Lexically Stable but Semantically Improvable: On the Editability of Machine Translation Drafts for LLM Post-Editing

Tianlin Xiao and Wenhan Chao
Beihang University, Beijing, China
{rookie,chaowenhan}@buaa.edu.cn

Abstract

Large language models (LLMs) perform strongly in machine translation, but their effectiveness as translation refiners depends on the draft source. We compare direct translation, self-refinement, and cross-system refinement on English-to-Chinese WMT test sets. Sockeye drafts obtain higher lexical-overlap scores but lower COMET than the stronger Qwen3 drafts, and receive larger gains after Qwen3 refinement. Additional controls qualify this pattern. Qwen3-0.6B drafts obtain the largest raw gain but lower final quality, while Qwen3-1.7B drafts start near Sockeye in COMET yet benefit less. Cross-family GLM-4-9B-Chat drafts also improve with Qwen3-8B, but achieve smaller gains and lower final quality than refined Sockeye drafts. Descriptive bucket, common-support, and regression analyses suggest that initial quality alone is an incomplete account of these differences. xCOMET-XL score and span diagnostics further characterize the contrast with Qwen3-8B self-refinement. We frame *translation editability* as a property of the evaluated draft–editor pairing, assessed jointly through quality gain, improvement probability, degradation risk, and final quality.

1 Introduction

Large language models (LLMs) are increasingly used as general-purpose machine translation (MT) systems, often producing fluent and semantically adequate translations without task-specific training (Jiao et al., 2023; Hendy et al., 2023). This raises a practical question for MT pipelines: as LLMs become stronger, do specialized MT systems still have a useful role, or should they simply be replaced by LLM translators? We argue that smaller or specialized systems may remain valuable not necessarily as final translators, but as *draft generators* whose outputs can be effectively revised by LLM post-editors. This framing suggests a parameter-scale implication: improving the draft source can be an alternative to simply increasing the size of the LLM translator or post-editor.

Automatic post-editing (APE) has traditionally studied how to correct MT outputs using source–hypothesis information (Chatterjee, 2019; Vu and Haffari, 2018). Recent LLM-based work has focused mainly on how to better guide the post-editor through prompts, feedback, or error annotations (Ki and Carpuat, 2024). We instead ask a complementary question: *what kinds of draft translations are suitable inputs for LLM post-editing?*

A natural assumption is that stronger translation systems should produce better drafts. Under this view, LLM-generated translations, which often obtain high neural quality scores, should be preferable inputs for LLM post-editing. However, this assumption overlooks *editability*. A draft may not have the highest initial semantic score, but it may preserve enough lexical and structural anchors for an LLM to make targeted corrections. Conversely, a fluent high-scoring LLM draft may leave little useful room for self-refinement.

We study this draft-source effect on English-to-Chinese WMT test sets by comparing direct translation, LLM self-refinement, and cross-system refinement with a Sockeye-based NMT system and Qwen-family LLMs. Figure 1 summarizes the settings. Our main analyses use COMET as the primary semantic metric and BLEU/chrF++ as surface-overlap metrics (Papineni et al., 2002; Popović, 2015, 2017; Rei et al., 2020), with reference-based xCOMET-XL as an additional score- and span-level check (Guerreiro et al., 2024). Additional weak-to-strong Qwen3 and cross-family GLM controls examine initial-quality headroom and comparisons beyond Qwen self-refinement.

Our results show that Sockeye drafts are lexically stable but semantically improvable: relative to the stronger Qwen3 drafts in the core comparison, they obtain strong BLEU/chrF++ but lower

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1236–1259
November 6-10, 2026 ©2026 Association for Computational Linguistics

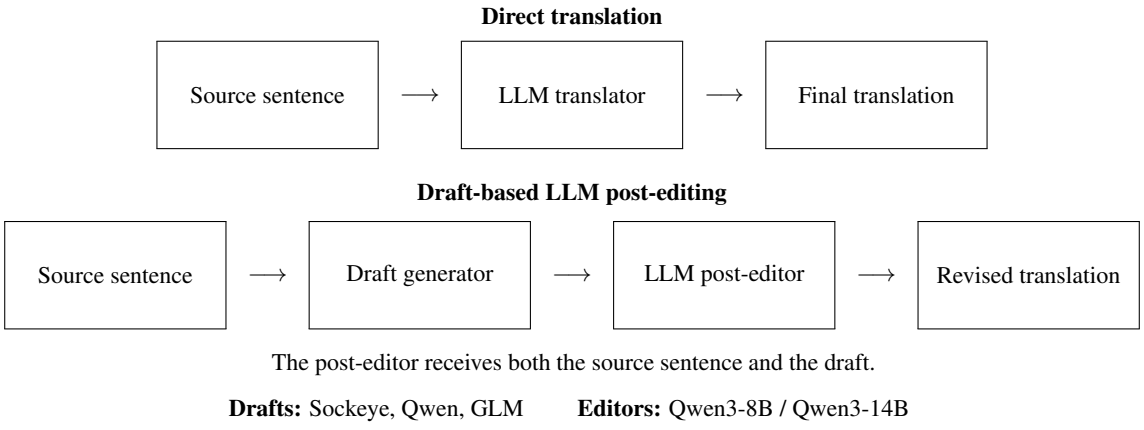


Figure 1: Study overview. We compare direct LLM translation with draft-based LLM post-editing to analyze how draft source affects post-editing gains.

COMET, and receive larger gains after refinement. The additional controls qualify this observation: the weakest Qwen3-0.6B drafts obtain the largest raw gain but remain worse in final quality, whereas similarly scoring Qwen3-1.7B drafts improve less than Sockeye. Cross-family GLM drafts also improve, without matching the final quality of refined Sockeye drafts. These findings motivate evaluating gains jointly with final quality and degradation risk.

To explain these observations, we use *translation editability* as a relational diagnostic perspective. Effective drafts are not simply high-overlap or low-scoring translations; they are drafts whose remaining errors are exposed in a locally correctable form for a complementary post-editor. This perspective reframes the role of specialized MT systems in the LLM era: they need not outperform LLMs as standalone translators to remain useful components of hybrid MT pipelines.

This paper makes the following contributions:

- We compare direct translation, same-model self-refinement, and cross-system LLM-based translation refinement across multiple WMT test sets.
- We use weaker Qwen3 and cross-family GLM drafts to assess initial-quality headroom and comparisons beyond Qwen self-refinement.
- We combine controlled sentence-level analyses with xCOMET-XL score/span diagnostics to characterize quality gains and harmful edits in the evaluated settings.
- We characterize realized translation editability as a draft–editor relation for assessing refinement benefits and risks.

2 Related Work

LLMs for Machine Translation. Recent work has evaluated ChatGPT, GPT-4, PaLM, and related models for machine translation, showing that instruction-following LLMs can produce fluent and semantically adequate translations (Jiao et al., 2023; Hendy et al., 2023; Vilar et al., 2023). Other studies examine prompting, in-context examples, instruction tuning, and multilingual variation in LLM translation (Zhang et al., 2023; Alves et al., 2023; Zhu et al., 2024). These studies mainly ask how to make LLMs better direct translators. We instead ask when LLMs are more useful as post-editors, especially given that LLM outputs may be semantically plausible while differing substantially in surface realization from references.

Automatic Post-Editing. Automatic post-editing (APE) corrects MT outputs using the source sentence and an initial hypothesis, traditionally with source–MT–post-edit training triples (Chatterjee, 2019). WMT APE shared tasks and later neural APE work show that gains depend strongly on the quality and type of the initial MT output (Chatterjee et al., 2018; Vu and Haffari, 2018; Chollampatt et al., 2020). Recent LLM-based APE studies use prompting, feedback, or error annotations to guide the post-editor, while also noting risks such as hallucinated edits (Raunak et al., 2023; Ki and Carpuat, 2024). In contrast, we study the draft source itself as a determinant of post-editing gains.

MT Evaluation Metrics. BLEU and chrF++ measure lexical or character-level overlap with references (Papineni et al., 2002; Popović, 2015, 2017), while neural metrics such as COMET better capture semantic adequacy and correlate more

strongly with human judgments (Rei et al., 2020; Freitag et al., 2022; Rei et al., 2022). xCOMET extends this line of learned metrics with fine-grained error detection and sentence-level scoring (Guerreiro et al., 2024). Prior work also shows that metric-based conclusions can vary with the systems being evaluated (Mathur et al., 2020). We therefore use COMET as the main semantic metric, BLEU/chrF++ as signals of surface stability, and xCOMET-XL as a score-level and span-level robustness check.

Metric Divergence and Draft Editability. For LLM-based translation, lexical and neural metrics may favor different systems. We use this divergence as a starting point for analyzing *translation editability*. In contrast to treating editability as a single draft score, we study it as a draft–post-editor relation: a useful draft should leave semantic headroom and expose errors that the post-editor can identify and correct. This connects to APE observations that post-editing gains depend on the baseline output, but we explicitly compare draft sources and self-refinement in LLM post-editing and treat editability as an explanatory diagnostic perspective.

3 Methodology

3.1 Task Formulation

Given a source sentence x, a draft system M first generates an initial translation:

$$y = M(x), \tag{1}$$

and an LLM post-editor P revises it conditioned on both the source and draft:

$$\hat{y} = P(x, y). \tag{2}$$

We study how the choice of M affects $\hat{y}$. We call this inference-time procedure *LLM-based translation refinement*: the LLM receives both the source and an existing translation and is prompted to improve it. We retain *post-editing* as an operational description of revising a draft. As in supervised APE, our editor conditions on both the source and draft, but we use it without training a dedicated model on source–MT–human-post-edit triples.

3.2 Cross-System Translation Refinement

We distinguish same-model *self-refinement*, within-family cross-model refinement, cross-family LLM refinement, and NMT-to-LLM refinement. Holding Qwen3-8B fixed as editor allows us to compare draft sources under a shared editing configuration. These comparisons diagnose source-dependent outcomes; they do not by themselves isolate a causal effect of model family.

3.3 Translation Editability

We use *translation editability* to describe when a draft translation can be improved reliably by a particular post-editor. This makes editability a relational property of a draft–post-editor pair, rather than an intrinsic property of the draft alone. A draft is editable for a post-editor when it provides semantic headroom, exposes errors in a locally correctable form, and preserves enough lexical or structural anchors to guide correction.

We characterize *realized editability* for a fixed draft–editor pair using four complementary reference-based quantities: expected quality gain, probability of improvement beyond $\epsilon = 0.001$, degradation risk below $-\epsilon$, and final refined quality. For a quality metric s, the gain is

$$\Delta s = s(\hat{y}) - s(y). \tag{3}$$

Final quality is the primary criterion for comparing pipelines; gain describes the conversion of residual error into improvement. Initial COMET, lexical overlap, edit rate, and predicted xCOMET-XL errors provide additional diagnostics. These observed quantities do not constitute a reference-free predictor available before refinement, and low initial quality alone is not sufficient evidence of useful editability.

4 Experimental Setup

4.1 Datasets

We conduct experiments on English-to-Chinese translation using three WMT test sets: WMT17, WMT21, and WMT22 (Bojar et al., 2017; Akhbardeh et al., 2021; Kocmi et al., 2022). These test sets contain 2,001, 1,002, and 2,037 sentence pairs, respectively. We use the official source sentences and reference translations for evaluation. All systems are evaluated on the same test inputs to ensure comparability.

4.2 Draft Translation Systems

We consider two groups of draft translation systems and one closed-source reference baseline.

Sockeye-based NMT system. We use a Sockeye-based Transformer neural MT system as the representative specialized NMT draft generator (Hieber et al., 2018). This system, referred to as Sockeye in the main tables, is the main non-LLM draft source in our core comparisons. We choose it because it achieves the strongest direct-translation performance among our available Sockeye variants, rather than because of its post-edited performance; architectural and training details are provided in Appendix C. We also conduct a source/reference exact-match overlap check between the Sockeye training corpus and the 5,040 WMT test sentences. The scan finds 9 unique overlapping test items (0.18%); all are source–reference pair matches involving short generic expressions such as greeting-like or formulaic sentences. Appendix C describes the procedure, and Appendix H reports an additional Sockeye-based draft system as a supplementary robustness check.

LLM translators. We evaluate Qwen3 models with different parameter sizes as the main LLM translation systems (Yang et al., 2025a). Additional Qwen2.5 direct translation results are reported in Appendix G. Qwen direct translation uses a zero-shot chat prompt asking the model to translate the source sentence into Chinese and output no other information; the exact template is provided in Appendix B. The prompt is wrapped with the tokenizer chat template using a generation prompt and Qwen thinking mode enabled. Since Qwen models may emit an internal thinking segment, we discard the thinking content and evaluate only the final decoded answer. We use no few-shot examples and set the maximum generation length to 32,768 new tokens; implementation-level decoding details are reported in Appendix C. Unlike these chat-based LLMs, the Sockeye model is a standard sequence-to-sequence MT system: it receives only the source sentence and directly outputs the translation.

Additional draft-source controls. We add Qwen3-0.6B and Qwen3-1.7B drafts to examine initial-quality headroom, and GLM-4-9B-Chat[1] drafts to test a cross-family LLM setting. These controls use Qwen3-8B as the fixed editor on the same WMT inputs and with the same refinement prompt and evaluation protocol. Their results are summarized in Section 6.1. Qwen-specific thinking and chat-template settings do not describe the GLM draft generator.

Closed-source API reference baseline. We include a closed-source API baseline under the provider's model label `gpt-5`, referred to as GPT-5 throughout the paper. The model was accessed on March 24, 2026 using the same direct-translation prompt as the open-source LLMs. It serves only as a direct-translation reference; API metadata are summarized in Appendix C.

4.3 LLM-Based Translation Refinement

Given a source sentence and a draft translation, the LLM post-editor is prompted to revise the draft while preserving the meaning of the source sentence. We use Qwen3-8B and Qwen3-14B as post-editors to study the effect of post-editor scale. The main prompt asks the model to improve the draft and return only the revised translation, without additional explanations. The full prompt is provided in Appendix A.

Post-editing is performed sentence by sentence in a one-pass setting. We apply the model tokenizer's chat template with a generation prompt and Qwen thinking mode enabled. Since Qwen3 can generate an internal thinking segment, we discard the thinking content and evaluate only the final decoded translation. Unless otherwise specified, post-editing generation uses temperature 0.2 and a maximum generation length of 32,768 new tokens; we do not use few-shot examples. Appendix C lists the implementation settings used in the scripts.

The core comparisons include same-model Qwen3 self-refinement, within-family Qwen3 cross-model refinement, and Sockeye-to-Qwen3 refinement. The additional GLM-to-Qwen3 setting is cross-family LLM refinement. Appendix D reports a prompt sensitivity analysis with a stronger direct-translation prompt and a more conservative refinement prompt; the alternatives do not improve the main experimental setting.

4.4 Evaluation Metrics

We evaluate translation quality using COMET, BLEU, and chrF++. COMET is used as the main semantic quality metric, while BLEU and chrF++ measure lexical and character-level overlap. BLEU and chrF++ are computed using sacreBLEU (Post, 2018); BLEU uses the Chinese `zh` tokenizer, and chrF++ uses word order 2. COMET scores are computed with the official command

[1]Official model repository: `https://huggingface.co/zai-org/glm-4-9b-chat`.

for `Unbabel/wmt22-comet-da`. We additionally evaluate representative systems with the reference-based xCOMET-XL model as a neural-metric robustness check; the results are reported in Appendix J. We use both scalar xCOMET-XL scores and automatically predicted error spans with severity labels. Following xCOMET-XL terminology, we group major and critical spans as severe errors. These span labels are treated as automatic diagnostic evidence rather than gold human error annotations. Scores, gains, and gaps are rounded independently. Command templates are provided in Appendix E.

For each post-editing system, we report its final score and gain over the corresponding draft, as defined in Equation 3.

4.5 Edit Distance Analysis

To measure how much the post-editor changes a draft translation, we compute the character-level edit distance between the draft output and the post-edited output. For each sentence pair, we define the edit rate as:

$$\mathrm{EditRate}(y, \hat{y}) = \frac{\mathrm{EditDistance}(y, \hat{y})}{\max(|y|, |\hat{y}|)}. \quad (4)$$

For Chinese outputs, we compute edit distance at the character level to avoid dependency on word segmentation. We report the average and median edit rate for representative systems. This analysis helps distinguish conservative post-editing from aggressive rewriting.

4.6 Sentence-Level Statistical Summaries

We report paired two-sided Wilcoxon signed-rank results and bootstrap 95% confidence intervals (CIs) for mean sentence-level COMET differences (Section 5.4). The bootstrap count, CI algorithm, and multiple-comparison treatment are not documented in the available records; these summaries are therefore interpreted per comparison, without claims of familywise error control.

5 Results

5.1 Direct Translation Results

Table 1 reports the core direct-translation results. Sockeye achieves strong BLEU and chrF++ but lower COMET than the stronger LLM translators in this table. This divergence suggests differences in surface overlap and semantic quality. It does not extend to every LLM draft: the additional Qwen3-0.6B and Qwen3-1.7B controls in Section 6.1 have lower initial macro-average COMET than Sockeye.

5.2 Translation Refinement Results

Table 2 shows that refinement benefits draft sources unequally. Sockeye obtains substantial COMET gains across all test sets, whereas the stronger Qwen3 drafts in this core comparison improve less. Section 6.1 qualifies this pattern with weaker Qwen3 and cross-family GLM drafts. Appendix H provides a supplementary check using one additional Sockeye system: its refinement gain is also large, although final quality still depends on the initial draft.

5.3 Effect of Post-Editor Scale

Increasing the post-editor from Qwen3-8B to Qwen3-14B brings only marginal and inconsistent changes. For Sockeye drafts, both post-editors achieve the same average COMET score of 0.8710; for Qwen3-14B drafts, the 14B post-editor improves over the 8B post-editor by only +0.0002 average COMET. Appendix J shows the same trend under xCOMET-XL: Qwen3-8B post-editing is slightly stronger than Qwen3-14B post-editing in our representative xCOMET-XL comparisons. Thus, variation across draft sources is substantially larger than the variation observed between the 8B and 14B post-editors in our evaluated settings. This also suggests that an 8B LLM post-editor applied to an editable NMT draft can be competitive with larger LLM translation or self-refinement baselines in translation quality. Appendix I further reports an exploratory Japanese-to-Chinese FLORES-200 check. In addition to COMET, the FLORES check reports BLEU, chrF++, and xCOMET, and shows the same draft-source pattern: Sockeye+Qwen3-8B obtains a much larger learned-metric gain over its draft than Qwen self-refinement and remains competitive with strong LLM baselines.

5.4 Sentence-Level Outcomes

Table 3 summarizes paired sentence-level COMET differences. We classify $\Delta\mathrm{COMET} > 0.001$ as improved, $\Delta\mathrm{COMET} < -0.001$ as degraded, and other outcomes as nearly unchanged. Sockeye+Qwen3-8B gains +0.0196 on average, with a 95% bootstrap CI of [0.0185, 0.0208]; 74.2% of sentences improve. Qwen3-8B self-refinement has a near-zero mean gain and a CI crossing zero. Refinement of Qwen3-14B drafts yields smaller

System	WMT17			WMT21			WMT22			Avg.
	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
Sockeye	0.8523	42.7	30.4	0.8365	36.3	27.4	0.8634	50.0	38.6	0.8507
Qwen3-8B	0.8683	38.6	25.6	0.8496	31.8	20.4	0.8671	40.8	28.9	0.8617
Qwen3-14B	0.8714	40.0	26.6	0.8546	32.4	20.9	0.8722	41.8	30.1	0.8661
Qwen3-32B	0.8715	40.3	26.7	0.8545	32.1	20.7	0.8712	41.7	29.8	0.8657
GPT-5	0.8779	42.2	28.7	0.8573	33.3	22.2	0.8801	41.8	31.8	0.8718

Table 1: Direct translation results on WMT17, WMT21, and WMT22. GPT-5 is included as a closed-source reference baseline. Avg. denotes the unweighted average of corpus-level COMET scores over WMT17, WMT21, and WMT22.

Draft	Post-Editor	WMT17		WMT21		WMT22		Avg.	
		COMET	Δ	COMET	Δ	COMET	Δ	COMET	Δ
Sockeye	–	0.8523	–	0.8365	–	0.8634	–	0.8507	–
Sockeye	Qwen3-8B	0.8756	+0.0233	0.8600	+0.0235	0.8775	+0.0141	0.8710	+0.0203
Sockeye	Qwen3-14B	0.8741	+0.0218	0.8596	+0.0231	0.8794	+0.0160	0.8710	+0.0203
Qwen3-8B	–	0.8683	–	0.8496	–	0.8671	–	0.8617	–
Qwen3-8B	Qwen3-8B	0.8685	+0.0002	0.8491	-0.0005	0.8679	+0.0008	0.8618	+0.0001
Qwen3-14B	–	0.8714	–	0.8546	–	0.8722	–	0.8661	–
Qwen3-14B	Qwen3-8B	0.8742	+0.0028	0.8575	+0.0029	0.8721	-0.0001	0.8679	+0.0019
Qwen3-14B	Qwen3-14B	0.8743	+0.0029	0.8564	+0.0018	0.8735	+0.0013	0.8681	+0.0020
Qwen3-32B	–	0.8715	–	0.8545	–	0.8712	–	0.8657	–
Qwen3-32B	Qwen3-8B	0.8747	+0.0032	0.8591	+0.0046	0.8734	+0.0022	0.8691	+0.0033

Table 2: COMET scores and post-editing gains for different draft sources. Δ denotes the COMET difference between the post-edited output and its corresponding draft. Avg. COMET and Avg. Δ are unweighted averages over WMT17, WMT21, and WMT22; sentence-level analyses in Table 3 are computed over pooled sentence-level differences.

gains. Sockeye+Qwen3-8B also has higher pooled mean COMET than Qwen3-14B direct translation and self-refinement.

5.5 Human Preference Evaluation

We report aggregate results from a stratified blind pairwise evaluation with two bilingual annotators over WMT17, WMT21, and WMT22. Each comparison contains 120 judgments, with the A/B order mirrored across annotators. Sockeye+Qwen3-8B receives 83 preferences against 23 for the original Sockeye draft, with 14 comparable judgments. Against Qwen3-14B self-refinement, it receives 61 preferences against 39, with 20 comparable judgments. Excluding ties, these correspond to 78.3% and 61.0%, respectively. These aggregate counts provide descriptive support rather than evidence of statistical superiority; the evaluation protocol, results, and record limitations are reported in Appendix F.

5.6 Comparison with a Closed-Source Baseline

The GPT-5 direct-translation baseline obtains 0.8718 average COMET, while Sockeye+Qwen3-8B and Sockeye+Qwen3-14B both obtain 0.8710 and higher average BLEU/chrF++ (41.9/28.9 and 42.4/29.4 vs. 39.1/27.6).

6 Analysis

6.1 Controls for Initial Quality and Draft Family

With Qwen3-8B fixed as editor, Table 4 compares drafts spanning initial quality and model family. Qwen3-0.6B has the lowest initial COMET and the largest gain, confirming that headroom matters; however, its final quality remains below the other refined systems. Qwen3-1.7B starts close to Sockeye but achieves a smaller gain and lower final quality. Thus, a large raw gain alone is not evidence of a better pipeline.

A common-support comparison with Qwen3-1.7B retains 94.9%–96.8% of sentences and shows larger Sockeye gains on all three test sets; an ex-

Comparison	Mean Δ	95% CI	Imp.	Unch.	Deg.	p
Sockeye+Qwen3-8B vs. Sockeye	+0.0196	[0.0185, 0.0208]	74.2	4.8	21.0	< 0.001
Qwen3-8B+Qwen3-8B vs. Qwen3-8B	+0.0003	[-0.0009, 0.0015]	47.5	7.4	45.1	0.279
Qwen3-14B+Qwen3-8B vs. Qwen3-14B	+0.0016	[0.0009, 0.0023]	42.2	24.0	33.8	< 0.001
Qwen3-14B+Qwen3-14B vs. Qwen3-14B	+0.0020	[0.0013, 0.0027]	41.3	28.0	30.7	< 0.001
Sockeye+Qwen3-8B vs. Qwen3-14B	+0.0049	[0.0038, 0.0060]	54.0	6.5	39.5	< 0.001
Sockeye+Qwen3-8B vs. Qwen3-14B+Qwen3-14B	+0.0029	[0.0018, 0.0040]	51.6	6.9	41.5	< 0.001

Table 3: Sentence-level COMET summaries over all test sets: mean paired differences, bootstrap 95% CIs, and reported two-sided Wilcoxon signed-rank p-values. Multiple-comparison treatment is unverified. Imp., Unch., and Deg. denote the percentages with $\Delta\text{COMET} > 0.001$, $[-0.001, 0.001]$, and < -0.001, respectively.

Draft	Initial	Final	Gain
Sockeye	0.8507	**0.8710**	+0.0203
Qwen3-0.6B	0.8138	0.8520	**+0.0381**
Qwen3-1.7B	0.8474	0.8594	+0.0121
GLM-4-9B-Chat	0.8582	0.8640	+0.0058

Table 4: Additional draft-source controls with Qwen3-8B as the fixed editor. Values are unweighted macro-averages of COMET over WMT17, WMT21, and WMT22, independently rounded to four decimals.

ploratory regression summary has the same direction (Appendix P). These descriptive controls are consistent with a draft-source association beyond initial-quality headroom, without establishing causality. GLM drafts also improve consistently, showing that beneficial refinement is not confined to NMT drafts. Nevertheless, refined Sockeye outputs have higher final COMET than refined GLM outputs on each test set. The contrast therefore extends beyond Qwen self-refinement.

6.2 Semantic Headroom Must Be Convertible

A natural concern is that the larger gains for Sockeye drafts may simply reflect their lower initial COMET scores. We therefore compare Sockeye post-editing and Qwen3-8B self-refinement within the same globally defined sentence-level buckets. The buckets use shared thresholds over initial COMET and sentence-level chrF++; thus, "low COMET" denotes relative semantic headroom under the same evaluator and test distribution, not an absolute quality label.

Table 5 shows different refinement outcomes within shared initial-quality ranges. In the two low-COMET buckets, Sockeye post-editing obtains large gains (+0.0306 and +0.0349), while Qwen self-refinement obtains only small gains (+0.0044 and +0.0069). The difference is also visible in stability: in the high-chrF++/low-COMET bucket, Sockeye post-editing improves 83.6% of sentences and degrades 14.7%, whereas Qwen self-refinement improves 55.7% and degrades 39.1%. The high-COMET buckets further show that self-refinement can be harmful even when the draft is already strong: Qwen self-refinement has negative mean ΔCOMET in both high-COMET buckets, while Sockeye post-editing remains positive. These results suggest that LLM post-editing requires *convertible semantic headroom*: the draft must not only leave room for improvement, but expose errors in a form that the post-editor can reliably repair. Appendix L reports descriptive gaps and a consistent positive-gap pattern using sentence-level BLEU.

We further repeat the key bucket analysis with xCOMET-XL scores as a robustness check. The pattern becomes even clearer: Sockeye+Qwen3-8B obtains a pooled mean gain of +0.0615 xCOMET-XL, improving 78.6% of sentences and degrading only 14.5%, whereas Qwen3-8B self-refinement obtains only +0.0028, with 47.3% improved and 41.0% degraded. In the high-chrF++/low-xCOMET bucket, which directly instantiates the *lexically stable but semantically improvable* regime, Sockeye post-editing improves by +0.1047, compared with +0.0243 for Qwen self-refinement. The full xCOMET-XL bucket table is provided in Appendix J.

Importantly, these results should not be read as showing that high lexical overlap is a necessary condition for editability. Lexical and structural anchors provide one productive mode of editability, especially when they are paired with correctable semantic gaps. However, editability ultimately depends on whether the remaining semantic headroom is locally correctable by the post-editor. This interpretation is also supported by the xCOMET-XL bucket analysis, where low-chrF++/low-xCOMET drafts improve substantially as well.

Bucket	Qwen n	Sock. n	Qwen Δ	Qwen I/D	Sock. Δ	Sock. I/D	Gap
high chrF++ / low COMET	524	1017	+0.0044	55.7/39.1	+0.0306	83.6/14.7	+0.0263
low chrF++ / low COMET	1879	1613	+0.0069	53.4/42.6	+0.0349	80.3/16.4	+0.0280
high chrF++ / high COMET	1680	1819	-0.0065	40.1/49.4	+0.0054	67.2/25.9	+0.0119
low chrF++ / high COMET	957	591	-0.0029	44.6/45.7	+0.0028	62.9/29.3	+0.0057

Table 5: Shared-threshold bucket analysis using global thresholds for initial COMET and sentence-level chrF++. Qwen denotes Qwen3-8B self-refinement; Sock. denotes Sockeye+Qwen3-8B. The bucket sizes differ because the same thresholds are applied to both systems rather than being fitted separately for each system. I/D reports the percentages of improved and degraded sentences; the omitted percentage is nearly unchanged. Gap is Sockeye+Qwen3-8B minus Qwen3-8B self-refinement.

6.3 xCOMET-XL Span Evidence for Headroom Conversion

The xCOMET-XL span analysis characterizes predicted error removal and harmful edits. Sockeye+Qwen3-8B reduces total predicted errors by 42.0% and severe errors by 31.4%, versus 5.5% and 4.7% for Qwen3-8B self-refinement. Its mean ΔxCOMET is larger (+0.0615 vs. +0.0028), its severe-error reduction is greater (Δsevere $=$ -0.490 vs. -0.061), and fewer sentences lose xCOMET while adding severe errors (7.9% vs. 19.2%). Appendix J reports bootstrap CIs, span tables, and examples.

6.4 Additional Diagnostics and Synthesis

Editing magnitude does not track gains: Sockeye+Qwen3-8B edits moderately (Avg. ER 0.262) and gains +0.0203 COMET, whereas Qwen3-8B self-refinement edits more (0.306) but gains +0.0001. Controlling for initial COMET, chrF++, edit rate, and length ratio yields a positive Sockeye-draft coefficient (+0.0157) in the descriptive regression (Appendices M and K). Qualitative examples show localized repairs and harmful adequacy shifts (Appendices N and J.3). Together, these diagnostics motivate editability as a draft–editor relation rather than a single draft metric.

7 Conclusion

This paper used LLM-based translation refinement to diagnose how draft source affects quality. With Qwen3-8B fixed as editor, Sockeye drafts improve more than the stronger Qwen3 drafts and the cross-family GLM drafts. The weaker Qwen3-0.6B drafts instead obtain the largest raw gain but lower final quality, confirming the need to assess gains jointly with final quality and degradation risk. Descriptive common-support, regression, bucket, and xCOMET-XL analyses suggest that initial quality alone is an incomplete account of the observed differences. These findings characterize realized editability as conditional on a draft–editor pairing.

Limitations

Our main experiments focus on English-to-Chinese translation. Their consistency across WMT test sets does not establish generalization to other language pairs, particularly low-resource languages or those with different morphology and word order. The exploratory Japanese-to-Chinese FLORES-200 check in Appendix I covers one test set, with COMET, BLEU, chrF++, and xCOMET but no human evaluation or span-level diagnostics.

We study inference-time LLM post-editing, not task-specific training of a dedicated APE model. Our conclusions rely mainly on automatic metrics, which may encode their own biases; xCOMET-XL spans are predicted errors, not gold annotations. The two-annotator human evaluation provides descriptive support, but unavailable item-level records prevent dependence-aware inference and agreement analysis. Larger multi-annotator evaluations are needed to verify adequacy, fluency, and error-type improvements.

GPT-5 serves only as a closed-source direct-translation reference and is not reproducible in the same way as the open-source systems. Refinement uses Sockeye, Qwen3, and GLM drafts with Qwen3 editors. These pairings do not establish an intrinsic NMT advantage or generalization across editor families.

The editor-size comparison covers 8B and 14B models, and the prompt check in Appendix D is limited. Model sizes, prompts, and decoding settings are not exhaustively evaluated. The timing and memory measurements in Appendix Q describe one hardware configuration, not a complete deployment benchmark or GPU-hour and monetary-cost analysis.

Some statistical specifications could not be recovered; the reported CIs have not been recomputed, and the controls are descriptive rather than causal. Realized editability is assessed after refinement with references, not by a reference-free predictor of benefits before refinement. Prediction and selective refinement remain open questions.

Ethical Considerations

We use publicly available MT evaluation datasets and do not collect new user-generated text. Two adult bilingual annotators evaluated anonymized translations and were informed that their judgments would be used for research. No sensitive personal data were collected, and no personally identifying information is reported. Recruitment and evaluation procedures are described in Appendix F.

Automatic post-editing can introduce adequacy, terminology, and style errors even when the output remains fluent. Our results diagnose draft-source effects under the evaluated settings and should not be treated as a deployment recommendation for high-stakes translation, where human review remains necessary.

Acknowledgments

AI assistance disclosure. ChatGPT assisted with data-processing scripts, manuscript organization and editing, and preparation of the English glosses and Pinyin transliterations. The authors conducted the experiments and are responsible for verifying all results, references, interpretations, and final text.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussà, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher M. Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, and 17 others. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Duarte M. Alves, Nuno M. Guerreiro, João Alves, José Pombal, Ricardo Rei, José G. C. de Souza, Pierre Colombo, and André F. T. Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Rajen Chatterjee. 2019. *Automatic Post-Editing for Machine Translation*. Ph.D. thesis, University of Trento.

Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018. Findings of the WMT 2018 shared task on automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 710–725, Belgium, Brussels. Association for Computational Linguistics.

Shamil Chollampatt, Raymond Hendy Susanto, Liling Tan, and Ewa Szymanska. 2020. Can automatic post-editing improve NMT? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2736–2746, Online. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The Sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 200–207, Boston, MA. Association for Machine Translation in the Americas.

Wenxiang Jiao, 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*.

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

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, Maja Popović, and Mariya Shmatova. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 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. 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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Vikas Raunak, 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. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, 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. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Thuy-Trang Vu and Gholamreza Haffari. 2018. Automatic post-editing of machine translation: A neural programmer-interpreter approach. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3048–3053, Brussels, Belgium. Association for Computational Linguistics.

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. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

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

A Post-Editing Prompt

We use the following prompt template for English-to-Chinese WMT post-editing across all draft sources and post-editors. The same language-neutral template is used for the Japanese-to-Chinese evaluation in Appendix I, with the corresponding source sentence and draft. The original Chinese prompt is followed by Pinyin and an English translation for readability; these reading aids are not supplied to the model. Pinyin omits tone marks.

> 原文：{source}
>
> 译文：{draft}
>
> 请优化译文，且只输出优化后的翻译结果，不输出任何其他信息。

Pinyin transliteration:

> Yuanwen: {source}
>
> Yiwen: {draft}
>
> Qing youhua yiwen, qie zhi shuchu youhua hou de fanyi jieguo, bu shuchu renhe qita xinxi.

English translation:

> Source: {source}
>
> Translation: {draft}
>
> Please improve the translation, and output only the improved translation result without any other information.

In implementation, the prompt is wrapped with the tokenizer chat template using `add_generation_prompt=True`. Qwen thinking mode is enabled, and only the final answer after the thinking segment is saved as the post-edited translation.

B Direct Translation Prompt

For Qwen direct translation on the English-to-Chinese WMT sets, we use the following zero-shot user prompt, reported exactly as used in the script. The source sentence is appended after the newline. The prompt is wrapped with the tokenizer chat template using `add_generation_prompt=True` and `enable_thinking=True`; only the final decoded answer after the thinking segment is evaluated.

> please translate this english text to Chinese,and don't output other information:
>
> {source}

C Implementation and Data-Overlap Details

Qwen inference. Table 6 lists the Qwen inference settings. Parameters not explicitly set in the script, such as `top_p`, `top_k`, and sampling flags, follow the inference library behavior. Only the final decoded answer after the thinking segment is evaluated.

Sockeye system details. The main Sockeye system uses Sockeye-trained BPE segmentation and a 6-layer encoder–decoder Transformer with hidden size 1024, feed-forward dimension 4096, and 16 attention heads. It is fine-tuned from a pretrained checkpoint using Adam with inverse square-root learning-rate scheduling, 4,000 warmup steps, dropout 0.1, label smoothing 0.1, a max-word batch size of 22,000, and a maximum sequence length of 95. The training data is mainly collected from OPUS (Tiedemann, 2012) and contains 275,916,636 sentence pairs, with 5,903,943,245 source-side tokens and 5,302,679,717 target-side tokens. During data preparation, 1,269,039 examples are filtered out because they do not fit the length and bucketing constraints.

Item	Setting
Backend	ModelScope `AutoModelForCausalLM` / `AutoTokenizer`
Model loading	`torch_dtype=auto`, `device_map=auto`
Chat template	`add_generation_prompt=True`
Qwen thinking mode	`enable_thinking=True`
Post-editing temperature	0.2
Maximum generation length	32,768 new tokens
Few-shot examples	None
Thinking handling	discard thinking segment; evaluate final answer only
Unspecified decoding parameters	not explicitly set in the script

Table 6: Qwen generation settings used in the implementation scripts.

Computational budget. We evaluate fixed translation and refinement systems without training new LLMs. A complete GPU-hour budget is not reported for the shared infrastructure; supplementary timing and memory measurements appear in Appendix Q.

Closed-source API reference baseline. The GPT-5 reference baseline was accessed on March 24, 2026 using the same direct-translation prompt as the open-source LLMs. The service metadata available at collection time listed OpenAI as the provider, `gpt-5` as the API model label, `aihubmix` as the service-side model family, a release date of 2025-08-07, a 400K-token context window, and a maximum output length of 128K tokens. The endpoint is closed-source and provider-dependent and serves only as a direct-translation reference.

Artifact licenses and terms of use. We use publicly available datasets, models, and evaluation tools for research evaluation, including WMT test sets, FLORES-200, Sockeye, Qwen and GLM models, sacreBLEU, COMET, and xCOMET. We cite the creators of these artifacts and use them according to their public licenses, access conditions, or API terms. The GPT-5 baseline is accessed under the API service terms. We do not redistribute restricted benchmark data or closed-source model artifacts in this paper.

Training/test overlap check. To reduce leakage concerns for the Sockeye system, we perform exact-match overlap checking between the Sockeye training corpus snapshot and the WMT17/WMT21/WMT22 test sets. We collect all WMT English sources and Chinese references, then scan the English and Chinese sides of each Sockeye training pair. A source overlap is recorded if the training English side exactly matches any WMT test source. A reference overlap is recorded if the training Chinese side exactly matches any WMT reference. If both sides match the same WMT test item, we treat it as a stronger source–reference pair overlap risk. The scan finds 9 unique affected WMT test items out of 5,040 (0.18%). All 9 are source–reference pair overlaps: the training English side matches the WMT source and the training Chinese side matches the corresponding WMT reference for the same test item. These items are short generic expressions, so we treat the overlap risk as limited and unlikely to drive the reported system-level trends.

D Prompt Sensitivity Analysis

We conduct a prompt sensitivity analysis to test whether the main results depend on the particular prompt templates used in Tables 1 and 2. We consider two alternatives: a stronger professional direct-translation prompt and a conservative post-editing prompt that asks the model to revise only when the draft contains mistranslations, omissions, terminology errors, or unnatural Chinese expressions.

The alternative prompts are:

> **Professional direct prompt.** You are a professional English-to-Chinese translator. Translate the following English sentence into accurate, fluent, and natural Chinese. Preserve all named entities, numbers, terminology, and factual meaning. Do not add explanations. Output only the Chinese translation.
>
> English: {source}
>
> Chinese:
>
> **Conservative post-editing prompt.** Source: {source}. Draft translation: {draft}. Revise the draft translation only when it contains mistranslations, omissions, terminology errors, or unnatural Chinese expressions. If the draft is already correct, keep it as unchanged as possible. Do not provide explanations. Output only the final translation.

For LLM self-refinement, the conservative post-editing prompt is applied to the fixed direct translations generated under the original direct-translation setting. Therefore, self-refinement gains in Table 8 are computed relative to the original direct draft, not relative to the newly generated direct translation. All averages in this section are unweighted averages over WMT17, WMT21, and WMT22.

Setting	Main	Alt.	Alt.–Main
Qwen3-8B direct	0.8617	0.8560	-0.0056
Qwen3-14B direct	0.8661	0.8597	-0.0064
Qwen3-8B self-refinement	0.8618	0.8621	+0.0003
Qwen3-14B self-refinement	0.8681	0.8669	-0.0012
Sockeye+Qwen3-8B	0.8710	0.8656	-0.0054

Table 7: Prompt sensitivity analysis using unweighted average COMET over WMT17, WMT21, and WMT22. Alternative direct rows use a professional direct-translation prompt; alternative post-editing rows use the conservative post-editing prompt on fixed drafts generated under the original setting.

Post-editing comparison	Conservative
Sockeye → Sockeye+Qwen3-8B	+0.0149
Qwen3-8B self-refinement on fixed draft	+0.0004
Qwen3-14B self-refinement on fixed draft	+0.0008

Table 8: Post-editing gains under the conservative prompt. For LLM self-refinement, gains are computed relative to the original direct draft, because the conservative prompt is applied to fixed original direct translations. The main-setting results are reported in Table 2.

The stronger direct-translation prompt does not improve direct translation: Qwen3-8B decreases from 0.8617 to 0.8560 average COMET, and Qwen3-14B decreases from 0.8661 to 0.8597. The conservative post-editing prompt also does not improve the main Sockeye-draft setting: Sockeye+Qwen3-8B decreases from 0.8710 under the main prompt to 0.8656 under the conservative prompt, although the conservative setting still yields a positive gain over the Sockeye draft. On fixed LLM direct drafts, the conservative prompt yields only marginal self-refinement gains (+0.0004 for Qwen3-8B and +0.0008 for Qwen3-14B). Thus, prompt choice affects the magnitude of realized gains, but the alternative prompts do not overturn the draft-source effect or provide a better main setting.

E Metric Command Templates

We use the following command templates for automatic evaluation. File names are placeholders for the corresponding source, hypothesis, and reference files.

```
sacrebleu ref.zh \
  -i hyp.zh -m bleu chrf \
  --chrf-word-order 2 \
  --tokenize zh

comet-score \
  -s src.en -t hyp.zh \
  -r ref.zh \
  --model \
  Unbabel/wmt22-comet-da
```

xCOMET-XL scores and predicted error spans are analyzed in Appendix J.

F Human Preference Evaluation Details

We report aggregate results from a blind pairwise human preference evaluation with two bilingual annotators. The evaluation covers two comparisons: Sockeye vs. Sockeye+Qwen3-8B, and Sockeye+Qwen3-8B vs. Qwen3-14B+Qwen3-14B. Each comparison contains 120 judgments over examples sampled from WMT17, WMT21, and WMT22. The systems are anonymized as two candidate translations, and the A/B order is mirrored across annotators. Annotators are instructed to choose the better translation in terms of overall adequacy and fluency, or to choose comparable quality if the two outputs are similar.

Recruitment, compensation, and consent. The annotators are two bilingual participants recruited by the authors for this small-scale research evaluation; no crowdsourcing platform is used. No monetary compensation is provided. Annotators were informed that their anonymized pairwise preference judgments would be used for research evaluation, and no personally identifying information is reported.

Data privacy and anonymization. We use existing public machine translation benchmarks and do not collect new user-generated text. The human-evaluation records contain only anonymized pairwise preference judgments over system outputs. System identities are hidden from annotators, and annotator names or personally identifying information are not included in the analyzed data. We do not perform additional offensive-content filtering

beyond the original benchmark construction, and we use the data only for research evaluation.

Comparison	Sockeye+8B	Other	Comp.
vs. Sockeye	83	23	14
vs. Qwen3-14B+14B	61	39	20

Table 9: Aggregate preferences, with 120 judgments per comparison. Sockeye+8B counts preferences for Sockeye+Qwen3-8B; Other denotes Sockeye in the first row and Qwen3-14B self-refinement in the second. Comp. denotes comparable outputs. Counts are descriptive; no significance test is reported.

The counts favor Sockeye+Qwen3-8B in both comparisons, more strongly against the original Sockeye draft. They correspond to 78.3% and 61.0% of non-tie judgments. The original item-level records are unavailable, so the number of unique evaluated sentences, sentence-level dependence, and inter-annotator agreement cannot be reconstructed from these totals. We therefore do not report significance tests or treat the aggregate preference percentages as a definitive system ranking.

G Additional Direct Translation Results

We also evaluate Qwen2.5 models as auxiliary direct-translation baselines (Yang et al., 2025b); they are not used in the post-editing analyses. As shown in Table 10, larger Qwen2.5 models generally obtain higher COMET scores, but they still follow the same overall pattern observed in the main results: LLM translators tend to achieve stronger COMET scores than the Sockeye draft, while their BLEU and chrF++ scores remain lower than the strongest lexical-overlap systems.

H Additional Sockeye Draft Results

We evaluate an additional Sockeye-based draft system to check whether the post-editing pattern extends beyond the main checkpoint described in Section 4.2.

The additional Sockeye system is trained on 112,089,861 sentence pairs, with 2,374,552,019 source-side tokens and 2,233,460,469 target-side tokens. During data preparation, 724,524 examples are filtered out because they do not fit the length and bucketing constraints.

As shown in Table 11, Qwen3-8B post-editing improves the additional Sockeye draft from 0.8453 to 0.8678 average COMET, corresponding to an average ΔCOMET of +0.0225. The gain is comparable in magnitude to the main Sockeye result in Table 2, although the final post-edited quality remains slightly lower. This suggests that the editability pattern is not limited to a single Sockeye checkpoint, while also showing that the initial draft quality still affects the final post-edited output.

I Exploratory FLORES-200 Japanese-to-Chinese Check

To probe whether the draft-source pattern is specific to English-to-Chinese WMT test sets, we conduct an exploratory FLORES-200 Japanese-to-Chinese evaluation on 1,012 devtest examples (NLLB Team et al., 2022). We reuse the English-to-Chinese experimental pipeline, replacing the evaluation inputs with Japanese sources and their Chinese references. The translation direction changes from English-to-Chinese to Japanese-to-Chinese; all other experimental conditions, including model configurations, decoding settings, and evaluation procedures, remain unchanged.

For direct translation, the source-language name in the prompt in Appendix B is changed from `english` to `japanese`:

> please translate this japanese text to Chinese,and don't output other information:
>
> {source}

The post-editing prompt in Appendix A is unchanged because it does not explicitly name the source language. It receives the Japanese source sentence and the Chinese draft. We report COMET, BLEU, chrF++, and xCOMET for this check; no human evaluation or span-level diagnostics are included.

The FLORES-200 results are consistent with the main draft-source pattern. Sockeye direct translation has strong lexical-overlap scores, including the highest chrF++ score (26.7435), but the lowest xCOMET score (0.8096). After Qwen3-8B post-editing, xCOMET rises substantially to 0.8523 and COMET rises from 0.8634 to 0.8805. The resulting system achieves the highest BLEU among the evaluated systems and remains close to Qwen3-32B direct translation under xCOMET (0.8523 vs. 0.8536). By contrast, Qwen3-8B and Qwen3-14B self-refinement obtain smaller xCOMET gains (+0.0138 and +0.0087). The decrease in chrF++ after Sockeye post-editing suggests that the post-editor does not merely preserve reference-overlapping surface forms; rather, it uses

System	WMT17			WMT21			WMT22			Avg.
	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
Qwen2.5-7B	0.8590	37.1	24.9	0.8403	30.2	20.4	0.8619	39.9	28.3	0.8537
Qwen2.5-14B	0.8638	37.5	25.6	0.8469	31.8	20.4	0.8639	40.4	29.4	0.8582
Qwen2.5-32B	0.8696	40.6	26.8	0.8500	32.7	20.9	0.8706	43.7	30.8	0.8634

Table 10: Additional direct translation results for Qwen2.5 models. These systems are included as auxiliary baselines and are not used in the main post-editing analysis.

Panel A: COMET by test set			
Dataset	**Additional Sockeye**	**+Qwen3-8B**	**ΔCOMET**
WMT17	0.8567	0.8725	+0.0158
WMT21	0.8332	0.8570	+0.0238
WMT22	0.8459	0.8739	+0.0280
Macro-average	0.8453	0.8678	+0.0225

Panel B: Macro-average surface-overlap scores		
Metric	**Additional Sockeye**	**+Qwen3-8B**
BLEU	40.0	40.5
chrF++	29.3	27.0

Table 11: Additional Sockeye-based draft results. +Qwen3-8B denotes post-editing the additional Sockeye draft with Qwen3-8B. Macro-averages weight WMT17, WMT21, and WMT22 equally; ΔCOMET is the gain over the corresponding draft. The additional system is a supplementary robustness check and is not used in the core comparisons.

System	COMET	BLEU	chrF++	xCOMET
GPT-5 API	**0.8852**	35.2505	22.8987	**0.8680**
Sockeye+Qwen3-8B	0.8805	**35.8641**	23.4265	0.8523
Qwen3-14B+Qwen3-14B	0.8773	33.4761	21.8195	0.8466
Qwen3-32B	0.8760	32.0428	20.6455	0.8536
Qwen3-8B+Qwen3-8B	0.8751	31.9855	20.8799	0.8385
Qwen3-14B	0.8750	33.2668	21.6989	0.8379
Qwen3-8B	0.8712	31.8285	20.9143	0.8247
Sockeye	0.8634	35.6746	**26.7435**	0.8096

Table 12: Exploratory FLORES-200 Japanese-to-Chinese results on 1,012 devtest examples. xCOMET is used as an additional learned-metric check; higher scores are better for COMET and xCOMET. The table is ordered by COMET.

Base	ΔCOMET	ΔBLEU	ΔchrF++	ΔxCOMET
Sockeye → Sockeye+Qwen3-8B	+0.0171	+0.1895	-3.3170	+0.0427
Qwen3-8B self-refinement	+0.0039	+0.1570	-0.0344	+0.0138
Qwen3-14B self-refinement	+0.0023	+0.2093	+0.1206	+0.0087

Table 13: FLORES-200 Japanese-to-Chinese post-editing gains. Sockeye drafts receive the largest COMET and xCOMET gains, while LLM self-refinement yields smaller learned-metric gains.

the draft as an anchor while making changes that substantially improve learned semantic quality. We therefore interpret this experiment as preliminary evidence that the headroom-conversion pattern is not limited to English-to-Chinese, while keeping it separate from the main evaluation because it covers only one FLORES language direction and one test set.

J xCOMET-XL Robustness and Span Diagnostics

To test whether the observed post-editing pattern is specific to COMET, we additionally evaluate representative systems with the reference-based xCOMET-XL model (Guerreiro et al., 2024). Absolute xCOMET-XL values should not be compared directly with COMET because they come from different learned metrics. Table 14 reports full scalar xCOMET-XL scores and their corresponding average gains. System-level averages in Table 14 are computed by averaging WMT17, WMT21, and WMT22 scores, whereas the sentence-level and bucket analyses in Table 15 report pooled sentence-level means over examples with available sentence-level xCOMET logs. We also analyze predicted error spans and severity labels as automatic diagnostics of error reduction and harmful rewrites. Paired bootstrap CIs and their documentation limits are reported in Section J.1.

The xCOMET-XL bucket analysis provides further metric-level evidence for the lexically stable but semantically improvable regime. In the high-chrF++/low-xCOMET bucket, Sockeye post-editing obtains a much larger gain than Qwen self-refinement (+0.1047 vs. +0.0243). However, low-chrF++/low-xCOMET drafts also improve substantially, so lexical overlap should be interpreted as a useful anchor signal rather than a necessary condition for editability.

The xCOMET-XL results reinforce the main COMET-based findings. Sockeye drafts receive large gains after Qwen post-editing, whereas Qwen self-refinement yields much smaller improvements. Notably, Sockeye+Qwen3-8B obtains the highest average xCOMET-XL score (0.8663), slightly above Sockeye+Qwen3-14B (0.8652) and Qwen3-14B+Qwen3-8B (0.8633). These results support two conclusions: the draft-source pattern is not specific to COMET, and increasing the post-editor from 8B to 14B does not guarantee better post-editing quality.

J.1 xCOMET-XL Bootstrap Intervals

Table 17 reports the original bootstrap CIs for Sockeye post-editing and Qwen3-8B self-refinement. The method note specifies paired resampling by dataset and segment ID, but incompletely documents the CI algorithm and bootstrap p-value calculation. The intervals are reported without recomputation, and the p-values are omitted.

J.2 xCOMET-XL Severity-Change Details

Table 18 reports severity-specific changes for the two most important comparisons. Sockeye post-editing reduces minor, major, and critical predicted errors much more strongly than Qwen3-8B self-refinement. In particular, Sockeye+Qwen3-8B reduces major errors by 30.7% and critical errors by 56.8%, whereas Qwen3-8B self-refinement reduces major errors by only 4.4%.

System	W17	W21	W22	Avg.	Δ
Sockeye	0.8010	0.7502	0.8528	0.8013	–
S+8B	0.8690	0.8318	0.8980	**0.8663**	+0.0650
S+14B	0.8694	0.8281	0.8980	0.8652	+0.0639
Q8	0.8513	0.8041	0.8802	0.8452	–
Q8+8B	0.8548	0.7997	0.8857	0.8467	+0.0015
Q14	0.8606	0.8124	0.8877	0.8536	–
Q14+8B	0.8716	0.8239	0.8945	0.8633	+0.0097
Q14+14B	0.8701	0.8212	0.8944	0.8619	+0.0083

Table 14: xCOMET-XL robustness results for representative systems. S denotes Sockeye; Q8 and Q14 denote Qwen3-8B and Qwen3-14B drafts; +8B/+14B denote the Qwen post-editor. Avg. Δ is computed relative to the corresponding draft system; averages are unweighted over WMT17, WMT21, and WMT22.

Panel A: Pooled sentence-level xCOMET gains			
Comparison	**Mean Δ**	**Imp.**	**Deg.**
Sockeye → S+8B	+0.0615	78.6%	14.5%
Q8 self-refinement	+0.0028	47.3%	41.0%
Panel B: xCOMET bucket analysis			
Bucket	**Q8 self**	**S+8B**	**Gap**
high chrF++ / low xCOMET	+0.0243	+0.1047	+0.0804
low chrF++ / low xCOMET	+0.0204	+0.0950	+0.0746
high chrF++ / high xCOMET	-0.0128	+0.0164	+0.0293
low chrF++ / high xCOMET	-0.0125	+0.0134	+0.0258

Table 15: Pooled xCOMET-XL gain and bucket summaries. Panel B uses initial xCOMET-XL and sentence-level chrF++; exact thresholds and per-bucket counts are unavailable, limiting reproducibility.

Comparison	ΔxCOMET	Err. red.	Sev. red.	Imp./Deg.	Sev. add.	Deg.+Sev. add.
Sockeye → Sockeye+8B	+0.0615	42.0	31.4	78.6/14.5	13.1	7.9
Sockeye → Sockeye+14B	+0.0609	40.3	29.9	78.7/15.0	12.6	7.9
Qwen3-8B self-refinement	+0.0028	5.5	4.7	47.3/41.0	20.5	19.2
Qwen3-14B → Qwen3-8B PE	+0.0094	6.9	8.4	43.5/27.4	11.5	10.2
Qwen3-14B self-refinement	+0.0082	5.9	7.4	40.3/25.6	9.0	8.1

Table 16: xCOMET-XL span-level error changes. Err. red. and Sev. red. are percentage reductions in total predicted errors and severe errors (major+critical), respectively. Sev. add. is the percentage of sentences where new severe errors are added after post-editing. Deg.+Sev. add. is the percentage of sentences whose xCOMET-XL score decreases while severe errors are added. The span labels are automatic diagnostics, not gold human annotations.

Metric	Sock.	S 95% CI	Q8 self	Q 95% CI	Gap
ΔxCOMET	+0.0615	[+0.0590,+0.0639]	+0.0028	[+0.0006,+0.0049]	+0.0587
Δsevere	-0.490	[-0.534,-0.448]	-0.061	[-0.097,-0.024]	-0.429
Severe added	13.1	[12.1,14.0]	20.5	[19.4,21.6]	-7.4
Deg. + severe added	7.9	[7.2,8.7]	19.2	[18.1,20.3]	-11.3

Table 17: Bootstrap 95% CIs for xCOMET-XL diagnostics. Sock. denotes Sockeye → Sockeye+Qwen3-8B; Q8 self denotes Qwen3-8B self-refinement. Gap is Sock. minus Q8 self. The last two rows are percentages, with gaps in percentage points.

Comparison	Severity	Initial	Edited	Δ	Reduction
Sockeye → Sockeye+8B	Minor	7858	3715	-4143	52.7
Sockeye → Sockeye+8B	Major	7674	5318	-2356	30.7
Sockeye → Sockeye+8B	Critical	199	86	-113	56.8
Qwen3-8B self-refinement	Minor	3764	3501	-263	7.0
Qwen3-8B self-refinement	Major	6428	6144	-284	4.4
Qwen3-8B self-refinement	Critical	96	74	-22	22.9

Table 18: xCOMET-XL severity-specific predicted error changes for the two core comparisons. Reduction is the percentage reduction relative to the initial error count.

J.3 xCOMET-XL High-Risk Examples

We manually inspect representative high-risk cases from the xCOMET-XL span analysis, defined as cases where post-editing decreases xCOMET-XL while adding severe predicted errors. Many such cases are concentrated in Qwen3-8B self-refinement. Table 19 shows one example: the draft correctly translates “shot” as being shot, but self-refinement changes it to a photography sense, introducing a severe adequacy error.

Field	Text
Source	Also shot were three men: two 29-year-olds and one 32-year old.
Reference	其他遭到枪击的三人中，两人29岁，一人32岁。 *Pinyin:* Qita zaodao qiangji de san ren zhong, liang ren 29 sui, yi ren 32 sui. *English gloss:* Of the other three people who were shot, two were 29 and one was 32.
Draft	另外还有三名男子被枪击，其中两名29岁和一名32岁。 *Pinyin:* Lingwai hai you san ming nanzi bei qiangji, qizhong liang ming 29 sui he yi ming 32 sui. *English gloss:* Three other men were also shot, including two aged 29 and one aged 32.
Self-refined output	另外还拍摄了三名男子：两名29岁的和一名32岁的。 *Pinyin:* Lingwai hai paishe le san ming nanzi: liang ming 29 sui de he yi ming 32 sui de. *English gloss:* Three other men were also photographed: two aged 29 and one aged 32.

Table 19: A high-risk Qwen3-8B self-refinement example identified by xCOMET-XL span diagnostics. The self-refined output introduces a harmful sense shift for "shot". Pinyin omits tone marks; English glosses preserve the meaning, including errors, of each Chinese string.

K Controlled Sentence-Level Regression

We report a diagnostic linear-regression summary with sentence-level ΔCOMET as the outcome. Its 10,080 observations comprise two outputs for each of 5,040 WMT source sentences: Sockeye+Qwen3-8B and Qwen3-8B self-refinement. The Sockeye indicator is 1 for the former and 0 for the latter. The available records do not specify how standard errors were handled with respect to within-source dependence; Table 20 therefore reports coefficients without standard errors or p-values.

Predictor	Coef.
Sockeye draft indicator	+0.0157
Initial COMET	-0.2210
Sentence chrF++ / 100	+0.0129
Edit rate	-0.0058
Length ratio	+0.0156

Table 20: Diagnostic sentence-level regression with ΔCOMET as the dependent variable. Sentence chrF++ is divided by 100. The positive Sockeye-draft coefficient remains after controlling for initial COMET, sentence-level chrF++, edit rate, and length ratio.

The negative initial-COMET coefficient is consistent with less headroom for higher-scoring drafts. The positive chrF++ and Sockeye coefficients describe conditional associations. Without verified uncertainty estimates or a causal design, these coefficients do not establish statistical significance or isolate a draft-source effect.

L Shared-Threshold Bucket Gap Analysis

As an additional diagnostic check, we split sentence-level drafts into four buckets using shared global thresholds over both Sockeye post-editing and Qwen3-8B self-refinement. For Table 21, the COMET threshold is 0.8789 and the chrF++ threshold is 27.9015. These are applied identically to both systems, so the bucket definitions are not system-specific. Within each bucket, we compute the mean ΔCOMET for Qwen3-8B self-refinement and Sockeye+Qwen3-8B, and report the gap between them.

All four mean gaps in Table 21 are positive, including the high-COMET buckets. Only descriptive gaps are reported, since the available records do not establish how dependence across shared sources and differing bucket memberships was handled. Shared thresholds do not exactly match initial-score distributions. Using sentence-level BLEU instead of chrF++ also gives positive descriptive gaps (+0.0256, +0.0284, +0.0117, and +0.0059).

M Edit-Rate Details

Table 22 reports the draft-to-post-edit edit-rate statistics used in Section 4.5 and the analysis section.

Bucket	Qwen n	Sockeye n	Qwen Δ	Sockeye Δ	Gap
high chrF++ / low COMET	524	1017	+0.0044	+0.0306	+0.0263
low chrF++ / low COMET	1879	1613	+0.0069	+0.0349	+0.0280
high chrF++ / high COMET	1680	1819	-0.0065	+0.0054	+0.0119
low chrF++ / high COMET	957	591	-0.0029	+0.0028	+0.0057

Table 21: Descriptive bucket-level ΔCOMET gaps under shared thresholds. Gap is Sockeye+Qwen3-8B minus Qwen3-8B self-refinement. Membership can differ between systems even though the test sources are shared.

Comparison	Type	Avg. ER	Med. ER	Len. Ratio	ΔCOMET
Sockeye+8B vs. Sockeye	Cross-system PE	0.262	0.235	0.868	+0.0203
Qwen3-8B+8B vs. Qwen3-8B	LLM self-refine	0.306	0.270	0.973	+0.0001
Qwen3-14B+8B vs. Qwen3-14B	LLM draft PE	0.126	0.089	0.985	+0.0019
Qwen3-14B+14B vs. Qwen3-14B	LLM self-refine	0.114	0.071	0.991	+0.0020

Table 22: Draft-to-post-edit edit rates for representative systems. Avg. ER and Med. ER denote the average and median character-level edit rates between the draft and the post-edited output. Len. Ratio denotes the average output-length ratio between the post-edited output and the draft.

N Qualitative Examples

Tables 23–25 provide representative sentence-level examples used in the qualitative discussion. The examples are selected to illustrate three recurring patterns: Sockeye drafts often preserve useful lexical or structural anchors, LLM post-editing can correct localized semantic or terminology errors, and LLM self-refinement may introduce harmful rewrites even when the output remains fluent.

Each example includes the source sentence, Chinese reference, draft, and edited output. Pinyin transliterations and English glosses are provided for readers; they are not model inputs or scored outputs. Pinyin omits tone marks and preserves numerals and Latin-script names. The glosses retain errors and awkward meanings in the Chinese strings so that the edits can be inspected in context. The main intended glosses are: “Tests” refers to cricket test matches rather than generic tests; “seconds, and thirds” refers to second and third servings; “going commercial” refers to commercialization rather than advertising; “customs” refers to customs duties rather than traditions; and “pressers” should be interpreted from the press/reporting context rather than as pressing devices.

Sockeye+8B WMT21-975 ΔCOMET: +0.2895	
Source	Pakistan and England will be facing each other in three Tests and as many T20Is.
Reference	巴基斯坦队和英格兰队将作为对手在三场测试赛和三场T20I 赛中相遇。 *Pinyin:* Bajisitan dui he Yinggelan dui jiang zuowei duishou zai san chang ceshisai he san chang T20I sai zhong xiangyu. *English gloss:* Pakistan and England will meet as opponents in three Test matches and three T20I matches.
Draft	巴基斯坦和英国将在三场测试和同样多的T20I 中对峙。 *Pinyin:* Bajisitan he Yingguo jiang zai san chang ceshi he tongyang duo de T20I zhong duizhi. *English gloss:* Pakistan and Britain will confront each other in three tests and as many T20Is.
Edited output	巴基斯坦和英格兰将在三场测试赛和同样数量的T20I比赛中交锋。 *Pinyin:* Bajisitan he Yinggelan jiang zai san chang ceshisai he tongyang shuliang de T20I bisai zhong jiaofeng. *English gloss:* Pakistan and England will face off in three Test matches and the same number of T20I matches.
Sockeye+8B WMT17-1488 ΔCOMET: +0.3525	
Source	Mr. Coe went back for seconds, and thirds.
Reference	科先生又吃了第二匙、第三匙。 *Pinyin:* Ke xiansheng you chi le di er chi, di san chi. *English gloss:* Mr. Coe ate a second spoonful and a third spoonful.
Draft	科先生又回去了几秒钟，又回去了三分之一。 *Pinyin:* Ke xiansheng you huiqu le ji miaozhong, you huiqu le san fen zhi yi. *English gloss:* Mr. Coe went back a few seconds, and went back one-third.
Edited output	科先生又吃了第二份和第三份。 *Pinyin:* Ke xiansheng you chi le di er fen he di san fen. *English gloss:* Mr. Coe had second and third servings.

Table 23: Representative Sockeye post-editing examples: sports terminology and idiomatic expressions.

Sockeye+8B WMT22-1022 ΔCOMET: +0.2256	
Source	Near the end, it touched on the Mindcrack going commercial.
Reference	临近结尾时，它提到了Mindcrack 的商业化。 *Pinyin:* Linjin jiewei shi, ta tidao le Mindcrack de shangyehua. *English gloss:* Near the end, it mentioned the commercialization of Mindcrack.
Draft	接近尾声时，它触及了Mindcrack 的广告。 *Pinyin:* Jiejin weisheng shi, ta chuji le Mindcrack de guanggao. *English gloss:* Near the end, it touched on advertisements for Mindcrack.
Edited output	接近尾声时，它提到了Mindcrack 商业化。 *Pinyin:* Jiejin weisheng shi, ta tidao le Mindcrack shangyehua. *English gloss:* Near the end, it mentioned Mindcrack going commercial.
Qwen3-8B self-refinement WMT22-1571 ΔCOMET: −0.0635	
Source	There may be local customs / VAT charges which are unknown to us and beyond our control.
Reference	当地可能会征收关税/增值税，这不在我们的控制范围内。 *Pinyin:* Dangdi keneng hui zhengshou guanshui/zengzhishui, zhe bu zai women de kongzhi fanwei nei. *English gloss:* Local customs duties/VAT may be levied; this is beyond our control.
Draft	可能会有当地关税/增值税费用，这些费用是我们无法预见且超出我们控制范围的。 *Pinyin:* Keneng hui you dangdi guanshui/zengzhishui feiyong, zhexie feiyong shi women wufa yujian qie chaochu women kongzhi fanwei de. *English gloss:* There may be local customs duties/VAT charges that we cannot foresee and that are beyond our control.
Edited output	可能存在我们未知且无法控制的当地习俗或增值税费用。 *Pinyin:* Keneng cunzai women weizhi qie wufa kongzhi de dangdi xisu huo zengzhishui feiyong. *English gloss:* There may be local customs (traditions) or VAT charges that are unknown to us and beyond our control.

Table 24: Further qualitative examples: a localized semantic repair for Mindcrack and a harmful shift from customs duties to local traditions.

Qwen3-8B self-refinement WMT21-167 ΔCOMET: -0.0977	
Source	“It’s not just a gotcha game,” Hume said of the intended purpose of the pressers.
Reference	“这不只是一场猫捉老鼠的游戏，” Hume 在谈到媒体人的预期目的时表示， *Pinyin:* “Zhe bu zhi shi yi chang mao zhuo laoshu de youxi,” Hume zai tandao meitiren de yuqi mudi shi biaoshi, *English gloss:* “This is not just a cat-and-mouse game,” Hume said when discussing the intended purpose of the media people,
Draft	休谟谈到记者们原本的意图时说：“这不仅仅是一个陷阱游戏。” *Pinyin:* Xiumo tandao jizhemen yuanben de yitu shi shuo: “Zhe bu jinjin shi yi ge xianjing youxi.” *English gloss:* Discussing the reporters’ original intentions, Hume said: “This is not merely a trap game.”
Edited output	“这不仅仅是一个陷阱游戏，”休谟说，他谈到按压器的预期用途。 *Pinyin:* “Zhe bu jinjin shi yi ge xianjing youxi,” Xiumo shuo, ta tandao anyaqi de yuqi yongtu. *English gloss:* “This is not merely a trap game,” Hume said, discussing the intended use of pressing devices.

Table 25: A Qwen3-8B self-refinement example in which a press-related expression is rewritten as a pressing device.

O Sentence-Level Gain Distributions

Table 26 reports the aggregate distribution of sentence-level ΔCOMET over WMT17, WMT21, and WMT22. We use the same threshold $\epsilon = 0.001$ as in the main analysis: degraded means $\Delta < -\epsilon$, nearly unchanged means $|\Delta| \leq \epsilon$, and improved means $\Delta > \epsilon$.

Comparison	Degraded	Nearly unchanged	Improved
Sockeye+Qwen3-8B vs. Sockeye	21.0% (1058)	4.8% (242)	74.2% (3740)
Qwen3-8B self-refinement	45.1% (2273)	7.4% (371)	47.5% (2396)
Qwen3-14B self-refinement	30.7% (1546)	28.0% (1411)	41.3% (2083)
Qwen3-14B+Qwen3-8B vs. Qwen3-14B	33.8% (1703)	24.0% (1212)	42.2% (2125)
Sockeye+Qwen3-8B vs. Qwen3-14B	39.5% (1991)	6.5% (326)	54.0% (2723)
Sockeye+Qwen3-8B vs. Qwen3-14B+Qwen3-14B	41.5% (2091)	6.9% (346)	51.6% (2603)

Table 26: Sentence-level ΔCOMET distributions over WMT17, WMT21, and WMT22. Values are percentages, with sentence counts in parentheses; each comparison contains 5,040 sentences. Counts combine the minor and large change subgroups, and percentages are calculated from these counts. The first four rows compare post-edited outputs with their corresponding drafts, while the last two rows compare Sockeye+Qwen3-8B against strong LLM baselines.

P Additional Draft-Source Control Details

All controls in this section use Qwen3-8B as editor and the same WMT test inputs, refinement prompt, and COMET evaluation protocol. The additional runs followed the existing inference pipeline with the draft model changed. Macro-averages weight the three test sets equally and differ from the pooled sentence-level means.

Reporting scope. The original scripts for these additional runs could not be recovered, and model-specific decoding details and statistical specifications (resampling, multiple-testing, and clustering) are incompletely documented. We therefore report aggregate scores, paired mean differences, and the regression point estimate as descriptive evidence, without confirmatory significance claims.

P.1 Weak-to-Strong Refinement and Common Support

Qwen3-0.6B improves from 0.813814 to 0.851956 (reported gain +0.038143), and Qwen3-1.7B from 0.847375 to 0.859434 (+0.012059). Sockeye improves from 0.850744 to 0.871048 (+0.020303). Scores and gains are rounded independently. The weakest draft has the largest gain but does not yield the best final output.

For the Sockeye–Qwen3-1.7B comparison, the analysis restricts source sentences to those for which both initial COMET scores fall within the overlapping 1st–99th percentile range. This retains 94.9%–96.8% of sentences across the three test sets. Table 27 reports paired mean gain differences in that region. Restricting support does not exactly match the initial-score distributions or control every possible source difference, so this is a descriptive sensitivity analysis.

The descriptive regression gives a Qwen3-1.7B coefficient of -0.009158 relative to Sockeye after controlling for initial COMET and dataset. The clustering unit is undocumented, so no uncertainty estimate is reported. This comparison differs from the original regression against Qwen3-8B self-refinement in Appendix K. The weaker-draft controls show that greater headroom can yield a larger raw gain, while similar initial macro-average quality need not yield similar refinement outcomes.

Dataset	Sockeye minus Qwen3-1.7B gain
WMT17	+0.011303
WMT21	+0.006060
WMT22	+0.006444

Table 27: Paired mean COMET gain differences within common support, with Qwen3-8B fixed as editor. Positive values favor Sockeye. Values are descriptive point estimates without uncertainty intervals.

P.2 Cross-Family GLM Drafts

GLM-4-9B-Chat provides a cross-family LLM draft source. Its macro-average COMET increases from 0.858213 to 0.863982 (+0.005769). Table 28 reports positive gains on each test set, the Sockeye–GLM differences in gains, and the differences in final quality. These paired mean differences are descriptive comparisons.

GLM starts above Sockeye in COMET on all three test sets, while the final ranking is reversed after refinement. Thus, GLM drafts also benefit from refinement, but refined Sockeye outputs score higher on all three test sets.

Q Supplementary Timing and Memory Measurements

The timing comparison uses a machine with three NVIDIA V100 32 GB GPUs, GPU-only inference, and no CPU or disk offloading. We report the

Dataset	GLM gain	Sockeye gain	Gain gap	Final-quality gap
WMT17	+0.007682	+0.023314	+0.015632	+0.005912
WMT21	+0.006630	+0.023479	+0.016849	+0.007964
WMT22	+0.002994	+0.014117	+0.011123	+0.007323

Table 28: Cross-family draft control under the fixed Qwen3-8B editor. Gain gap is Sockeye gain minus GLM gain; final-quality gap is refined Sockeye COMET minus refined GLM COMET. The gaps are paired by source sentence.

average translation time per sentence for each pipeline. The full Sockeye+Qwen3-8B pipeline averages 1.8490 seconds per sentence, compared with 2.0835 seconds for direct Qwen3-14B and 3.4803 seconds for direct Qwen3-32B translation. The stage-level averages are approximately 0.05 seconds for Sockeye draft generation and 1.7990 seconds for Qwen3-8B refinement. Table 29 also reports memory figures. Under simultaneous model residency, the aggregate memory estimate for Sockeye+Qwen3-8B is approximately 16.83 GiB.

System	Seconds/sentence	GiB
Sockeye draft only	≈0.05	≈1.48
Sockeye+Qwen3-8B	1.8490	16.83†
Qwen3-14B direct	2.0835	27.61
Qwen3-32B direct	3.4803	61.17

Table 29: Average per-sentence times for the respective experiments on the V100 setup. Sockeye+Qwen3-8B time covers the complete translation pipeline. †Aggregate residency estimate, not a measured per-device peak. Sockeye memory is 1,512 MiB.

These measurements describe the tested V100 configuration. Timing boundaries, batch size, warm-up, run-to-run variability, and memory-measurement definitions are incompletely documented. We therefore report supplementary averages rather than a standardized deployment benchmark. Aggregate residency does not represent per-device peaks or total GPU-hours; these results do not establish a general quality–efficiency frontier.