Document Summarization for AI-based Post-Editing

Vera Senderowicz Guerra vera.senderowicz@welocalize.com
Dimitrios Pavlou dimitrios.pavlou@welocalize.com
Peter Bourgonje peter.bourgonje@welocalize.com
Olesia Khrapunova olesia.khrapunova@welocalize.com
Konstantinos Karageorgos konstantinos.karageo@welocalize.com
Aaron Schliem aaron.schliem@welocalize.com
Welocalize

Abstract

Post-Editing (PE) is typically performed on isolated segments or small batches, without access to broader document context. In this paper, we investigate whether pre-generated, document-level summaries can improve PE quality. Using a purpose-built summarization prompt evaluated across nine LLMs from OpenAI and Google, we select two models with contrasting summary styles for downstream experiments on 448 documents covering 37 target locales and 13 content domains. Summaries generated by `gemini-2.5-flash-lite`, which are directive and domain-specific, bring output closer to the human reference and yields modest gains in COMET, whereas those generated by `GPT-4o`, which tend to be more generic and descriptive, degrade performance across most metrics. The positive effect appears most pronounced in terminologically dense domains and lower-resource locales. A qualitative analysis shows that improvements arise when summaries provide specific, actionable guidance on terminology, domain conventions, and style, and that performance decreases when summaries are underspecified or conflicting. These findings suggest that summary *specificity and actionability*, rather than the mere addition of context, determine whether document-level information benefits post-editing.

1 Introduction

A growing body of work demonstrates that AI Post-Editing (AIPE)[1] enhances Machine Translation (MT) output and Translation Memory (TM) fuzzy matches (Koehn and Senellart, 2010; Raunak et al., 2023; Uguet et al., 2024). This effectiveness is further amplified when prompted with language-specific conventions and critically selected historical human-approved translations (Moslem et al., 2023; Mu et al., 2023; Chitale et al., 2024). Building on this evidence, we hypothesize that AIPE further benefits from document-relevant context, and we experiment with providing the PE model with a pre-generated, high-level summary of the entire source document, including its content type and domain.

Our contributions are as follows:

- We propose augmenting AIPE prompts with automatically generated document-level summaries, giving the model access to global context when making segment-level post-editing decisions.
- We describe a summary generation pipeline in which we first generate summaries across multiple LLMs, use human evaluation to identify the target quality and tone, recover a reusable prompt via cross-model reverse prompt engineering, and then shortlist two models with contrasting summary styles for downstream evaluation.

[1]Throughout this paper, AIPE refers to the automatic editing of MT output or TM matches by an LLM, as distinct from MT post-editing (MTPE), in which a human editor revises MT output.

- We evaluate our approach on 448 documents covering 37 target locales and 13 domains. We find that summaries improve PE when they provide specific, actionable guidance on terminology, domain conventions, and style, but degrade it when underspecified or conflicting. The effect is further modulated by content domain and target language.

2 Related Work

Translating segments in isolation leads to discourse-level errors such as inconsistent terminology and pronoun misalignment (Voita et al., 2019). While document-level MT approaches address this by providing broader context, Herold and Ney (2023) show that volume alone does not help and can introduce noise, motivating *distilled* over raw context. Mohammed and Niculae (2025) further demonstrate that Large Language Models (LLMs) are sensitive to whether context is genuinely informative, and uninformative context can harm performance. Together, these findings argue in favor of providing a structured summary rather than adjacent segments.

With respect to PE, Moslem et al. (2023) show that in-context signals such as style guides and fuzzy matches enable LLMs to outperform dedicated MT engines on domain-specific content without fine-tuning. Our AIPE baseline is based on this LLM-based paradigm. Ki and Carpuat (2024) establish that signal *specificity* is decisive: Fine-grained feedback improves PE quality, while coarse signals do not. This aligns with our findings that more directive summaries improve but more generic summaries decrease performance.

3 Method

Our set-up consists of the two main components described below: the generation of document-specific summaries (Section 3.1), and subsequently using these tailored summaries during PE (Section 3.2).

3.1 Summary Generation

We used two non-reasoning (`GPT-4o`, `GPT-4.1`) and two reasoning (`o1`, `o3`) models through the OpenAI GUI[2] to iteratively generate summaries for two English source documents. From our pool of documents (see Section 4), we randomly selected one from the Marketing domain and another consisting of Technical Documentation, to cover diverse document genres. For each document, all source segments were concatenated into a single input text and passed to the summary generation model. We shared the output summaries, along with the original source documents, with a team of professional linguists, and asked them to rank the summaries by their perceived usefulness in the context of a translation/post-editing task. Overall agreement favoured the summaries generated by the `o1` model. We then reverse-prompted `GPT-4o` with the preferred summary, instead of `o1`, to avoid model-specific bias in the recovered prompt, and obtain a generation instruction that is not overfit to a single model. Li and Klabjan (2025) demonstrate that reverse prompt engineering yields semantically faithful results even across different generator–inverter model pairs, supporting the feasibility of this cross-model approach. The final prompt to create the summaries is included in Appendix B.

This prompt was then used to automatically produce summaries for a wider set of documents representing all content types present in our evaluation data (see Section 4). At this stage, we experimented with several OpenAI and Google models[3], spanning a range of cost, latency, and reasoning profiles. To determine their suitability for use in AIPE prompting, we had the resulting summaries qualitatively evaluated by a panel of linguists and engineers, who ranked them according to post-editing utility and usability in prompting contexts, respectively. Appendix A describes the panel's assessment and verdict reasons in more detail, but the final choice was to keep one model per provider: `GPT-4o` and `gemini-2.5-flash-lite`.

3.2 Post-Editing

Our AIPE system operates on both MT output and TM fuzzy matches, leveraging publicly available LLMs prompted with the target-locale-specific style guide and in-context source–target pairs retrieved from previously approved projects or translation memories. In our experiments, we augment this baseline by prepending to the prompt a document-level summary (generated as described in Section 3.1) for the document from which each seg-

[2] https://openai.com/, accessed in July of 2025.

[3] https://cloud.google.com/vertex-ai/generative-ai/docs/model-garden/explore-models, accessed in July of 2025.

ment originates. The summary thus provides the model with global context that is otherwise absent in segment-level PE. The same post-editing model is used across all three configurations, so that the only variable under comparison is the presence and style of the document-level summary. We do not disclose the specific model for the reasons noted in the Limitations section. Since the output of AIPE is either an unchanged or a modified target segment, evaluation can be done using standard translation quality metrics (see Section 5).

An alternative to segment-level PE is to post-edit an entire document in a single pass, which recent long-context models make technically feasible. We operate at the segment level with a pre-computed summary because this matches our current production translation workflows, where segments are processed independently, arrive incrementally, and must remain individually traceable for review and TM storage. A pre-computed summary is also amortized across all segments of a document and all its target locales, whereas document-level PE would repeat the full source context for every target language. A direct comparison between the two approaches is left to future work.

4 Data

We evaluate on 448 proprietary documents drawn from 103 client accounts across real-world projects. The documents contain a total of 331,151 segments, comprising approximately 3.98 million source words.

Languages. The source language is English (en-US and en-GB) for all documents. Our dataset covers 37 target locales, including languages such as Arabic, Chinese, French, German, Greek, Japanese, Korean, Polish, Portuguese, Spanish, Thai, Turkish, Ukrainian, and Vietnamese. Locale variants of the same language (e.g. pt-PT and pt-BR) are treated as separate conditions throughout.

Content Types. The dataset spans 13 translation domains or content specialties, with the most frequent being Product Service, Marketing and Chemistry.

Segment Composition. Each document contains source segments, pre-translations, and human-approved reference translations. Pre-translations originate from machine translation (40.3%; engines include Google NMT, Systran, Microsoft Translator, and DeepL), leveraged TM matches (22.1%), fuzzy matches (16.3%), repetitions (12.3%), and in-context exact matches (5.7%). This mix reflects realistic production conditions where AIPE operates on heterogeneous input quality.

5 Evaluation

We evaluate the impact of document summaries on AIPE performance using three complementary quality metrics:

- *Edit Distance Relative Improvement* (based on Levenshtein (1966)): the relative reduction in normalized Levenshtein distance to the human reference, compared to the pre-translation baseline:

$$\Delta_{\text{ed}} = \frac{d(MT, \text{ref}) - d(PE, \text{ref})}{d(MT, \text{ref})} \times 100$$

  where d is the normalized character-level indel distance (Levenshtein, 1966), computed with unit cost for insertions and deletions; substitutions are treated as a deletion followed by an insertion (cost 2);

- *COMET* (Rei et al., 2020);

- *BLEU* (Papineni et al., 2002).

All metrics are computed against the human-approved reference translations. For every metric, we provide the relative change compared to the pre-translation. We report results for three configurations: the baseline AIPE system with no summaries (as described in Section 3.2), AIPE with `gemini-2.5-flash-lite`-generated summaries, and AIPE with `GPT-4o`-generated summaries.

Additionally, we report the behavioural metric *Edit Effort Change*, which measures how much a post-editing system rewrites the pre-translation relative to the original quality gap:

$$\Delta_{\text{effort}} = d(MT, PE) - d(MT, \text{ref})$$

where d is the normalized character-level distance defined above. A negative value indicates the system edited less than the MT-to-reference gap; a positive value indicates it edited more. This metric captures editing volume, not whether the edits translate into a quality improvement; read alongside the

quality metrics, it allows us to measure editing efficiency: fewer edits for the same quality gain indicates a more efficient and targeted post-editing.

While these quantitative metrics provide a useful aggregate picture, they do not fully capture the quality of individual post-editing decisions. As Freitag et al. (2022) demonstrate, surface-level metrics like BLEU can diverge considerably from human quality judgments, and Bentivogli et al. (2018) show that PE-based and direct human assessment can yield complementary (and sometimes conflicting) views of translation quality. Moreover, Sarti et al. (2025) find that domain, language, and editor characteristics strongly influence PE outcomes, underscoring that aggregate scores may mask important variation across conditions. We therefore complement them with a qualitative human review of a collection of samples (Section 5.4).

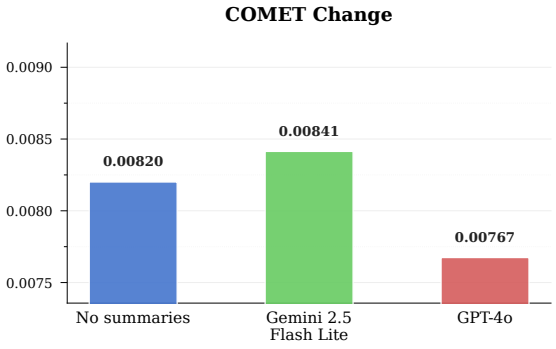

Figure 2: Mean COMET Change score per configuration. Higher is better. Gemini summaries achieve the highest score (0.00841), followed by the baseline (0.00820), with GPT summaries scoring lowest (0.00767).

5.1 Overall Results

Figures 1 to 3 report the four metrics mentioned above across the three configurations.

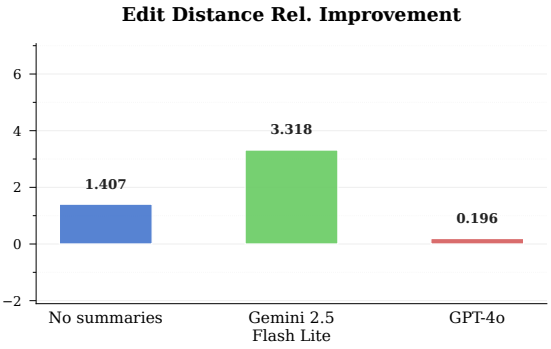

Figure 1: Edit distance relative improvement over the pre-translation. Higher is better. Gemini summaries yield a relative improvement of 3.32, compared to 1.41 for the baseline and 0.20 for GPT summaries.

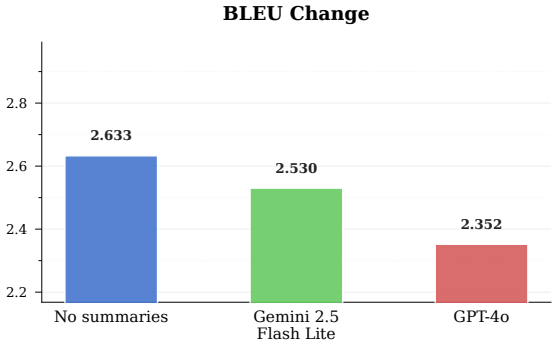

Figure 3: Mean BLEU Change score per configuration. Higher is better. BLEU deteriorates under both summary conditions (Gemini: 2.530; GPT: 2.352) relative to the baseline (2.633), though Gemini's drop is smaller.

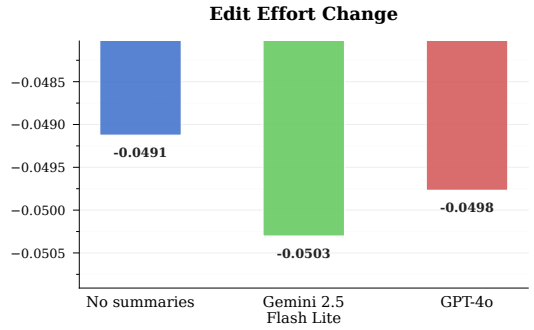

Figure 4: Edit effort change per configuration. All three AIPE systems edit the pre-translation by a similar amount: differences are within 0.0012 (Baseline: -0.0491; Gemini: -0.0503; GPT: -0.0498). This indicates that summaries do not significantly alter how many edits the system produces, only how effectively those edits improve quality.

The results when using summaries produced by `gemini-2.5-flash-lite` show the best performance overall. By contrast, `GPT-4o` summaries degrade performance on all quality metrics. Edit distance relative improvement (Figure 1) shows that Gemini-augmented AIPE moves the output closer to the human reference, while GPT-4o summaries do the opposite. Edit effort change (Figure 4) is essentially flat across the three configurations, with all values falling within 0.0012 of one another; this indicates that summaries do not alter how much the system rewrites the pre-translation, only how effectively those edits improve quality. In other words, the Gemini setup achieves higher quality with the same volume of editing, suggesting more targeted edits rather than more extensive rewriting. Conversely, GPT summaries appear to misdirect the same volume of editing, degrading quality without reducing effort. COMET change (Figure 2) reaches 0.00841 with the Gemini setup, marginally above the baseline (0.00820), while GPT drops to 0.00767. BLEU change declines modestly from 2.633 to 2.530 with Gemini, a pattern consistent with the well-known tendency of neural MT/PE to introduce surface variation that disrupts n-gram overlap without necessarily reducing translation quality (Freitag et al., 2022). The larger BLEU decline to 2.352 with GPT, however, aligns with its degradation across all other quality metrics.

5.2 Impact by Content Type

A breakdown by content specialty reveals that the impact of document summaries is not uniform across domains. Figures 5, 6, and 7 show BLEU change, COMET change, and edit distance relative improvement, respectively, across the 13 content specialties in our dataset.

Summaries seem to be most beneficial in two scenarios. First, in terminologically dense or highly regulated domains such as Biology and Legal, where domain-specific conventions, regulatory constraints, and precise terminology are critical, both summary configurations—and `gemini-2.5-flash-lite` in particular—yield directionally positive results, though not consistently across metrics. Second, in the catch-all "Other" category, where documents are difficult to classify by domain, summaries appear to compensate for the lack of strong domain priors, with Gemini showing a marked gain in edit distance.

A further pattern emerges in domains where the baseline AIPE configuration degrades translation quality (i.e., moves output further from the human reference). In categories such as Marketing and Medical Devices, where baseline edit distance is negative, summaries tend to mitigate this degradation, suggesting that document-level context helps the model avoid harmful edits when the task is otherwise under-constrained. For general or mixed content types, results are less consistent, with gains and losses varying across metrics and configurations.

5.3 Impact by Target Locale

Breaking the same metrics down by target locale (Figures 8, 9, and 10) tells a story that is broadly consistent with the per-content analysis. The effect of document summaries is small, but directionally positive in most languages. For high resource European pairs (`de-DE`, `fr-FR`, `es-ES`, `nl-NL`), all three configurations sit close together, suggesting that AIPE already has a strong direction from the style guide and translation pairs alone and that the document summary contributes little additional information. The largest positive deltas under the Gemini configuration appear in languages where the baseline itself is weaker (e.g. `th-TH`, `id-ID`, `pt-PT`, and `zh-TW`), suggesting that summaries might also help the most with languages that are more challenging for MT. A small number of target locales

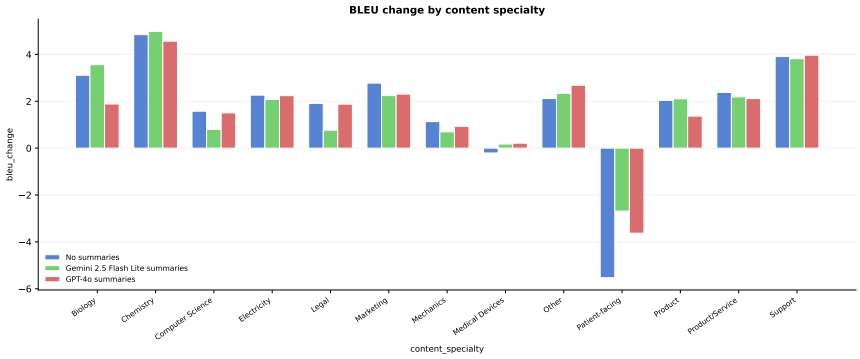

Figure 5: BLEU change by content specialty for the three configurations.

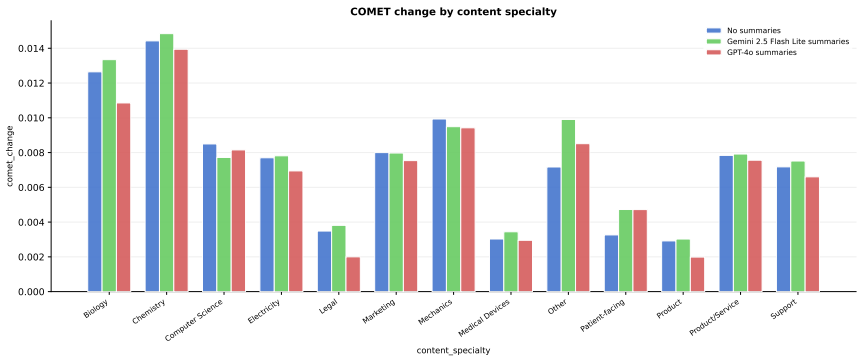

Figure 6: COMET change by content specialty for the three configurations.

(`it-IT`, `pt-BR`, `ja-JP`, `tr-TR`) show a marginal regression for both summary configurations relative to the no-summary baseline.

Although individual differences across content types and target locales are not statistically significant—unsurprisingly, given the small per-cell sample sizes that result from slicing 448 documents across 37 target locales and 13 domains—the pattern is not homogeneous but directionally consistent: `gemini-2.5-flash-lite` summaries improve performance or prevent quality degradation across the majority of domains and some of the most challenging languages for MT, while `GPT-4o` summaries mostly degrade it. We take this consistency as supportive evidence that summary specificity, rather than summary presence alone, drives the effect.

5.4 Human Review

We conducted a qualitative analysis of a sample of cases with notable differences in edit distance between the `gemini-2.5-flash-lite`-augmented and baseline configurations, focusing on Gemini as the better-performing summary model overall, according to the automatic metrics. We examined segments where summaries led to improvement, deterioration, or no change, and also reviewed the model's outputted rationale to identify which summary elements drove these effects and inspect how the model interacts with the information the summaries provide.

No Impact (Majority of Cases). In most cases, summaries did not alter the PE outcome: the differences reported by the automatic metrics can have zero impact on quality according to human reviewers. When the pre-translated segment already met expectations in terms of tone, structure, and terminology, the model did not introduce any unnecessary edits. We consider this a reassuring finding: as suggested by the edit effort change results, summaries by themselves do not encourage over-editing.

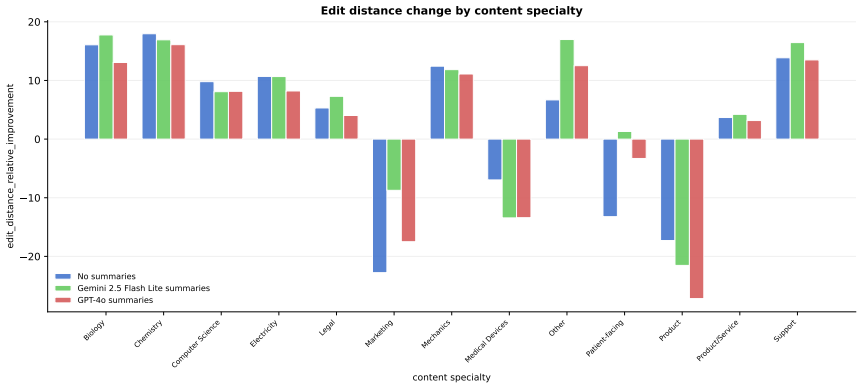

Figure 7: Edit distance relative improvement by content specialty for the three configurations.

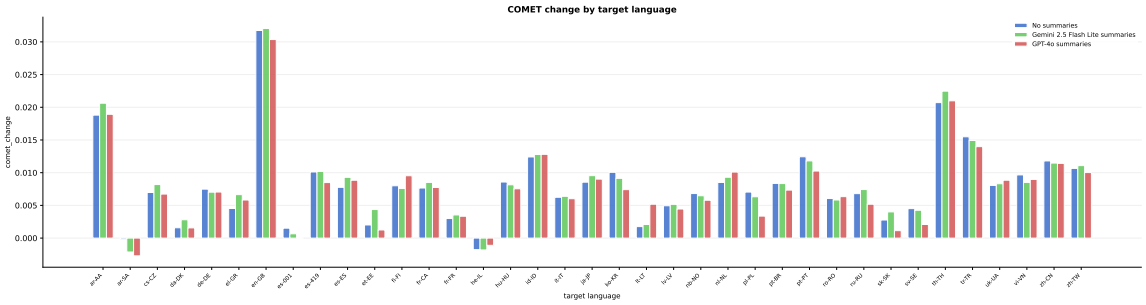

Figure 8: COMET change by target locale for the three configurations.

Improvements. Summaries drove improvements primarily in the following three areas:

1. **Attention to technical and domain-specific elements**: Summaries that flagged proper nouns, placeholders, and capitalization conventions, as well as domain standards, helped the model focus on these aspects during editing.
2. **Detection of incomplete content**: When the summary emphasized thoroughness or completeness, the model identified and corrected translations with missing or overlooked content more reliably than in the baseline configuration.
3. **Rationale quality**: The model's stated rationale for its edits was more precise and context-specific when informed by a summary, suggesting that document-level context helps the model make more targeted editing decisions.

All three mechanisms share a common property: the summary supplies concrete, checkable constraints rather than general description. This is consistent with our overall finding that actionability (a characteristic of the summaries generated by `gemini-2.5-flash-lite`), rather than the mere presence of document context, is determinant of a summary's utility.

Deterioration. In a minority of cases, summaries led to a performance decrease. This typically occurred in three scenarios: when the summary flagged certain terms (e.g., named entities or Do-Not-Translate items) without specifying how to handle them, causing the model to make incorrect decisions; when summary guidance contradicted the style guide in high-stakes domains; or when short or ambiguous segments were disproportionately influenced by summary content, resulting in unnecessary edits. We consider this information to be extremely helpful to inform new experiment iterations.

Taken together, document summaries helped most when they provided specific, actionable direction on terminology, domain conventions, and style,

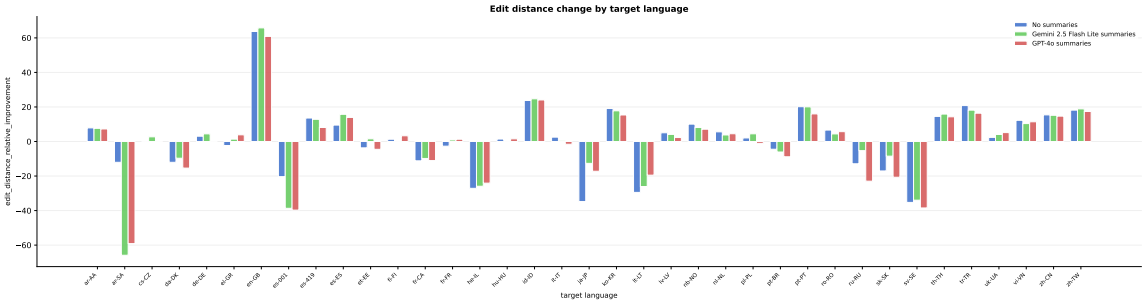

Figure 9: Edit distance relative change by target locale for the three configurations.

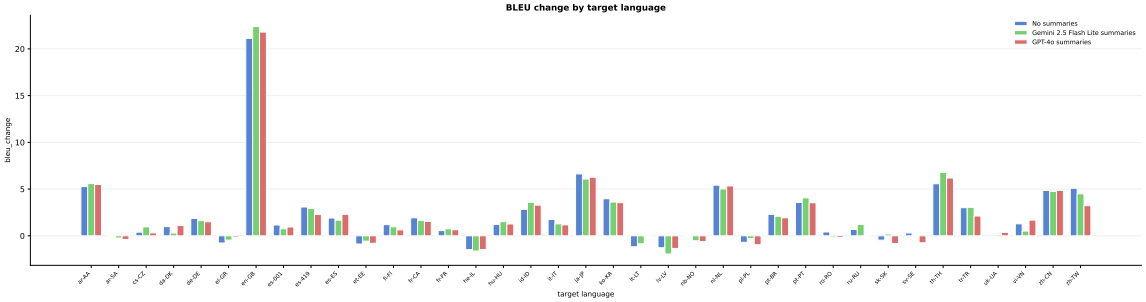

Figure 10: BLEU change by target locale for the three configurations.

and caused harm primarily when they were underspecified, conflicting, or generic. When the pretranslation already met quality expectations and the summary added no document-specific direction, it had no effect.

5.5 Cost Analysis

Figure 11 shows the mean PE cost in token usage per configuration in USD. The configurations using summaries incur a higher cost due to the additional (summary) tokens in the prompt, as well as to the summary generation itself. As illustrated, the increase for `gemini-2.5-flash-lite` is higher than that for `GPT-4o`. Despite the higher token cost, `gemini-2.5-flash-lite` outperforms the baseline on edit distance relative improvement and COMET, and outperforms `GPT-4o` on both of these as well as BLEU. We consider this a favorable cost–quality trade-off.

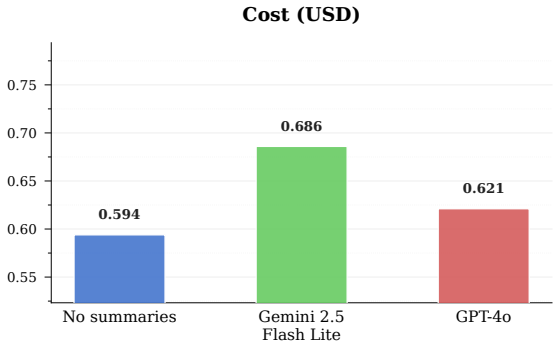

Figure 11: Mean PE cost in USD per configuration across all 448 documents.

6 Conclusion & Future Work

We investigate whether augmenting AIPE with document-level summaries to have access to global context improves segment-level post-editing quality. After a multi-stage pipeline to design a summary generation prompt and shortlist two models, we compare `gemini-2.5-flash-lite` and `GPT-4o` summaries against a no-summary baseline across 448 documents, 37 target locales, and 13 content do-

mains.

Our results show that summary effectiveness depends on specificity and actionability: under the same prompt conditions, `gemini-2.5-flash-lite` produces directive, domain-grounded summaries that yield improvements in edit distance and COMET, while `GPT-4o`'s summaries tend to be more generic and descriptive, causing a degradation of performance across most metrics. A qualitative analysis confirms that summaries help most when they provide concrete guidance on terminology, domain conventions, and style, and cause harm when they are underspecified or conflicting. The benefit appears most pronounced in terminologically dense domains and lower-resource locales, and in cases where the baseline AIPE itself degrades quality, summaries can help mitigate harmful edits.

Several directions for future work emerge. First, rather than providing a summary for every document, a natural extension is to predict whether a summary is likely to help for a given document—potentially based on content type, language pair, or baseline confidence; or even on historical data—and only inject one when beneficial, reducing unnecessary token cost. Moreover, systematically comparing summaries across a wider range of model families (reasoning-oriented, non-reasoning, and open-weight) would help disentangle the contribution of model capability from summary content.

Relatedly, we hold the post-editing model fixed throughout; how the choice of post-editor interacts with the choice of summarizer model—for instance, whether stronger post-editors benefit less from directive summaries—remains an open question. It also remains open whether a model's general summarization ability predicts its utility as an AIPE summarizer, which would offer a cheaper model-selection procedure than the downstream evaluation used here. Finally, comparing summary-augmented segment-level PE against document-level PE in a single long-context pass would help establish whether distilled context is preferable to full context when both are available.

Limitations

Our experiments make use of both a proprietary PE platform and proprietary customer data which we cannot disclose, limiting reproducibility. Additionally, we only use English as a source language; results may differ for other source languages. Our edit-based metrics are computed against a single human reference translation, which assumes one preferred rendering per segment. For longer or more nuanced source sentences, several equally acceptable translations may exist, so a reduction in edit distance does not necessarily indicate an improvement in quality; this motivates our inclusion of COMET alongside the edit-based metrics, which nonetheless doesn't have an equal performance across all languages. Relatedly, our human review examines post-editing decisions and model rationales rather than providing a full human quality assessment of the final post-edited output. Finally, although directional patterns are consistent, individual differences across content types and target locales are not statistically significant, owing to the small per-cell sample sizes that result from slicing the data across 37 target locales and 13 domains.

References

Bentivogli, L., Cettolo, M., Federico, M., and Federmann, C. (2018). Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment. In *Proceedings of the 15th International Conference on Spoken Language Translation (IWSLT)*, pages 62–69.

Chitale, P., Gala, J., and Dabre, R. (2024). An empirical study of in-context learning in LLMs for machine translation. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7384–7406, Bangkok, Thailand. Association for Computational Linguistics.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., and Martins, A. F. T. (2022). Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages

46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Herold, C. and Ney, H. (2023). Improving long context document-level machine translation. In Strube, M., Braud, C., Hardmeier, C., Li, J. J., Loaiciga, S., and Zeldes, A., editors, *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 112–125, Toronto, Canada. Association for Computational Linguistics.

Ki, D. and Carpuat, M. (2024). Guiding large language models to post-edit machine translation with error annotations.

Koehn, P. and Senellart, J. (2010). Convergence of translation memory and statistical machine translation. In Zhechev, V., editor, *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 21–32, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Li, H. and Klabjan, D. (2025). Reverse prompt engineering: A zero-shot, genetic algorithm approach to language model inversion. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26223–26245, Suzhou, China. Association for Computational Linguistics.

Mohammed, W. and Niculae, V. (2025). Context-aware or context-insensitive? assessing LLMs' performance in document-level translation. In Bouillon, P., Gerlach, J., Girletti, S., Volkart, L., Rubino, R., Sennrich, R., Farinha, A. C., Gaido, M., Daems, J., Kenny, D., Moniz, H., and Szoc, S., editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 126–137, Geneva, Switzerland. European Association for Machine Translation.

Moslem, Y., Haque, R., Kelleher, J. D., and Way, A. (2023). Adaptive machine translation with large language models. In Nurminen, M., Brenner, J., Koponen, M., Latomaa, S., Mikhailov, M., Schierl, F., Ranasinghe, T., Vanmassenhove, E., Vidal, S. A., Aranberri, N., Nunziatini, M., Escartín, C. P., Forcada, M., Popovic, M., Scarton, C., and Moniz, H., editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Mu, Y., Reheman, A., Cao, Z., Fan, Y., Li, B., Li, Y., Xiao, T., Zhang, C., and Zhu, J. (2023). Augmenting large language model translators via translation memories. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10287–10299, Toronto, Canada. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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. Association for Computational Linguistics.

Raunak, V., Sharaf, A., Wang, Y., Awadalla, H., and Menezes, A. (2023). Leveraging GPT-4 for automatic translation post-editing. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (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. Association for Computational Linguistics.

Sarti, G., Zouhar, V., Chrupała, G., Guerberof-Arenas, A., Nissim, M., and Bisazza, A. (2025). QE4PE: Word-level quality estimation for human post-editing. *Transactions of the Association for Computational Linguistics*, 13:1410–1435.

Uguet, C., Bane, F., Aymo, M., Torres, J., Zaretskaya, A., and Blanch Miró, T. B. M. (2024). LLMs in post-translation workflows: Comparing performance in post-editing and error analysis. In Scarton, C., Prescott, C., Bayliss, C., Oakley, C., Wright, J., Wrigley, S., Song, X., Gow-Smith, E., Bawden, R., Sánchez-Cartagena, V. M., Cadwell, P., Lapshinova-Koltunski, E., Cabarrão, V., Chatzitheodorou, K., Nurminen, M., Kanojia, D., and Moniz, H., editors, *Proceedings of*

the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pages 373–386, Sheffield, UK. European Association for Machine Translation (EAMT).

Voita, E., Serdyuk, D., Lample, G., Titov, I., and Sennrich, R. (2019). Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1264–1274. Association for Computational Linguistics.

A Summary Generation Procedure

In Section 3.1 we discussed how we used Google and OpenAI models and a qualitative analysis to arrive at a prompt for summary generation. `o1` and `o3`'s summaries came close to the type of summary that we would consider helpful for post-editing. They were well structured, with an appropriate level of detail and specificity; however, they were overly verbose. `gemini-2.5-flash-lite` performed the best overall among the Google models. Its summaries were concise yet exhaustive, and their assertive, instructional framing matched the target tone well. By contrast, `GPT-4o` and `GPT-4o-mini` produced shorter, rather generic summaries that read more as neutral descriptions than as directive guidance, which made us doubt their capacity to provide AIPE with the specific instructions and the actionability it would need to improve its performance. Most summaries produced by `gemini-1.5-pro` also lacked certain specific instructional elements, without offering a compensating reduction in token cost. `gemini-2.0-flash` summaries were phrased tentatively and contained rather hedging language that might not be ideal to guide AIPE. `gemini-2.5-flash` and `gemini-2.5-pro` tended to produce exhaustive summaries that risked being too long and repetitive, and also deferred ultimate decisions to external resources (the client's decision, glossaries, TMs, etc), inaccessible by the system at editing time. Based on this analysis, we proceeded with only the summaries generated by `GPT-4o` and `gemini-2.5-flash-lite` for our PE experiments (lower-cost models, one per provider), the former as a representative of the more generic, descriptive style, and the latter as our best match for the assertive, directive tone.

B Summary Generation Prompt

The following prompt was used to generate document summaries for all models in our experiments.

> You are a Translation Quality Specialist who creates detailed summaries and guidelines to support professional translators and post-editors. You specialize in analyzing documents of all types and identifying what translators need to know to produce accurate, culturally appropriate, and well-formatted translations. You are highly attentive to tone, terminology, structure, formatting, and cross-cultural or compliance-related concerns.
>
> You receive a source document that will be translated into a target language. Provide a summary of this document specifically for use by a translation or post-editing team. Your summary should address the following areas:
>
> **Document Type & Format**
>
> - Describe what kind of document it is (e.g., handbook, contract, technical manual, brochure).
> - Summarize the document's structure (e.g., sections, headings, tables, bullet lists, embedded forms).
> - Indicate the intended audience (internal staff, external partners, general public, specialists, etc.).
>
> **Domain / Subject Area**
>
> - Identify the main field, industry, or topic (e.g., legal, HR, marketing, software, healthcare).
> - Note any legal, regulatory, or compliance-related context that could influence translation.
>
> **Content Style & Tone**
>
> - Characterize the writing style (e.g., formal, instructional, promotional, legalistic).
> - Indicate the tone (e.g., neutral, persuasive, authoritative) and whether it must be preserved.
>
> **Translation-Relevant Considerations**

- Highlight key terminology, repeated jargon, acronyms, or proprietary labels requiring consistent rendering.
- Identify culturally specific or locale-bound elements (e.g., legal systems, currency, measurement units, references to laws/agencies, contact details).
- Note formatting or structural components that should be preserved (e.g., bullet lists, numbering, forms, tables, signature fields).
- Indicate if the document includes non-text elements, embedded graphics, or multilingual content that need to be tagged or processed separately.
- Flag any potential translation challenges, such as highly regulated content, idiomatic expressions, region-specific rules, or text that must remain literal.

Overall Guidance

- Summarize the document's overall purpose (e.g., compliance, employee instruction, customer onboarding).
- Provide advice relevant to the translation brief: tone fidelity, legal accuracy, formatting consistency, localization vs. direct translation choices.
- Note any areas requiring special attention for target language or culture, including terminology integrity and usability.

Your summary should be clear, structured, neutral, and practical, enabling translators or post-editors to accurately and efficiently work with the document. Focus on providing context and direction, not replicating the content.