experiments and analysis are not disclosed. Consequently the replicability of this work is affected.

Ethics statement

To ensure ethical standards and protect client anonymity, all examples in this work have been modified. Care has however been taken to ensure that these modifications do not alter the core content that underpins the analysis and conclusions.

References

Alves, Duarte M, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Chen, Pinzhen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190.

Dong, Qingxiu, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA, November. Association for Computational Linguistics.

Finkelstein, Mara, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, et al. 2026. Translategemma technical report. *arXiv preprint arXiv:2601.09012*.

Ki, Dayeon and Marine Carpuat. 2024. Guiding large language models to post-edit machine translation with error annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4253–4273.

Luo, Yun, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. corr abs/2308.08747 (2023). *arXiv preprint arXiv:2308.08747*.

Nunziatini, Mara and Sara Diego. 2024. Implementing gender-inclusivity in mt output using automatic post-editing with llms. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 580–589.

Papineni, Kishore, 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.

Popović, Maja. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Raunak, Vikas, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024.

Snover, Matthew, Bonnie Dorr, Richard 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.

Uguet, Celia, Fred Bane, Mahmoud Aymo, João Torres, Anna Zaretskaya, and Tània Blanch Miró Blanch Miró. 2024. Llms in post-translation workflows: comparing performance in post-editing and error analysis. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 373–386.

Vidal, Blanca, 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.

Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu. 2020. Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhai, Yuexiang, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2024. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, pages 202–227. PMLR.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhu, Shaolin, Menglong Cui, and Deyi Xiong. 2024a. Towards robust in-context learning for machine translation with large language models. In Calzolari, Nicoletta, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629, Torino, Italia, May. ELRA and ICCL.

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

A Appendix

This shows an example of a section of the style guide that was automatically converted to an APE instruction and then manually validated.

1. **Style guide section: Inclusive writing**

    - Exercise caution when referring to cities, countries or regions that are politically contested. When giving examples involving multiple regions, use comparable references. For example, avoid mixing countries with states or continents.
    - Do not generalise or stereotype people based on their region, culture, age or gender, not even with positive stereotypes.
    - Do not use rude, derogatory or racially biased terms.
    - Accessibility: Avoid using words that imply pity or illness. Instead of 'victim of', 'suffering from', 'afflicted by' or 'crippled by', use 'person who has' or 'person with'. Instead of 'the disabled', 'the handicapped', 'the blind' or 'the deaf', use 'disabled people', 'blind people', 'deaf people' or 'people with hearing loss'.
    - Gender-neutral language: Avoid gendered terms (postal workers instead of postmen, firefighter instead of fireman). Use 'they' when a person's gender is unknown or irrelevant (Ask your concierge if you need help. They are there to support you.)

2. **LLM conversion:**

    - Use inclusive, gender-neutral language and avoid stereotypes.

3. **Manually validated instructions:**

    - Avoid stereotypes, generalizations, and offensive or pitying terminology.
    - Use gender-neutral language (postal workers not postmen, they for unknown gender).
    - Use person-first language (person with... not victim of...).