

Simple Beats Spend: Cost-Performance Trade-Offs for Multi-Agent Machine Translation of Financial Long Documents

Arturo Oncevay Zhijin Guo Elena Kochkina Damiano Fassina*
Simerjot Kaur Charese H. Smiley Xiaomo Liu Manuela Veloso*

JPMorgan AI Research
first.last@jpmorgan.com

Abstract

Agentic MT frameworks often adopt human-inspired roles, layered post-editing, and complex workflows. To stress-test these designs, we examine cost–performance trade-offs across five multi-agent MT workflows on specialized long documents. Using six language pairs from two financial test sets and LLMs spanning lower to higher capacity and cost tiers, we evaluate translation quality and terminology accuracy relative to token-level cost. We find: (i) lean pipelines match the quality of more complex systems at lower cost; (ii) terminology accuracy depends more on model capacity than on workflow complexity; (iii) document-length resilience is not universal, as some simple and some complex workflows maintain quality on longer documents; and (iv) post-editing agents offer a better return on model capacity than base translation. Edited-span analysis further suggests that many post-edits are stylistic or lexical alternatives rather than consistently preferred improvements.

1 Introduction

Recent advances in large language models (LLMs) have enabled multi-agent systems for machine translation (MT), opening new avenues for improving translation quality (Briva-Iglesias, 2025). Most existing frameworks adopt human-inspired roles, including pre-drafting (Briakou et al., 2024), proofreading (Kim, 2025a), and roles grounded in MQM such as style reviewer (Wang et al., 2025a); whereas others use agents’ memory components to tackle document-level MT sentence-by-sentence (Wang et al., 2025c). Despite these innovations, the trade-off between translation performance and cost remains underexplored.

Besides, agent-based approaches have not proven robust for general MT, as observed in the most recent WMT 2025 General MT Shared Task

(Kocmi et al., 2025), where agent-based submissions (Kim, 2025a; Grubišić and Korencić, 2025) underperformed. However, in light of the shared task’s takeaway—“Time to Stop Evaluating on Easy Test Sets”—we are motivated to stress-test agentic MT on financial long documents, where translation quality and terminology fidelity are critical (Semenov et al., 2025; Oncevay et al., 2025b,a).

Here we systematically investigate the cost–performance trade-offs in multi-agent document-level translation. We compare five agentic workflows (§3.2) across six language pairs from two financial document datasets (§3.1); using four LLMs that vary in capacity and cost, and a suite of metrics covering translation quality, terminology, and model cost (§3.3). Our research questions: (i) Are simpler multi-agent MT workflows cost-effective without quality loss? (§4.1) (ii) For specialized terminology accuracy, does model capacity matter more than workflow complexity? (§4.2) (iii) How resilient are different workflow designs to increased document length? (§4.3) (iv) Where should model capacity be allocated for best cost–performance: base translation or post-editing? (§4.4). We also perform edited-span analysis to test whether post-editing improves the spans it modifies (§4.5).

Throughout this paper, we use “multi-agent MT” following the terminology of the systems we evaluate. Most are lightweight multi-role LLM workflows: they chain prompted roles such as translation, refinement, review, or proofreading, with limited deterministic processing and no autonomous orchestration, routing, tool use, or stopping policies. Our claims therefore concern multi-role agentic MT workflows, not the full space of autonomous agent systems.

2 Related Work

Most agentic MT systems adopt prompt chaining, where the output of one role feeds the next.

*Contribution done while working at JPMorgan.

Examples, ordered by agent count, include **IRB** (Grubišić and Korencić, 2025) (translation then self-refinement), **MAMT** (Kim, 2025a,b) (translation, post-editing, proofreading), and **STEP-BY-STEP** (Briakou et al., 2024) (pre-drafting research, drafting, refinement, proofreading). More complex pipelines introduce parallel review, as in **MAATS** (Wang et al., 2025a), which combines translation with seven MQM-based reviews followed by final refinement. Beyond chaining, segment-by-segment processing with memory agents at the sentence level has been explored alongside a translator agent (**DELTA**; Wang et al., 2025c).¹

Cost-performance trade-offs remain underexplored. Among these systems, only **MAMT** reports a cost-related ablation.² **STEP-BY-STEP** and **DELTA** evaluate at document level, but the former uses short general-domain documents (178 tokens on average), while the latter focuses on long literary texts. In contrast, we compare reproducible agentic MT workflows in a shared setup on financial documents, where terminology fidelity is central.

Concurrent work on AIDATerm (Di Rosa, 2026) studies terminology-constrained MT with Analysis, Translation, Post-editing, and Review agents. It reports strong segment-level results through reasoning-enabled review. Our work instead focuses on document-level MT and the cost of workflow complexity across multiple agentic systems.

3 Experimental Setup

3.1 Data: Financial Long Texts

We use two document-level test sets in the financial domain. First, **DOLFIN** (Nakhle et al., 2025) contains translated documents across multiple financial types, including contracts, profiling, and periodic reports. We filter to four language pairs: English (En)→German (De), En→Spanish (Es), En→French (Fr), and En→Italian (It). We include only texts whose English side is between 1k and 3k tokens, counted with the tiktoken tokenizer.³

Second, we use the document-level test set from the WMT25 Terminology Shared Task (Track 2) (**WMT25-Term**; Semenov et al., 2025), comprising translated annual reports from the Hong Kong Monetary Authority (macroeconomics and policy-making topics) in English↔Traditional Chinese

(Zht). Each document exceeds 1k tokens on the English side and includes terminology annotations that we leverage for further analysis. Statistics for both datasets appear in Table 1 in the Appendix.

3.2 Agentic Systems

We experiment with the **IRB** (Grubišić and Korencić, 2025), **MAMT** (Kim, 2025a,b), **STEP-BY-STEP** (Briakou et al., 2024), **MAATS** (Wang et al., 2025a) and **DELTA** (Wang et al., 2025c) agentic MT systems, illustrated in Figure 1 (for DELTA diagram, see the original publication). DELTA differs from the other workflows in that it is memory-augmented: it translates sentence by sentence while retrieving and updating multi-level document memory. The remaining systems are primarily prompt-chained multi-role workflows, sometimes with deterministic processing between steps, such as integrating post-edits before a final review in MAMT.

These are reproducible systems (code or prompts published) with high performance. We also consider a **zero-shot** translation prompt baseline as provided in the WMT25 General MT Shared Task (Kocmi et al., 2025).

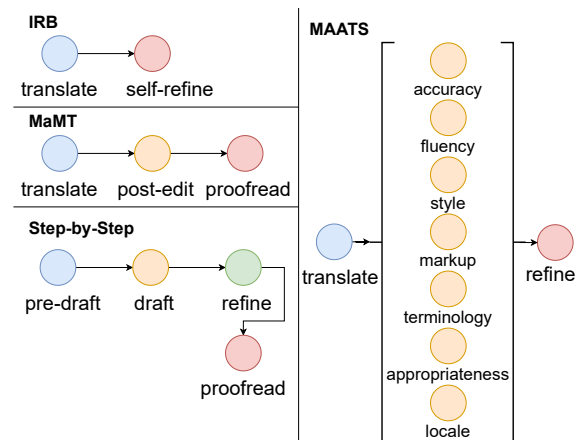


Figure 1: Diagrams of multi-agent translation systems.

3.3 Models, Cost, and Inference

We evaluate models spanning capacity and cost tiers, from higher to lower price-per-token: GPT-4.1, Qwen3-235B, Qwen3-32B (Yang et al., 2025), and GPT-4.1 nano. Prices appear in Appendix Table 2. Unless otherwise stated, cost refers to API token cost for inference over the test sets, computed from input and output tokens.

Workflows are implemented with LangChain; APIs via AWS Bedrock for Qwen 3 and OpenAI for GPT-4.1/nano. We set the temperature to 0 for

¹TransAgents (Wu et al., 2024) also emulates a translation company with specialized roles, but we do not include it because its prompts and document test sets are proprietary.

²Appendix A discusses broader cost-performance studies.

³<https://github.com/openai/tiktoken>

reproducibility, except where workflows specify otherwise (Kim, 2025a).

3.4 Evaluation Metrics

We use **MetricX-24**⁴ (Juraska et al., 2024), an embedding-based metric, as the primary quality metric, where we leverage a SEGAL-like approach (Wang et al., 2025b) for long documents (see Appendix C). We compute **Terminology Accuracy** (TermAcc) with the WMT25 Terminology Shared Task implementation (Semenov et al., 2025) and complement results with **chrF++** (Popović, 2017). For ranking, we adopt **Pareto efficiency** over performance vs. token-level cost.

We also run **paired two-sided Wilcoxon signed-rank tests** on per-sample scores, with paired approximate randomization as a sensitivity check. Each agentic workflow is compared against the zero-shot baseline using the same underlying model and language pair.

4 Cost-Performance Trade-Offs

Figures 2 and 3 present the cost-performance trade-off results.⁵ For legibility, the plots distinguish workflows by color but omit per-model markers; full per-model results, including statistical tests, are in Appendix Tables 4a and 4b.

4.1 Simple Beats Spend

We observe three consistent patterns across both datasets: (i) No agentic workflow consistently outperforms the others once model capacity is taken into account. Points for different workflows largely cluster, and rank-1 systems include lean agent pipelines (IRB, STEP-BY-STEP) alongside the zero-shot baseline. (ii) The zero-shot baseline is competitive at a fraction of the cost, often near the Pareto frontier with GPT-4.1 and Qwen3-235B. (iii) Adding agents or complex designs raises token cost but does not yield clear gains in MetricX. Dataset effects are also evident: WMT25-Term is more challenging than DOLFIN, with higher MetricX ($\downarrow$) and a wider vertical spread. These patterns are consistent with the different language pairs across the two test sets.

Statistical tests support this pattern: on MetricX, only 6% of DOLFIN and 12% of WMT25-Term

⁴MetricX-24-hybrid-large-v2p6; reference-based; 512-token context window; range 0-25; lower is better ($\downarrow$).

⁵Across our analyses, MetricX and chrF++ yield the same conclusions, so we report only MetricX in the main text.

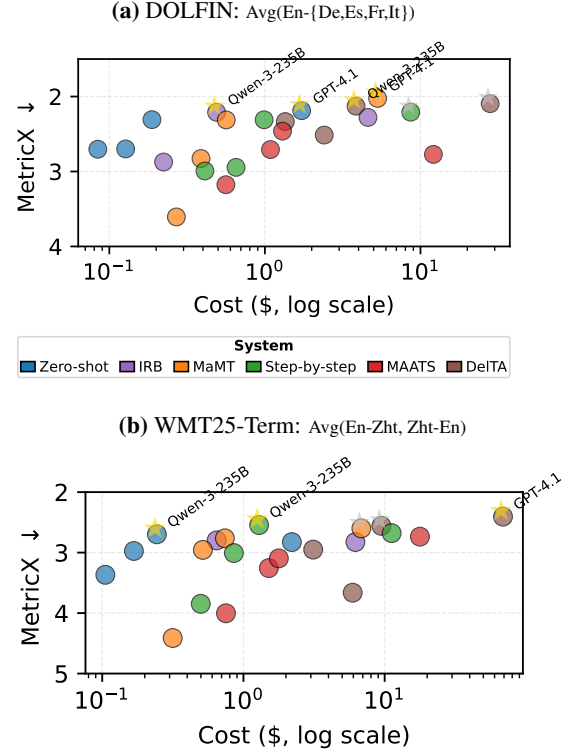


Figure 2: MetricX ($\downarrow$) performance–cost trade-offs on DOLFIN (a) and WMT25-Term (b). Pareto-optimal systems are marked $\star$ (rank 1) and $\star$ (rank 2). Base-model labels appear only on rank-1 systems to reduce clutter. Ranking considers only the top 25% by MetricX.

agentic settings significantly improve over the zero-shot counterpart with the same model and language pair; the rest are indistinguishable from or worse than zero-shot. See Appendix §D.1 for details.

4.2 Terminology Accuracy

We assess whether model choice or workflow choice has a greater impact on terminology accuracy. Figure 3 summarizes the main results. The points partition into two clear capacity tiers. High-capacity models (GPT-4.1, Qwen3-235B) form an upper cluster with higher TermAcc, while lower-capacity models (GPT-4.1 nano, Qwen3-32B) form a lower cluster. This separation persists across workflows, indicating that terminology accuracy depends more on model capacity than on workflow complexity. Thus, in our setting, terminology fidelity behaves more like a model-capacity problem than a workflow-design problem.

Within the upper tier, lean pipelines are competitive. MAMT (three agents) achieves strong terminology accuracy at moderate cost. By contrast, MAATS, despite including a specialized terminology review agent, does not produce a consistent

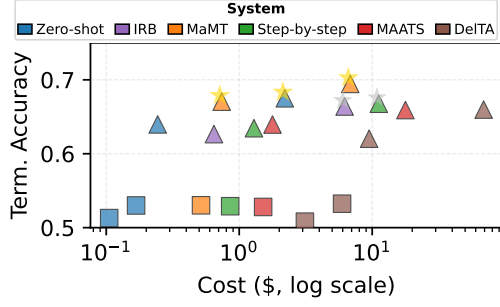


Figure 3: TermAcc vs. cost (log scale) on WMT25-Term. Shapes indicate capacity tiers: $\blacktriangle$ higher-capacity models (GPT-4.1, Qwen3-235B) and $\blacksquare$ lower-capacity models (GPT-4.1 nano, Qwen3-32B). Colors indicate workflows. Pareto-optimal systems: rank 1 $\star$, rank 2 $\star$

advantage across models. STEP-BY-STEP devotes its first agent to pre-drafting for idioms and terminology; it yields moderate gains (e.g., a Rank 2 result with Qwen3-235B, as seen in Table 5), but not a systematic lead over simpler designs.

4.3 Resilience to Document Length

We assess sensitivity to document length by partitioning each test set at the median number of English tokens and comparing performance on the shorter half vs. the longer half. Figure 4 illustrates the results for WMT25-Term with Qwen3-235B (shorter half marked by $\times$, longer half by $\star$); other models appear in Appendix Figure 7.

Three main observations emerge: (i) Longer documents often hinder MetricX and TermAcc, but the effect on the former varies by workflow and model. (ii) Notably, lean 2-3 agent workflows, such as IRB with all models and MAMT with GPT-4.1 show better MetricX performance in the longer half. (iii) The long-document-specific workflow (DELTA) does not show any improvement on longer documents and it largely holds steady. Other simpler, lower-cost pipelines remain comparably resilient.

4.4 Post-Editing vs. Base Translation

We study heterogeneous settings, comparing where to allocate higher model capacity in three workflows with clear role separation (IRB, MAMT, MAATS), using two models: GPT-4.1 (high capacity) and GPT-4.1 nano (low). We report four allocations: (i) High-only: GPT-4.1, (ii) Low-only: GPT-4.1 nano, (iii) Low $\rightarrow$ High: nano for translation, GPT-4.1 for post-editing, (iv) High $\rightarrow$ Low: GPT-4.1 for translation, nano for post-editing.

Figure 5 shows MetricX vs. cost on the two

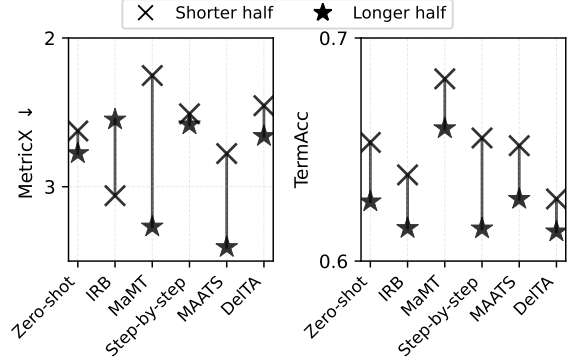


Figure 4: Length analysis on WMT25-Term with Qwen3-235B. The test set is halved by token length.

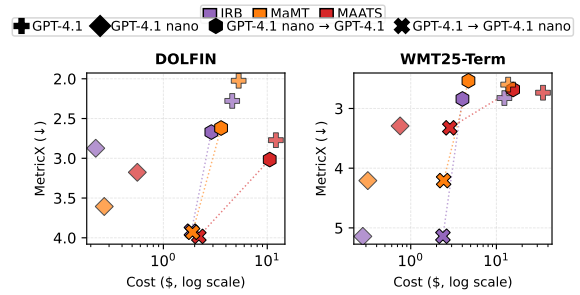


Figure 5: MetricX vs. cost for IRB, MAMT, and MAATS under four capacity allocations. Marker shapes encode allocations. Dotted arrows connect the two mixed allocations per workflow to highlight the advantage of placing the higher-capacity model in post-editing.

datasets. Two findings are consistent across workflows and datasets: (i) Low $\rightarrow$ High outperforms High $\rightarrow$ Low. Placing the higher-capacity model in post-editing yields better quality per dollar than placing it in the first-pass translation. (ii) Low $\rightarrow$ High approaches High-only quality at a fraction of the cost and achieves comparable performance on WMT25-Term.

4.5 Edited-Span Analysis

Since document-level metrics cannot capture the subtle benefits of different workflows, we examine whether post-editing agents improve the spans they modify. Appendix E expands technical details.

Appendix Table 7 reports metrics indicating the amount of edits performed across four workflows with refine or post-edit steps (IRB, MAMT, STEP-BY-STEP, MAATS), such as normalized edit distance (NED) (%) and percentage of documents with at least one edit. Editing ranges from near-zero (IRB with GPT-4.1: Normalized Edit Distance $< 1\%$) to aggressive rewriting (IRB with

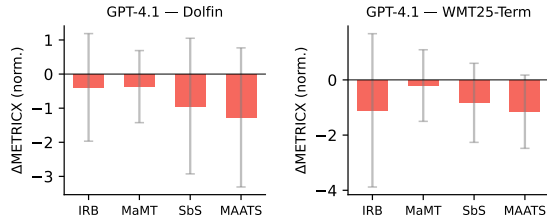


Figure 6: Character-weighted mean Δ MetricX per workflow (positive = improvement) using GPT-4.1. Other models are included in Appendix Figure 8.

Qwen3-32B: NED > 18%), yet in all cases over 90% of document text remains unchanged.

We extract edited spans by applying Myers diff (Myers, 1986) between the initial translation and the final output. To focus on substantive local changes, we retain only *replacement* edits of at least 10 characters that cover at most 3 sentences. For each qualifying span, we compute MetricX for both stages—initial and post-edited—against the aligned reference span (Appendix E.1).

Figure 6 reports the character-weighted mean Δ MetricX per workflow. Across both datasets, deltas are consistently negative or near zero, suggesting that post-editing does not reliably improve span-level MetricX, though MetricX may have limited sensitivity to small stylistic or lexical edits.

LLM Judge. To validate these findings with a judgment-based metric, we conduct an LLM-as-judge preference evaluation (Appendix E.2) over the same qualifying spans (three independent replicates per span). Appendix Table 7 reports the edit win rate (E%) and inter-replicate agreement. Across all models the average E% remains below 50%, confirming that the initial translation is preferred over the post-edited version more often than not. IRB, the leanest workflow, achieves the highest edit win rate (42% on average), while MAMT edits are the least preferred (29%). These results suggest that post-editing agents do not systematically improve translation quality at the span level; rather, their modifications reflect alternative stylistic or lexical choices that are not consistently favored by the judges.

Human Validation. As a small validation of the LLM-as-judge results, one Spanish native speaker with finance experience annotated blind pre-edit vs. post-edit pairs for GPT-4.1 on DOLFIN English→Spanish. We sampled 20 spans each from IRB and STEP-BY-STEP. The annotator pre-

ferred the edited span in 40% of IRB cases and 25% of STEP-BY-STEP cases, with human/LLM agreement of 70% and 85%, respectively. The annotator noted that many pairs were acceptable in both versions and often differed only in style, literalness, or lexical choice. Although limited in scale, this check is consistent with our conclusion that post-editing does not yield systematic span-level gains; full details are in Appendix E.3.

Together with §4.4, these findings support the capacity-allocation result: post-editing can improve system-level cost-performance, but not every local edit is beneficial. Many edits appear to be selective corrections or stylistic/lexical alternatives. Appendix F discusses broader qualitative patterns.

5 Conclusion and Recommendations

To our knowledge, we present the first unified cost-performance analysis of multi-agent MT on financial long documents. Our results do not imply that multi-agent workflows are never useful; rather, they support a simple design principle: start with a lean workflow and add roles only when a specific need is demonstrated, since marginal quality gains rarely justify added token cost. For terminology fidelity, prioritize model capacity over workflow complexity; terminology-focused agents alone do not compensate for lower-capacity models. For long documents, specific designs are not uniquely robust. Finally, when budget permits only one higher-capacity slot, assign it to post-editing rather than to the first-pass translation.

These recommendations do not rule out richer agentic or model-level features such as orchestration or reasoning. Concurrent work (Di Rosa, 2026) shows that reasoning-enabled review can be effective for segment-level terminology-constrained MT. Our document-level results complement this finding: such additions may improve specialized cases, but their value should be assessed relative to model capacity and marginal token cost.

Limitations

Agentic Workflows To our knowledge, we include most reproducible multi-agent MT systems available at the time of submission. However, the workflow architecture is narrow: most use prompt chaining, one uses parallel review, and another uses memory agents. None of these systems include orchestrators (e.g., deciding when to stop post-

editing) or tool-calling capabilities (e.g., querying terminology databases).

Model Selection We cover models across lower to higher capacity from two providers, constrained by availability in our research environment. The goal is not to benchmark models or identify a single best model. Reasoning-centric models that use separate “reasoning tokens” are outside the scope and would require additional cost accounting.

Prompts and Temperature We use prompts as provided in prior code repositories or papers. Prompt tuning can help, especially for lower-capacity models. For example, Qwen3-32B underperformed for IRB because it produced very short self-refinement outputs (e.g., “this part of the translation remains unchanged”), leading to under-translation. To maintain comparability, we keep prompts and settings fixed across systems. We set temperature to 0 for reproducibility, except where prior work specifies otherwise.

Cost Analysis We consider only token cost (input and output) from API calls. We do not include costs related to compute infrastructure nor do we measure latency or energy consumption. Extending the analysis to these dimensions is future work; token cost is a representative proxy for our setting.

Language and Domain Diversity Our test sets cover six language pairs, mostly high-resource in NLP, and focus on financial documents (including multiple subdomains). This reflects the limited availability of specialized, document-level test sets for low-resource language pairs. We encourage the community to develop broader resources to stress-test agentic MT in challenging scenarios.

Acknowledgments

We are thankful to Samuel Mensah, Joy Sain, and other members of the JPMorgan AI Research team for their insightful feedback since early stages of this work. We also appreciate the feedback of the anonymous reviewers and meta-reviewer.

Disclaimer

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates (“JP Morgan”) and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims

all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

© 2026 JPMorgan Chase & Co. All rights reserved.

References

- Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. [Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.
- Vicent Briva-Iglesias. 2025. [Are AI agents the new machine translation frontier? challenges and opportunities of single- and multi-agent systems for multilingual digital communication](#). In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 365–377, Geneva, Switzerland. European Association for Machine Translation.
- Emanuele Di Rosa. 2026. [Multi-agent orchestration for terminology-constrained machine translation in industrial localization](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 917–926, San Diego, California, USA. Association for Computational Linguistics.
- Tobias Domhan and Dawei Zhu. 2025. [Same evaluation, more tokens: On the effect of input length for machine translation evaluation using large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7929–7947, Suzhou, China. Association for Computational Linguistics.
- Jared Fernandez, Clara Na, Vashisth Tiwari, Yonatan Bisk, Sasha Luccioni, and Emma Strubell. 2025. [Energy considerations of large language model inference and efficiency optimizations](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32556–32569, Vienna, Austria. Association for Computational Linguistics.
- Ivan Grubišić and Damir Korencic. 2025. [IRB-MT at WMT25 translation task: A simple agentic system using an off-the-shelf LLM](#). In *Proceedings of the*

- Tenth Conference on Machine Translation*, pages 753–764, Suzhou, China. Association for Computational Linguistics.
- Hemang Jain, Shailender Goyal, Divyansh Pandey, and Karthik Vaidhyanathan. 2025. [Dissecting transformers: A clear perspective towards green ai](#). *Preprint*, arXiv:2510.02810.
- Fleur Jeanquartier, Claire Jean-Quartier, Paul Rieder, Vedad Misirlić, Christian Pasero, Richard Hohensinner, Heimo Müller, and Andreas Holzinger. 2026. [Assessing the carbon footprint of language models: Towards sustainability in ai](#). *Resources, Conservation and Recycling*, 226:108670.
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. [MetricX-24: The Google submission to the WMT 2024 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.
- Ahrii Kim. 2025a. [Multi-agentMT: Deploying AI agent in the WMT25 shared task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 769–777, Suzhou, China. Association for Computational Linguistics.
- Ahrii Kim. 2025b. [A preliminary study of AI agent model in machine translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 583–586, Suzhou, China. Association for Computational Linguistics.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakounga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- Eugene W Myers. 1986. An $O(ND)$ difference algorithm and its variations. *Algorithmica*, 1(1):251–266.
- Mariam Nakhle, Marco Dinarelli, Raheel Qader, Emmanuelle Esperança-Rodier, and Hervé Blanchon. 2025. [DOLFIN - document-level financial test-set for machine translation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5544–5556, Albuquerque, New Mexico. Association for Computational Linguistics.
- Nienke Nijkamp, June Sallou, Niels van der Heijden, and Luís Cruz. 2024. [Green ai in action: Strategic model selection for ensembles in production](#). In *Proceedings of the 1st ACM International Conference on AI-Powered Software*, AIware 2024, page 50–58, New York, NY, USA. Association for Computing Machinery.
- Arturo Oncevay, Elena Kochkina, Keshav Ramani, Toyin Aguda, Simerjot Kaur, and Charese Smiley. 2025a. [Translating domain-specific terminology in typologically-diverse languages: A study in tax and financial education](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35018–35032, Suzhou, China. Association for Computational Linguistics.
- Arturo Oncevay, Charese H. Smiley, and Xiaomo Liu. 2025b. [The impact of domain-specific terminology on machine translation for finance in european languages](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, New Mexico, US. Association for Computational Linguistics.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Jon Saad-Falcon, Avanika Narayan, Hakki Orhun Akengin, J. Wes Griffin, Herumb Shandilya, Adrian Gamarra Lafuente, Medhya Goel, Rebecca Joseph, Shlok Natarajan, Etash Kumar Guha, Shang Zhu, Ben Athiwaratkun, John Hennessy, Azalia Mirhoseini, and Christopher Ré. 2025. [Intelligence per watt: Measuring intelligence efficiency of local ai](#). *Preprint*, arXiv:2511.07885.
- Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. [Findings of the WMT25 terminology translation task: Terminology is useful especially for good MTs](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 554–576, Suzhou, China. Association for Computational Linguistics.
- Daria Smirnova, Hamid Nasiri, Marta Adamska, Zhengxin Yu, and Peter Garraghan. 2025. [Energy-aware data-driven model selection in llm-orchestrated ai systems](#). *Preprint*, arXiv:2512.01099.
- George Wang, Jiaqian Hu, and Safinah Ali. 2025a. [MAATS: A multi-agent automated translation system based on MQM evaluation](#). *Preprint*, arXiv:2505.14848.
- Kuang-Da Wang, Shuoyang Ding, Chao-Han Huck Yang, Ping-Chun Hsieh, Wen-Chih Peng, Vitaly Lavrukhin, and Boris Ginsburg. 2025b. [Extending automatic machine translation evaluation to book-length documents](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32311–32327, Suzhou, China. Association for Computational Linguistics.
- Yutong Wang, Jiali Zeng, Xuebo Liu, Derek Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025c. [DeTA: An online document-level translation agent based on multi-level memory](#). In *International Conference on Representation Learning*, volume 2025, pages 15708–15731.

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024. [TransAgents: Build your translation company with language agents](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Zeyu Yang and Wesley Armour. 2025. [The hidden joules: Evaluating the energy consumption of vision backbones for progress towards more efficient model inference](#). In *Forty-second International Conference on Machine Learning*.

A Related Work: Cost-Performance Trade-Off

Recent research increasingly emphasizes the importance of evaluating machine learning systems not only by task performance, but also by associated costs such as energy consumption, computational resources, latency, and environmental impact. Several studies have explored the energy–performance trade-offs in large language models and inference optimization (Fernandez et al., 2025; Jain et al., 2025; Yang and Armour, 2025), proposed energy-aware model selection strategies (Smirnova et al., 2025; Nijkamp et al., 2024), and introduced unified metrics for benchmarking cost–accuracy relationships across different hardware configurations (Saad-Falcon et al., 2025; Jeanquartier et al., 2026). While significant progress has been made in understanding cost–performance trade-offs across various machine learning domains, a unified analysis in multi-agent MT has yet to be established. We therefore present what is, to the best of our knowledge, the first comprehensive cost–performance analysis of multi-agent MT.

B Experimental Setup

Table 1 includes data statistics; Table 2 presents the model pricing (per token); Table 3 summarizes the workflow steps.

	DOLFIN			WMT25-Term		
	En-De	En-Es	En-Fr	En-It	En-Zht	Zht-En
Docs	13	17	31	13	55	56
Avg Sent/Doc	30.5	22.5	27.3	19.8	46.9	47.6
Tokens _(Src)	19.0k	21.2k	47.0k	19.3k	103.0k	82.4k
Tokens _(Tgt)	32.0k	33.4k	72.2k	32.2k	77.2k	108.7k
Avg Tok _(Src)	1.5k	1.2k	1.5k	1.5k	1.9k	1.5k
Avg Tok _(Tgt)	2.5k	2.0k	2.3k	2.5k	1.4k	1.9k

Table 1: Dataset statistics.

Model	Input	Output
GPT-4.1 ₍₂₀₂₅₋₀₄₋₁₄₎	\$0.00200	\$0.00800
Qwen3-235B	\$0.00022	\$0.00088
Qwen3-32B	\$0.00015	\$0.00060
GPT-4.1 nano ₍₂₀₂₅₋₀₄₋₁₄₎	\$0.00010	\$0.00040

Table 2: Model pricing per 1,000 tokens (standard tier). Source: AWS Bedrock and OpenAI’s official sites, accessed on: 01/01/2026.

Step	IRB	MaMT	SBS	MAATS	DELTA
Pre-drafting			✓		
Translation	✓	✓		✓	✓
Drafting translation			✓		
Self-Refine	✓				
Refinement			✓	✓	
Post-editing		✓			
Proofreading		✓	✓		
MQM annotation				✓	

Table 3: Agentic System Workflow Steps (Abbreviated System Names)

C Document-level MetricX

MetricX-24 (Juraska et al., 2024) supports a context window of 512 tokens, which is insufficient for evaluating documents up to 2-3k tokens end-to-end. We therefore apply MetricX at the segment level.

We build on SEGAL (Wang et al., 2025b), which segments, aligns, and scores system outputs in chunks and is robust to under- and over-translation—an important property for our analysis. Instead of running segmentation and alignment as separate steps, we fuse them into a single LLM-based procedure inspired by Focus-Sentence Prompting (FSP) (Domhan and Zhu, 2025). Concretely, we first provide the full source, system output, and reference to the LLM. In subsequent turns, the LLM extracts a sequence of aligned tuples, where each tuple corresponds to a discourse unit of 1–3 sentences.

After extraction, we automatically verify coverage of both the source and the system output across the aligned tuples. If coverage falls below 95%, we re-run the alignment for that sample. Finally, we score each aligned tuple with MetricX. If a tuple contains an empty alignment on either side (i.e., indicating under- or over-translation), we assign the maximum penalty (25) for that segment. For this experiment, we use Claude Sonnet 4.6 as the aligner. Appendix G includes the prompt.

D Complementary Results

Tables 4a and 4b present the main results of the cost-performance trade-off using MetricX, chrF++ and Terminology Accuracy (for WMT25-Term), including statistical significance tests. Figure 7 shows the length analysis for three models. Figure 8 shows the span-level Δ MetricX for three models.

D.1 Statistical Significance Tests

We ran paired two-sided Wilcoxon signed-rank tests on per-sample MetricX-24 and TermAcc scores, using paired approximate randomization as a sensitivity check. Each agentic workflow was compared against the zero-shot baseline with the same underlying model and language pair. We use $\alpha = 0.05$.

For DOLFIN MetricX-24, out of 80 agentic comparisons (workflow $\times$ model $\times$ language pair), only 5 (6%) show a statistically significant improvement over zero-shot: 3 from DELTA and 2 from MaMT. Another 2 comparisons are significantly worse, and 73 are not significantly different.

For WMT25-Term MetricX-24, out of 40 comparisons, only 5 (12%) are significantly better than zero-shot. These wins are spread across DELTA, MAATS, MaMT, and Step-by-step, with at most 2 wins per workflow. Another 9 comparisons are significantly worse, and 26 are not significantly different.

For Term Accuracy in WMT25-Term, out of 48 comparisons, only 2 (4%) are significantly better than zero-shot, both from IRB. Another 22 comparisons are significantly worse, and 24 are not significantly different. Thus, no agentic workflow shows a systematic Term Accuracy gain over its zero-shot counterpart.

Overall, the additional complexity and token cost of agentic workflows do not translate into systematic, statistically significant gains over zero-shot prompting. The few significant wins are scattered across workflows and models, while most settings are indistinguishable from or worse than zero-shot.

System		MetricX ↓					chrF++ ↑					Cost (\$)
		Avg	En-De	En-Es	En-Fr	En-It	Avg	En-De	En-Es	En-Fr	En-It	Total
Zero-shot	GPT-4.1	2.19★	1.40	2.26	2.67	2.43	73.0★	72.0	72.3	74.5	73.0	1.72
	Qwen 3 235B	2.31	1.76	2.35	2.86	2.27	70.3	69.2	70.8	69.8	71.2	0.19
	Qwen 3 32B	2.70	2.19	2.68	3.11	2.81	67.2	65.2	68.0	67.4	68.2	0.13
	GPT-4.1 nano	2.70	2.07	3.34	2.89	2.51	69.4	67.7	68.9	69.4	71.7	0.08
IRB	GPT-4.1	2.28	1.57	2.49	2.60	2.44	73.0★	72.3	72.2	74.4	73.0	4.60
	Qwen 3 235B	2.21★	1.87	2.19	2.68	2.11	69.7	67.3	71.2	69.6	70.6	0.49
	Qwen 3 32B	6.32‡	6.21	6.37	5.62	7.09	54.7	54.2	56.0	56.8	51.9	0.32
	GPT-4.1 nano	2.87	2.08	3.13	3.31	2.97	68.7	67.6	68.8	68.1	70.2	0.22
MaMT	GPT-4.1	2.03‡★	1.23	2.15	2.53	2.19†	74.0★	73.4	72.9	75.2	74.6	5.30
	Qwen 3 235B	2.31	2.08	2.45	2.61	2.11†	70.4†	69.3	71.4	70.5	70.2	0.56
	Qwen 3 32B	2.83	2.23	3.26	3.22	2.60	67.4	65.9	68.0	67.5	68.3	0.39
	GPT-4.1 nano	3.61†	4.49	3.39	3.80‡	2.75	67.6	66.1	67.6	67.2	69.7	0.27
Step-by-step	GPT-4.1	2.21★	1.48	2.40	2.57	2.38	72.3‡★	71.7‡	71.6‡	73.2‡	72.7‡	8.66
	Qwen 3 235B	2.31	1.68	2.34	2.90	2.32	68.0‡	68.0‡	68.5‡	67.6‡	68.1‡	0.99
	Qwen 3 32B	2.95	3.36	2.72	3.05	2.65	67.3‡	65.4‡	67.9‡	68.0‡	68.0‡	0.65
	GPT-4.1 nano	2.99	1.96	3.44	3.59	2.98	67.8‡	67.4‡	67.2‡	67.5‡	69.2‡	0.41
MAATS	GPT-4.1	2.77‡	1.92	3.28	2.89†	2.99	70.1	68.5	70.2	71.4	70.3	12.11
	Qwen 3 235B	2.46	2.08	2.68	2.87	2.21	70.0	68.6	70.6	69.8	70.9	1.30
	Qwen 3 32B	2.71	2.18	2.77	3.27	2.60	66.5	65.2	67.5	65.3	67.8	1.09
	GPT-4.1 nano	3.18	2.70	3.15	4.02	2.84	68.1	67.4	68.0	66.5	70.5	0.56
DeITA	GPT-4.1	2.10★	1.38	2.17	2.60	2.22	72.3	70.3	72.6	72.9‡	73.4	28.07
	Qwen 3 235B	2.13†★	1.41	2.35	2.48†	2.27	71.0†★	69.7	71.0	70.6	72.5	3.85
	Qwen 3 32B	2.51	1.72†	2.61	3.00	2.73	68.3‡	66.8‡	69.0†	67.8	69.7	2.41
	GPT-4.1 nano	2.33‡	1.86	2.32‡	2.68	2.48	69.8‡	67.8	69.9†	69.9†	71.7	1.35

(a) DOLFIN dataset.

System		MetricX ↓			TermAcc ↑			chrF++ ↑			Cost (\$)
		Avg	En-Zh	Zh-En	Avg	En-Zh	Zh-En	Avg	En-Zh	Zh-En	Total
Zero-shot	GPT-4.1	2.83	2.91	2.75	0.68★	0.61	0.74	57.7★	40.8	74.5	2.20
	Qwen 3 235B	2.70★	2.35	3.05	0.64	0.62	0.65	56.8★	41.8	71.7	0.24
	Qwen 3 32B	2.97	2.77	3.18	0.53	0.49	0.57	53.0	35.7	70.3	0.17
	GPT-4.1 nano	3.37	3.32	3.41	0.51	0.45	0.58	49.7	32.1	67.2	0.11
IRB	GPT-4.1	2.83	2.46	3.19	0.66★	0.62†	0.71	56.3	40.9	71.7	6.20
	Qwen 3 235B	2.80	2.42	3.18	0.63	0.62	0.64	55.5	41.2	69.9	0.65
	Qwen 3 32B	10.74‡	7.77‡	13.71‡	0.33	0.38†	0.28	31.5	28.0	35.0‡	0.38
	GPT-4.1 nano	6.04	4.54†	7.55	0.44	0.42	0.46	42.7	30.1	55.3	0.28
MaMT	GPT-4.1	2.60†★	2.44†	2.76	0.69‡★	0.64‡	0.75	58.3‡★	41.7†	74.9‡	6.84
	Qwen 3 235B	2.76‡	2.65‡	2.88†	0.67‡★	0.67‡	0.67‡	58.0‡★	43.6‡	72.5	0.74
	Qwen 3 32B	2.96	2.78	3.13	0.53‡	0.49‡	0.57†	53.2	36.2	70.3	0.52
	GPT-4.1 nano	4.41	4.52	4.30	0.46‡	0.39‡	0.53†	45.2‡	27.7‡	62.8‡	0.32
Step-by-step	GPT-4.1	2.68	2.52	2.84†	0.67‡★	0.62‡	0.72‡	57.2‡	40.7‡	73.6‡	11.21
	Qwen 3 235B	2.55★	2.44	2.65	0.63‡	0.61‡	0.65‡	55.5‡	40.5‡	70.5‡	1.29
	Qwen 3 32B	3.01	2.74†	3.28	0.53‡	0.48‡	0.58‡	52.6‡	35.0‡	70.1‡	0.86
	GPT-4.1 nano	3.85†	4.15	3.55†	0.45‡	0.39‡	0.52‡	45.9‡	28.6‡	63.3‡	0.50
MAATS	GPT-4.1	2.74‡	2.59	2.88‡	0.66	0.60	0.71	56.5	39.5	73.5	17.77
	Qwen 3 235B	3.10	3.20	2.99	0.64	0.62	0.66	56.5★	41.3	71.7	1.78
	Qwen 3 32B	3.26	3.35†	3.16	0.53	0.48	0.57	52.9	35.5	70.3	1.51
	GPT-4.1 nano	4.00†	3.25‡	4.76	0.43	0.43	0.44	41.1	31.1	51.1	0.75
DeITA	GPT-4.1	2.41★	2.45	2.36	0.66‡	0.59‡	0.73	56.5‡	39.1‡	73.8	68.87
	Qwen 3 235B	2.56★	2.55‡	2.56	0.62‡	0.60‡	0.64†	55.7‡	39.7‡	71.7	9.48
	Qwen 3 32B	3.66	3.27	4.05	0.53	0.49	0.58	51.6‡	33.3‡	69.9†	5.94
	GPT-4.1 nano	2.95‡	2.80	3.10‡	0.51	0.45	0.57	50.5‡	32.2	68.8‡	3.12

(b) WMT25-Term dataset.

Table 4: Main results for the DOLFIN and WMT25-Term datasets. Top-performing systems are labelled as rank 1★ and rank 2★ according to Pareto optimality (only systems above the 75th percentile in performance are considered for Pareto ranking). Significance markers (†: $p < 0.05$, ‡: $p < 0.01$) come from paired Wilcoxon signed-rank tests against the Zero-shot baseline using the same model; the Avg column uses a pooled test across language pairs.

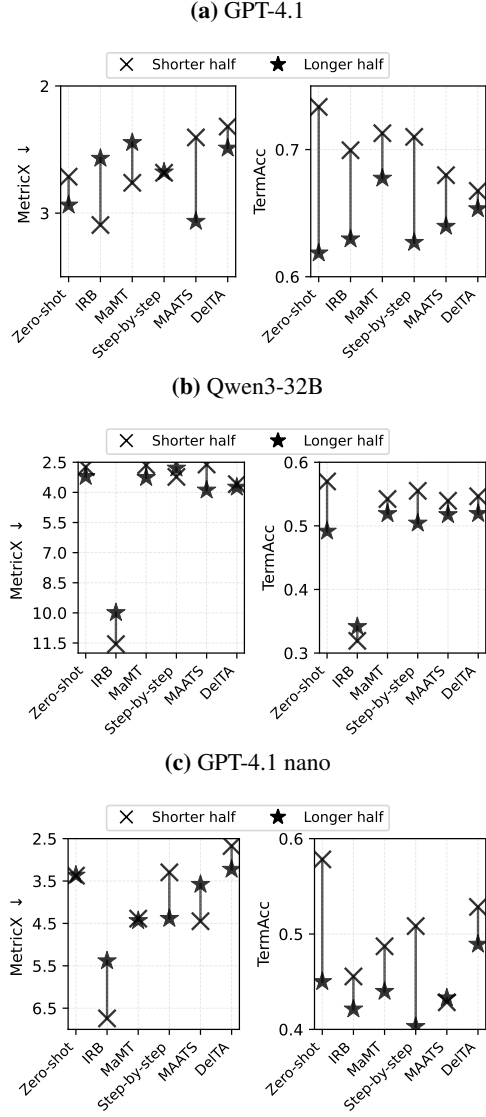


Figure 7: Length analysis on WMT25-Term with GPT-4.1 (a), Qwen3-32B (b) and GPT-4.1 nano (c). The test set is halved by token length.

E Span-Level Analysis

In this section, we complement details for the analysis of post-editing behavior at the span level by quantifying how much each workflow edits and whether those edits are preferred. Table 7 summarizes editing effort (e.g., NED, % of documents edited, and edited-span statistics) as well as LLM-as-judge outcomes (E% and agreement).

E.1 Span-Level Alignment and Evaluation

Following our document-level setup (§C), we use an FSP-style multi-turn prompt to align span-diff edits to the source and reference. We first identify edited spans between the initial and final translations using the Myers diff algorithm. To focus

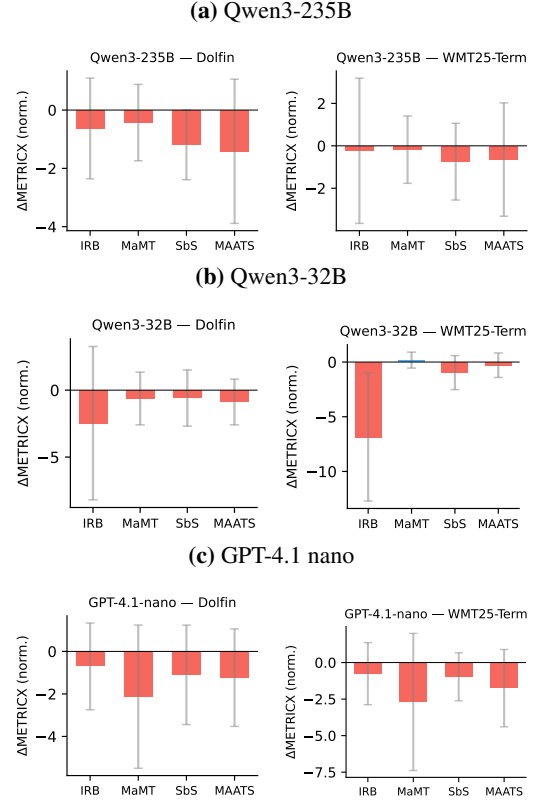


Figure 8: Character-weighted mean Δ MetricX per workflow (positive=improvement) using Qwen3-235B, Qwen3-32B and GPT-4.1 nano.

on substantive changes, we retain only *qualifying spans*: replacement edits (not pure insertions or deletions) of at least 10 characters that cover at most 3 sentences. The LLM then receives the full source, initial translation, and reference as context, and in subsequent turns extracts aligned tuples for each qualifying span (i.e., the edited target span paired with its corresponding source and reference spans). We score these aligned tuples with MetricX. We use Claude Sonnet 4.6 as the aligner. Appendix G includes the prompt.

E.2 LLM-as-Judge Preference Evaluation

To complement the reference-based MetricX scores, we conduct a preference evaluation using an LLM judge. For each qualifying span, the judge receives the source sentence, the initial translation span, and the post-edited span in a randomized A/B order (to mitigate position bias), and must select which translation is better—no ties are permitted. We use a multi-turn prompt: the first turn provides document-level context (source, reference, and full translations), and subsequent turns elicit binary preferences for individual spans.

We run three independent replicates per span at temperatures 0, 0.3, and 0.5, aggregating via majority vote. We report two metrics: **E%**, the edit win rate (percentage of spans where the post-edited version is preferred), and **Agr**, the inter-replicate agreement (percentage of spans where all three replicates concur). We use Claude Sonnet 4.6 as the judge. Appendix G includes the prompt.

E.3 Human Annotation and Judge-to-Human Agreement

To sanity-check the LLM-as-judge span evaluation, we conducted a small manual annotation on DOLFIN English→Spanish for GPT-4.1. One Spanish native speaker with finance experience annotated a blind, randomly ordered sample of pre-edit vs. post-edit span pairs. We selected IRB and STEP-BY-STEP, sampling 20 spans per workflow. The annotator chose which version was better, allowing blank or tie labels; all 40 judgments were decisive.

Table 5 summarizes the results. For IRB, the human preferred the edited translation in 40% of cases, compared with 60% for the LLM judge on the same subset. For STEP-BY-STEP, the human preferred the edited translation in 25% of cases, close to the LLM judge’s 30% on the same subset. Agreement was 70% for IRB and 85% for STEP-BY-STEP, with Cohen’s κ of 0.42 and 0.63, respectively. The annotator noted that many pairs were acceptable in both versions, with decisions often driven by minor stylistic, literalness, or lexical-preference differences. This small check does not change our main conclusion: post-editing does not yield systematic span-level quality gains in this setting.

Illustrative disagreements. Table 6 shows representative human/LLM disagreements from the manual check. These examples are not intended as additional quantitative evidence; they illustrate that disagreements often arise from capitalization, lexical preference, or small semantic shifts rather than clear-cut errors.

F Reported Qualitative Patterns

Automatic scores and LLM judges may miss some qualitative advantages of agentic translation workflows. We therefore summarize strengths and weaknesses reported in the original papers to contextualize our findings.

	IRB	Step-by-Step
# spans	20	20
Decisive / blank / tie	20 / 0 / 0	20 / 0 / 0
<i>Human judge</i>		
Init / Edit picks	12 / 8	15 / 5
Edit-win rate	40.0%	25.0%
Binom. p vs. 50%	0.50	0.04
<i>LLM judge</i>		
Edit-win rate (same spans)	60.0%	30.0%
Edit-win rate (all spans)	57.1%	37.7%
<i>Human vs. LLM</i>		
Agreement	70.0%	85.0%
Cohen’s κ	0.42	0.63

Table 5: Human vs. LLM edit-effort results on DOLFIN English→Spanish for GPT-4.1. Edit-win rate denotes preference for the post-edited span; p is a two-sided exact binomial test against 50%.

Prior work reports improvements in the settings for which these methods were originally proposed, including edits that address omissions (Wang et al., 2025a), capture nuanced phenomena such as slang, indirect translation, and humor (Briakou et al., 2024), or improve naturalness (Wang et al., 2025a). Some approaches also report gains in document-level phenomena such as proper-noun consistency, pronoun usage, and co-reference resolution (Wang et al., 2025c). However, common issues persist: hallucinations, added content not present in the source (Kim, 2025a; Briakou et al., 2024), and omissions, particularly in longer sentences (Kim, 2025a). Models can also be overly sensitive to narrow instructions, leading to overgeneralization in rare cases (Briakou et al., 2024).

Workflow	Annotator-preferred span	LLM-preferred span	Pattern
IRB	<i>La sociedad gestora</i> podrá implementar...	<i>La Sociedad Gestora</i> podrá implementar...	Capitalization-only edit; the annotator preferred the reference-aligned lower-case form.
IRB	Empresas que el Gestor de Inversiones <i>determine que incumplen...</i>	Empresas que, <i>a juicio del Gestor de Inversiones, incumplan...</i>	The edit rewrites the clause; the annotator preferred the version closer to the reference wording.
IRB	...en <i>historias de emisores individuales</i> , rotación sectorial...	...en <i>operaciones sobre emisores individuales</i> , rotación sectorial...	The edit is fluent but shifts the meaning of “single names stories” toward transactions.
STEP-BY-STEP	...el mercado <i>recuperó el optimismo...</i>	...el mercado finalmente <i>volvió al optimismo...</i>	Here the annotator preferred the edit: it is more idiomatic and closer to the reference verb, though the LLM kept the explicit rendering of “finally”.
STEP-BY-STEP	... <i>nombrar</i> subdepositarios...; esté de acuerdo con los Requisitos del Banco Central; podrá ser <i>terminado...</i>	... <i>designar</i> subdepositarios...; cumpla con los Requisitos del Banco Central; podrá ser <i>rescindido...</i>	Long legal span with mixed local effects: some edits improve register, but others mainly substitute near-synonyms. The annotator preferred the initial overall, while the LLM preferred the edited span.

Table 6: Representative annotator/LLM disagreements in the DOLFIN English→Spanish manual check; columns show the span preferred by each judge.

Table 7: Post-editing effort between initial translation (Draft for SbS) and final output. NED: normalized edit distance (%); %D: documents with at least one edit; Sp: mean edited spans per document (Myers diff); $\bar{\ell}$: mean span length (chars); QS: qualifying spans for Δ analysis (≥ 10 chars, ≤ 3 sentences, replacements only); E%: edit or final output win rate w/ LLM judges; Agr: inter-replicate agreement (3 independent runs per sample).

Model	Lang	NED	%D	Sp	$\bar{\ell}$	QS	E%	Agr
IRB								
GPT-4.1	en→de	0.9	92	5.8	561	13	42	75
	en→es	1.4	88	5.3	944	35	57	80
	en→fr	0.8	90	3.9	1082	39	44	62
	en→it	0.9	92	3.8	783	22	45	86
	en→zh	0.2	31	0.8	44	1	0	100
	zh→en	3.8	62	2.9	3387	44	55	73
Qwen3-2.5B	en→de	6.7	100	6.2	970	22	67	57
	en→es	2.9	100	5.1	737	34	50	65
	en→fr	5.6	100	5.1	916	65	55	66
	en→it	6.2	100	4.7	1138	28	36	71
	en→zh	2.8	75	2.2	165	10	70	100
	zh→en	3.3	82	3.0	589	44	20	80
Qwen3-3.2B	en→de	18.5	85	2.0	3153	7	57	100
	en→es	19.7	100	2.4	3607	12	58	75
	en→fr	18.0	87	2.8	2741	25	60	48
	en→it	26.2	100	3.8	1856	22	50	95
	en→zh	23.6	91	4.1	761	4	0	100
	zh→en	54.3	95	5.3	4869	41	32	88
GPT-4.1-nano	en→de	1.3	100	4.8	1945	16	47	73
	en→es	0.3	71	2.5	523	3	100	67
	en→fr	5.2	100	3.8	1145	48	17	77
	en→it	2.1	92	3.0	1317	18	17	67
	en→zh	7.3	78	3.5	941	1	0	100
	zh→en	17.8	98	5.6	1718	72	33	74
avg		9.6	88	3.9	1496	26	42	78
MaMT								
GPT-4.1	en→de	4.5	100	10.6	941	71	22	82
	en→es	3.1	100	7.6	780	70	27	61
	en→fr	2.9	100	8.2	1213	144	45	61
	en→it	4.8	100	6.7	1436	80	35	67
	en→zh	3.3	100	14.0	118	17	6	81
	zh→en	4.0	100	13.9	500	365	25	80
Qwen3-2.5B	en→de	3.7	92	6.5	1708	41	34	95
	en→es	2.7	94	6.4	1209	49	37	71
	en→fr	3.6	100	8.1	1235	116	37	66
	en→it	4.7	100	6.7	1407	56	30	66
	en→zh	3.9	100	14.6	128	24	17	88
	zh→en	2.0	100	9.6	510	248	35	76
Qwen3-3.2B	en→de	0.9	92	3.4	965	14	54	54
	en→es	0.7	65	3.4	574	10	50	80
	en→fr	0.7	87	3.3	1409	45	44	56
	en→it	0.8	85	3.5	2199	10	40	70
	en→zh	1.3	87	6.2	104	4	50	75
	zh→en	0.6	77	3.3	429	27	56	67
GPT-4.1-nano	en→de	22.0	92	7.9	2338	17	8	92
	en→es	14.3	94	9.1	930	31	3	94
	en→fr	19.9	97	6.7	2043	79	10	79
	en→it	14.7	92	5.7	3701	44	9	93
	en→zh	10.9	93	5.1	213	12	0	92
	zh→en	12.8	89	6.5	914	89	21	80
avg		5.9	93	7.4	1125	69	29	76
SbS								
GPT-4.1	en→de	1.4	100	6.5	994	25	35	70
	en→es	2.6	100	7.9	941	54	38	74
	en→fr	2.6	100	9.7	1255	116	49	59
	en→it	1.6	100	5.2	1768	36	64	53
	en→zh	1.5	100	9.5	112	1	0	0
	zh→en	3.0	100	18.1	419	257	26	73
Qwen3-2.5B	en→de	6.4	100	8.5	1436	58	8	83
	en→es	8.6	100	8.4	1586	87	25	71
	en→fr	10.7	100	11.0	1463	178	26	66
	en→it	8.5	100	4.9	2917	72	21	67
	en→zh	2.9	100	10.1	142	23	48	100
	zh→en	7.8	100	14.8	494	297	16	82
Qwen3-3.2B	en→de	0.3	85	1.7	726	2	50	100
	en→es	0.2	47	2.4	441	5	40	60
	en→fr	0.9	74	3.6	440	20	53	74
	en→it	0.2	62	0.9	369	5	75	75
	en→zh	0.8	87	4.0	115	3	0	100
	zh→en	0.3	71	2.1	415	25	36	56
GPT-4.1-nano	en→de	1.1	92	5.2	471	8	29	100
	en→es	3.0	94	6.0	1626	31	24	69
	en→fr	3.6	97	6.4	556	62	20	78
	en→it	2.1	100	4.4	1060	21	25	65
	en→zh	4.2	91	3.8	198	1	—	—
	zh→en	5.3	100	9.7	475	118	30	81
avg		3.3	92	6.9	851	63	32	72
MAATS								
GPT-4.1	en→de	8.8	100	10.4	1695	63	34	83
	en→es	11.2	100	12.1	1330	110	56	76
	en→fr	8.0	100	9.5	1397	166	30	75
	en→it	9.5	100	6.3	3857	65	28	88
	en→zh	9.8	100	18.7	125	25	24	92
	zh→en	6.6	100	28.0	442	403	26	82
Qwen3-2.5B	en→de	3.7	100	8.2	2769	33	48	83
	en→es	3.5	100	7.2	1155	54	30	80
	en→fr	2.4	100	7.9	1038	108	42	69
	en→it	2.2	100	6.9	1308	43	28	93
	en→zh	2.1	100	9.3	133	9	44	100
	zh→en	1.4	96	8.4	503	116	28	87
Qwen3-3.2B	en→de	2.8	100	7.6	1445	35	53	84
	en→es	4.0	100	6.1	1409	67	19	87
	en→fr	5.5	100	6.4	2289	115	27	84
	en→it	3.0	100	5.9	2001	48	40	85
	en→zh	4.2	100	12.9	135	6	17	83
	zh→en	1.9	100	9.2	553	137	35	77
GPT-4.1-nano	en→de	1.8	100	5.2	931	21	11	72
	en→es	4.2	100	5.1	2010	40	10	70
	en→fr	6.8	97	7.1	1416	91	20	70
	en→it	1.9	100	5.2	2513	50	34	62
	en→zh	4.9	91	5.0	138	5	0	80
	zh→en	24.1	96	15.9	436	90	21	83
avg		5.6	99	9.4	1293	79	29	81

G Prompts

We present the core prompts used in our document-level MetricX (§C) and span-level evaluation pipeline (§E.1, §E.2). For brevity, we show only the key instructions; full document context (source, translation, reference) is passed in XML-delimited fields.

Document-Level Alignment (Multi-Turn)

System: You are an expert in translation alignment. You will receive a source text, its machine translation, and a reference translation. Segment and align them into corresponding discourse units, one segment at a time.

Guidelines: 1. Each segment is a coherent discourse unit (1–3 sentences). 2. Segments must align semantically across source, translation, and reference. 3. Each field must not exceed 512 SentencePiece tokens. 4. ALL text from ALL three documents must appear in segments. 5. Under-translation: set "translation": "". Over-translation: set "source": "" and "reference": "".

Turn 1 (user): [source, translation, reference in XML tags] Return one JSON object per turn: {"source": "...", "translation": "...", "reference": "..."} When done: {"done": true}

Subsequent turns (user): next

Span-Level Alignment (Single-Turn)

System: You are an expert translation alignment assistant. For each excerpt from the initial translation, find and return the corresponding SOURCE and REFERENCE text.

Rules: - Return EXACT text from the source/reference. - Expand to full sentence boundaries if excerpt spans partial sentences. - If no corresponding text exists, return empty string. - Return ONLY the JSON array.

User: [source, initial translation, reference in XML tags]

For each of the following N excerpt(s), find the corresponding source and reference text. <excerpt_1> ... </excerpt_1>

Return a JSON array of N objects: [{"source": "...", "reference": "..."}, ...]

LLM-as-Judge Preference (Multi-Turn)

System: You are an expert multilingual translation quality assessor. Compare two candidate translations and choose which is better – no ties allowed.

Turn 1 (user): [source, translation A, translation B in XML tags] I will ask you to compare specific excerpts. For each, choose A or B considering: accuracy, fluency, naturalness, and terminology.

Turn 1 (assistant): Understood. I will compare each span and choose A or B. Please show me the first span.

Per-span turn (user): Source: {source_span} Translation A: {span_a} Translation B: {span_b} Which is better, A or B? Reply with ONLY: A or B

Per-span turn (assistant): A or B

Three replicates per sample (temp. 0, 0.3, 0.5) with randomized A/B assignment; majority vote determines winner.