AI Post-Editing in Production: A 71,262-Segment Evaluation Across Five Domains, Ten Languages and Five Systems

Mara Nunziatini
Welocalize
mara.nunziatini@welocalize.com

Mercedes Speroni
Welocalize
mercedes.speroni@welocalize.com

Abstract

This study evaluates an AI post-editing (AIPE) system in a professional translation setting, covering translation from English into ten target languages across five domains. We evaluate the system using automatic metrics on 71,262 production segments and human evaluation on a stratified sample of 6,618 segments (approximately 600 segments per target language) assessed by 60 professional translators. AIPE refines machine translation output using a secure publicly available LLM, retrieving language-specific style guides and high-quality bilingual examples to guide edits. We compare it with direct LLM translation (LLMT), Google Translate, and DeepL. The two AIPE configurations evaluated consistently outperform the generic translation baselines in terms of quality. LLMT does not match this quality, though it may suit less quality-sensitive domains. We observe how AIPE's gains vary according to pre-translation type, with fuzzy translation memory matches overrepresented among severe errors, and discuss deployment implications.

1 Introduction

The adoption of large language models (LLMs) in professional translation workflows has accelerated rapidly, raising practical questions for language service providers (LSPs) about where and how to deploy them. Two distinct paradigms have emerged: AI post-editing, where an LLM refines the output of an existing MT engine, and LLM-based direct translation, where the LLM translates from source without a pre-translation step. Both approaches promise quality improvements over generic neural MT, but independent, production-scale evidence comparing them is scarce. Most published evaluations focus on research-grade datasets or assess single language pairs and domains (Raunak et al., 2023), but industry deployments involve a far messier reality: heterogeneous content types and language pairs, mixed pre-translation input types (translation memory matches, generic MT, custom MT), and stringent business constraints on cost and turnaround. This paper reports on an evaluation designed to answer three questions: 1. Does AIPE deliver measurable quality improvements over generic MT baselines alone? 2. Is LLM-based direct translation a quality-equivalent and cheaper alternative to MT + AIPE? 3. How does pre-translation quality moderate AIPE's gains? A methodological note is needed. AIPE and direct LLM translation use additional resources, such as high-quality bilingual examples and style guides, while Google Translate and DeepL were not trained with glossaries or translation memories for this experiment. AIPE also has access to a pre-translation, which it refines rather than generating text from scratch. This difference is intentional and reflects real production conditions. As a result, the quality differences we observe come from the full process, including the initial translation, the model, the prompting strategy and the additional resources.

2 Background and Related Work

Post-editing MT output is well-established as a productivity measure in professional translation

(Koponen, 2016). The advent of LLMs has prompted researchers to explore their use both as post-editors and as direct translators. Jiao et al. (2023) show that LLMs can achieve competitive translation quality on standard benchmarks. Moslem et al. (2023) demonstrate that LLMs can effectively adapt MT output when provided with in-context bilingual examples, closely analogous to the Translation Pairs mechanism used in our AIPE solution. Their findings on single language pairs motivated our investigation of whether this approach scales across ten target languages and diverse professional content types. Our study complements this prior work with production-scale evidence grounded in real client workflows with heterogeneous input types. We calculate automatic metrics on the 71,262-segment dataset, and human annotation of a stratified 6,618-segment subset (around 600 per language) by 60 professional translators using rankings and the DQF-MQM framework (Lommel et al., 2014). Lommel et al. (2014b) demonstrate its reliability across annotator profiles, supporting our use of both internal and external specialist linguists.

3 Systems

3.1 AIPE

AIPE is an LLM-based post-editing system that takes existing translations and improves them. The LLM receives the source and pre-translated text together with relevant context retrieved via a retrieval-augmented generation (RAG) strategy. This context includes human-reviewed bilingual segments (referred to as Translation Pairs throughout this paper) and language-specific style rules, which guide the LLM to correct errors, adjust terminology, and align the output with the expected tone and brand voice. AIPE targets MT outputs and fuzzy translation memory (TM) matches, skipping Exact and ICE (Internal Context Exact)[1] matches to preserve pre-approved content. The system respects inline tags and file structure, while also storing detailed metadata and post-editing results for tracking, analysis, and continuous improvement. Typically, this feature is used in conjunction with other technologies under the OPAL Enable product offering (Nunziatini et al., 2025), such as customized MT and AI Quality Estimation, to maximise its efficiency. Two AIPE configurations were evaluated:

- Generic MT + AIPE: using Google Translate output as the pre-translation, and
- Production AIPE: follows a typical real-life production setup in which we leverage TM(s) down to 75% fuzzy matches first, and then apply MT (generic or customized, depending on the use case) on the rest of the content. Then, AIPE is applied to all the segments, skipping exact and ICE matches. In more detail, pre-translation input types are distributed as follows:
  - MT segments (generic or customized): 45%
  - fuzzy TM matches equal or above 75%: 19%
  - ICE segments: 19%
  - exact TM matches: 17%

In both cases, Translation Pairs were retrieved where available.

3.2 LLMT

LLMT uses the same secure publicly available LLM, but translates directly from source, without any MT pre-translation. It receives the same style guides and Translation Pairs as AIPE. This configuration is cheaper and faster to operate since it eliminates the cost of an MT pre-translation step, and was hypothesized to approximate AIPE quality at lower cost.

3.3 Baseline MT Engines

Google Translate and DeepL served as generic MT baselines, translating directly from source with no content-specific resources such as glossaries, translation memories or style guides.

4 Experimental Setup

4.1 Automatic Scoring

The first part of the experiment is an automatic scoring exercise. In this exercise we compare, segment by segment, the translations generated by the different systems mentioned above against the gold standard (reference) human translation. The automatic evaluation dataset comprised 71,262 segments (553,518 words) spanning:

[1] An ICE match is a 100% match where the preceding and the following segments that are in the TM are the same as the previous and next segment in the translation. Since the segment matched as well as the segments before and after that match are identical to the earlier translation, the translation quality has already been verified.

Table 1: Number of segments per target language and domain

Target Language	Prod/Serv	Product	Marketing	Support	Med. Dev.	Total	%
fr-FR (French)	3,668	749	1,873	1,245	228	7,763	10.89%
es-ES (Spanish)	3,668	749	1,873	1,245	228	7,763	10.89%
de-DE (German)	3,668	749	1,873	1,245	228	7,763	10.89%
zh-CN (Chinese)	3,668	749	1,873	1,245	228	7,763	10.89%
ja-JP (Japanese)	3,668	749	1,873	1,245	228	7,763	10.89%
pt-BR (Portuguese)	3,668	749	1,873	1,245	228	7,763	10.89%
ko-KR (Korean)	3,668	749	1,873	1,245	228	7,763	10.89%
it-IT (Italian)	3,668	749	1,873	1,245	228	7,763	10.89%
ru-RU (Russian)	846	749	1,873	881	228	4,577	6.42%
tr-TR (Turkish)	850	749	1,873	881	228	4,581	6.43%
Total	31,040	7,490	18,730	11,722	2,280	71,262	100%
%	43.56%	10.51%	26.28%	16.45%	3.20%		

- 10 target target languages: en-US into fr-FR, es-ES, de-DE, zh-CN, ja-JP, pt-BR, ko-KR, it-IT, ru-RU, tr-TR
- 5 domains: Product/Service (44%), Product (11%), Marketing (26%), Support (17%), Medical Devices (3%)

Table 1 shows the distribution of segments across target languages and domains for the full automatic evaluation dataset.

4.2 Human Evaluation

A stratified sample of approximately 600 source segments per language (equivalent to 6,618 total source segments) was selected for human evaluation across the same accounts and target languages, with slightly smaller samples for ru-RU and tr-TR, reflecting their lower volume of translation requests across accounts. Five system outputs (Generic MT + AIPE, Production AIPE, LLMT, Google Translate, and DeepL) were assessed per segment by three professional translators per language, for a total of 60 annotators (30 specialist linguists for medical device content and 30 for the rest of the domain). The evaluation was conducted blindly and with system order randomized. Annotators performed two tasks:

- Error annotation using DQF-MQM, covering Accuracy, Fluency, Terminology, Style, Design, Verity and Locale Convention error types, each rated as Neutral, Minor, Major, or Critical.
- Preference ranking into three positions (Rank 1 = best, Rank 3 = worst), with ties permitted. Annotators could assign multiple systems to the same rank position, and no rank position was mandatory, meaning that if all outputs were deemed of average or insufficient quality, they could leave the Rank 1 position empty.

Linguists were provided with glossaries, style guides, and TMs as a reference if available.

5 Results

5.1 Automatic Metrics

We report edit distance (HTER-style, computed against human post-edits), ChrF (Popović, 2015), BLEU, and COMET (Rei et al., 2020).

System	HTER ↓	ChrF ↑	BLEU ↑	COMET ↑
Generic MT+AIPE	0.12	82.90	66.70	0.92
Production AIPE	0.11	82.96	70.02	0.92
Google Translate	0.21	66.57	42.10	0.89
LLMT	0.20	68.96	46.56	0.89
DeepL	0.22	64.96	41.29	0.88

Automatic metrics consistently rank the two AIPE-based configurations above all other systems, with Production AIPE achieving the lowest HTER (0.11) and highest chrF (82.96), BLEU (70.02), and COMET (0.92), followed closely by Generic MT+AIPE, while LLMT offers marginal gains over the Google Translate baseline and DeepL performs comparably to it. To validate whether these automatic metric differences reflect genuine translation quality gaps as perceived by expert translators, we conducted human evaluations.

5.2 System Ranking

Human rankings were consistent across all aggregation methods (total counts, fractional tie-aware scoring, majority vote, unanimous vote). The order from best to worst was: Generic MT + AIPE $>$ Production AIPE $>$ Google Translate $\approx$ LLMT $>$ DeepL.

Table 2: Row-normalized rank distribution by system

System	Rank 1	Rank 2	Rank 3
Generic MT + AIPE	57%	29%	14%
Production AIPE	47%	30%	23%
Google Translate	44%	34%	22%
LLMT	41%	32%	26%
DeepL	38%	33%	29%

As shown in Table 2, Generic MT + AIPE received Rank 1 in 57% of all its annotations, Production AIPE received 47% and DeepL 38%, which makes a 19-percentage-point gap between strongest and weakest. This ranking held across the individual target languages and domains. Minor variation occurred in the relative positions of LLMT and Google Translate depending on aggregation method, but neither system consistently broke from the middle tier.

5.3 Error Annotation

Error counts mirrored rankings. As we can see in Table 3, Generic MT + AIPE accounted for 15.8% of all annotated errors, Production AIPE for 19.1%, and DeepL for 23.6%, the highest of any system.

Table 3: Error distribution by systems (all errors)

System	% of Total Errors
DeepL	23.6%
LLMT	21.1%
Google Translate	20.4%
Production AIPE	19.1%
Generic MT + AIPE	15.8%

An important nuance concerns error severity. Production AIPE and Generic MT + AIPE share the same AIPE component; the only difference is the pre-translation it processes. Despite this, Production AIPE shows a higher share of major and critical errors (see Table 4). A possible explanation is the variability in pre-translation quality. In the Production AIPE setup, the inputs include fuzzy TM matches, which account for 19% of all segments but 33% of severe errors. This may suggest that when the input is less reliable or more heterogeneous, AIPE is more likely to preserve or amplify those issues.

Table 4: Error distribution by system (major and critical)

System	% Major & Critical
DeepL	25.0%
Production AIPE	21.7%
Google Translate	20.1%
LLMT	19.7%
Generic MT + AIPE	13.7%

Regarding the distribution of error types (rows normalized by system), Style, Terminology, and Accuracy errors are the most frequent categories across all systems (see Table 5). Within its own error distribution, Google Translate shows a higher share of Terminology errors (32%), as expected from a generic model, while, within its own distribution, Production AIPE shows a higher share of Accuracy errors (29%). This pattern may also be related to the characteristics of its pre-translation inputs (which includes fuzzy TM matches), although this remains a tentative interpretation that needs further investigation.

Table 5: Row-normalized error types by system (%)

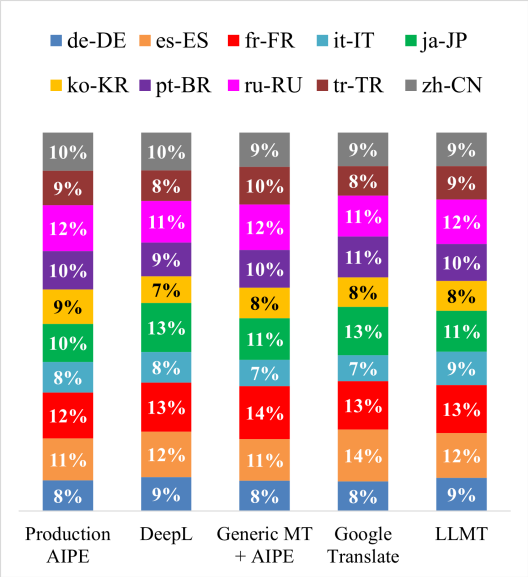
Error Type	Prod. AIPE	DeepL	Gen. MT + AIPE	Google Trans.	LLMT
Style	27.1	28.9	30.9	29.3	30.6
Terminology	22.2	29.0	24.3	32.3	19.9
Accuracy	29.1	24.3	24.4	22.9	25.1
Fluency	16.7	14.8	14.7	13.8	18.7
Design	3.7	1.9	4.5	0.8	4.6
Other	0.8	0.6	0.8	0.6	0.6
Locale conv.	0.4	0.4	0.3	0.2	0.3
Verity	0.1	0.1	0.1	0.1	0.1

5.4 Target Locale-Level Variation

The overall system ranking (Generic MT + AIPE lowest error share, followed by Production AIPE, with DeepL highest) holds across all ten target target languages, though the magnitude of performance differences varies by language. Total error

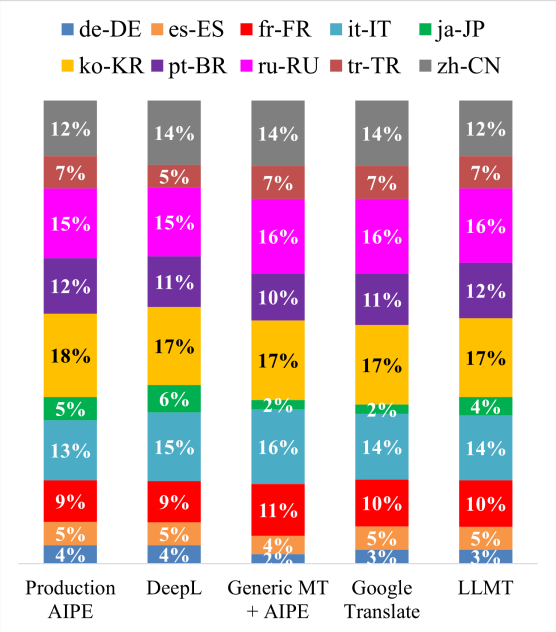
volume and major/critical error concentration do not always move together across target languages, which has practical implications for how deployment risk is assessed. es-ES, fr-FR, and ru-RU show relatively high total error shares (11–14%), while de-DE, it-IT, ko-KR, and tr-TR show lower total shares (7–10%) (see Table 6).

Table 6: Column-normalized error share per system by language



The severity picture (Table 7) differs: ko-KR, despite low total errors, shows a relatively high concentration of major and critical errors, while es-ES shows the reverse (high total errors but a lower share of severe ones). These patterns suggest that locale-level deployment decisions benefit from severity-weighted evaluation rather than total error counts alone. Results for ru-RU and tr-TR should be interpreted with additional caution given their smaller representation in the human evaluation dataset.

Table 7: Column-normalized major + critical error share per system by language



5.5 Content-Level Variation

Content type modulates AIPE's gains, though the dataset distribution means that findings for smaller specialties should be treated as directional rather than definitive. Medical Devices presents a distinct profile: across all systems, this specialty exhibits a higher share of major and critical errors relative to total errors than any other content type, consistent with the precision demands of regulated content (see Table 8).

Table 8: Mean major + critical errors per segment by system and content specialty

Content	Gen. MT + AIPE	Prod. AIPE	Google Trans.	LLMT	DeepL
Marketing	0.06	0.09	0.09	0.11	0.12
Med. Devices	0.22	0.33	0.33	0.30	0.35
Product	0.08	0.14	0.11	0.13	0.15
Prod./Service	0.09	0.15	0.15	0.14	0.20
Support	0.07	0.13	0.14	0.09	0.14

6 Conclusions

This paper presents production-scale evidence that AI post-editing (AIPE) consistently outperforms generic MT baselines and LLM direct translation (LLMT) across ten target languages and five domains. The quality gap between the two AIPE configurations further reveals that pre-translation quality is a key deployment variable: using generic MT alone as input produces fewer and less severe errors than combining MT with fuzzy TM matches,

suggesting that a single, consistent pre-translation source is preferable to mixed input types of varying quality. LLMT may suit domains where top-tier quality is not required, but it does not offer a quality-equivalent alternative. These findings provide concrete guidance for AIPE deployment decisions in professional localization workflows.

6.1 Deployment Implications

The findings have several concrete implications for LSPs considering similar deployments:

1. AIPE on generic MT is a strong default configuration. The marginal cost of a Google Translate pre-translation is low, and the quality gain over standalone LLM or MT translation is consistent and meaningful across target languages and domains.

2. LLMT has a role in cost-tiered workflows. It delivers quality comparable to standalone Google Translate while costing roughly 60% less (approximately $0.40 vs. $1.00 per language per 600 segments, based on actual production spend[2]). For content types where top-tier quality is not critical, LLMT represents a viable and more economical option that also benefits from style guides and Translation Pairs, which generic MT does not receive.

3. Fuzzy TM inputs require careful handling. They are over-represented among severe errors and may benefit from pre-filtering or a tailored prompting strategy before being passed to AIPE.

7 Limitations and Next Steps

The dataset, whilst large, was not perfectly balanced: Product/Service dominated the volume (44%). The inability to systematically distinguish custom from generic MT inputs in the AIPE condition limited the precision of comparisons across pre-translation types. Furthermore, both AIPE and LLMT are powered by a single LLM; other LLMs or generic MT engines may behave differently. Results reflect a specific LSP context and may not generalise directly to other deployment architectures or domains. Next steps include collecting a more balanced dataset, conducting match-type-controlled experiments to isolate the effect of pre-translation quality on AIPE performance, and evaluating alternative LLMs and MT engines to test the generalisability of these findings.

[2]Pricing as of November 2025

References

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Koponen, Maarit. 2016. Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25:131–148.

Lommel, Arle, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 12:455–463.

Lommel, Arle, Maja Popović, and Aljoscha Burchardt. 2014. Assessing inter-annotator agreement for translation error annotation. In *Proceedings of the Workshop on Automatic and Manual Metrics for Operational Translation Evaluation (MTE)*, pages 24–30, Reykjavik, Iceland.

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

Nunziatini, Mara, Konstantinos Karageorgos, Aaron Schliem, and Mikaela Grace. 2025. OPAL Enable: Revolutionizing localization through advanced AI. In *Proceedings of Machine Translation Summit XX: Volume 2*, pages 83–85, Geneva, Switzerland. European Association for Machine Translation.

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, Lisbon, Portugal.

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, Singapore.

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

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 (AMTA)*, pages 223–231, Cambridge, MA.