Feedback as an Interface: Controlling Form and Detail in LLM Translation Refinement

Yiqi Liu[1], **Flavjo Tomori**[2],
Nafise Sadat Moosavi[3], **Steffen Eger**[4], **Chenghua Lin**[1]*

[1]University of Manchester, [2] Osapiens,
[3]University of Sheffield, [4]University of Technology Nuremberg
{yiqi.liu, chenghua.lin}@manchester.ac.uk flavjotomori@gmail.com
n.s.moosavi@sheffield.ac.uk steffen.eger@utn.de

Abstract

Feedback presentation shapes how LLMs use evaluator judgments during translation refinement. We introduce a machine translation protocol that keeps the evaluator score, main issue, and recommendation fixed while varying feedback form and supporting detail. Across seven models and six translation directions, low-load feedback outperforms high-load feedback in all 42 translator-direction settings under xCOMET. On the 29,400-record rewriting subset, mean xCOMET scores multiplied by 100 are 83.30 for low-load feedback and 80.67 for high-load feedback; the full-matrix original-translation reference mean is 83.46. MetricX-24 supports the low-load advantage on a stratified subset. A no-feedback second pass yields a small improvement over the original translations, and direct evaluator prompting reproduces the low-load advantage without a separate rewriting stage. These results show that compact feedback limits the quality loss from evaluator-guided refinement in the evaluated high-resource MT settings.

1 Introduction

When an LLM evaluator identifies an accuracy error in another model's translation, the translator model is prompted to revise, and the resulting output may be unchanged, improved, or degraded. The same judgment can arrive as a terse instruction, a structured checklist, or a long explanation with qualifiers. These presentations may preserve the same score and core recommendation, yet an LLM reviser may respond to them differently. Feedback acts as an interface between evaluator judgment and revision.

This distinction matters because LLM refinement through iterative feedback is often brittle (Shinn et al., 2023; Madaan et al., 2023; Chen et al., 2024b; Kamoi et al., 2024). Across tasks, critique can help, do little, or degrade outputs (Pan et al., 2024; Wang et al., 2025; Wan et al., 2025; Huang et al., 2024; Tyen et al., 2024; Jiang et al., 2025). The unresolved question is which part of the feedback signal a model uses when revising.

Existing work usually varies both the identity of the critic (i.e., the evaluator model used to generate feedback) and the way feedback is expressed (Wang et al., 2025; Gou et al., 2024; Paul et al., 2024; Subramaniam et al., 2025). As a result, it is difficult to separate *who* critiques from *how* the same core critique is presented.

We isolate this sensitivity with a controlled protocol. For each source sentence and evaluator comment, we keep the score, main issue, and recommendation fixed while changing the feedback's wording, format, or amount of exposed detail. The final suite tests two surface-form dimensions, tone and response format, plus one content-density dimension, information load. Because rewrites can alter the underlying critique, rewrite faithfulness and style compliance are checked explicitly.

We therefore ask three questions. RQ1: Does changing feedback form or information load affect revision quality when the evaluator's score, main issue, and recommendation are held fixed? RQ2: How **robust** are the observed effects across metrics and validation checks? RQ3: Does direct evaluator prompting reproduce the low-load advantage without a separate rewriting stage? We use evaluator identity as a baseline, with feedback form and density as the main variables of interest. Analyses of tone, response format, and a second refinement

*Corresponding author.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1204–1221
November 6-10, 2026 ©2026 Association for Computational Linguistics

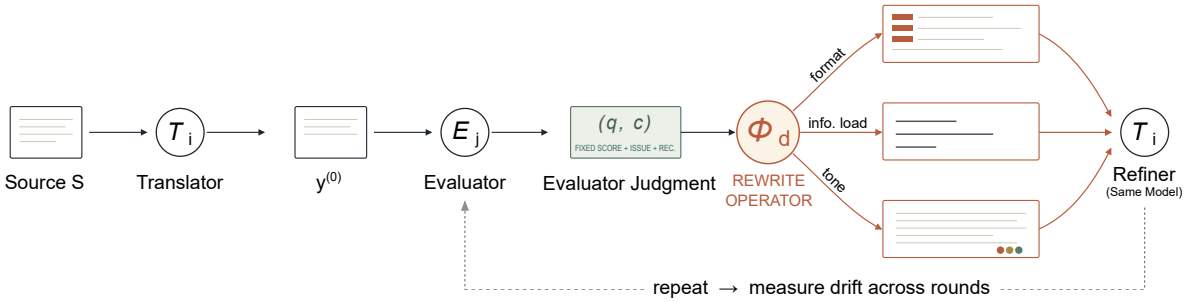


Figure 1: Controlled comparison for feedback presentation. The score, main issue, and recommendation are fixed, while the feedback is rewritten by tone, format, or information load before revision by the same translator.

round are reported in Appendices G and H.

Contributions. We make three contributions. First, we introduce an MT refinement evaluation protocol and dataset artifacts that separate critic identity from feedback form and supporting detail across seven models and six translation directions. Second, we show that low-load feedback limits quality degradation relative to high-load feedback, with support from xCOMET and a MetricX-24 subset evaluation. Third, we evaluate a no-feedback second pass and direct evaluator prompting to examine revision without evaluator feedback and to test compact feedback without an additional rewriting stage.

2 Related Work

LLM Judging and Refinement Self-refinement and self-correction studies report both gains and failures (Kamoi et al., 2024; Pan et al., 2024; Huang et al., 2024; Tyen et al., 2024; Jiang et al., 2025). Later variants add reranking, tools, confidence estimates, or reasoning diversity (Yan and Xu, 2023; Yao et al., 2023b; Wang et al., 2023; Yao et al., 2023a). Cross-refinement moves critique to a different model and often improves over pure self-correction (Wang et al., 2025; Gou et al., 2024; Paul et al., 2024). Multi-agent and ensemble pipelines add debate, adjudication, or consensus (Irving et al., 2018; Subramaniam et al., 2025; Rawat et al., 2025; Omar et al., 2025; Motwani et al., 2024; Chen et al., 2024a). These systems show that refinement is conditional and that critic identity can matter. They also tend to vary critic identity, critique content, roles, instructions, evidence, and wording together, leaving the feedback interface itself uncontrolled.

Varying the Presentation. Prompt-steerability and evaluator-bias studies show that wording and presentation can steer model judgments (Miehling et al., 2025; Wu and Aji, 2025; Fu and Liu, 2025; Hong et al., 2026). Related feedback-design work also varies instruction style or feedback phrasing (Zhou et al., 2023; Jia et al., 2024; Leiter and Eger, 2024). These studies motivate the interface view, but their outcomes usually concern judgments, preferences, or aggregate system performance. Our setting asks a more systematic and controlled question about revision: if the evaluator score, main issue, and recommendation are fixed, can feedback form or the amount of detail exposed to the reviser still change the revised translation?

We pair controlled rewrites with explicit faithfulness and style-compliance checks (Jia et al., 2024). This lets us test feedback presentation as a controllable variable under semantic constraints.

3 A Controlled Feedback Framework for LLM Refinement

This section defines the paper's central comparison: critic identity varies at the evaluator stage, while feedback form and exposed detail vary through controlled rewrites of the same evaluator judgment. Surface-form variants change how the critique looks; information-load variants change how much of the critique the reviser sees. Both preserve the score, main issue, and recommendation.

Table 1 summarizes the three dimensions, and Table 2 gives a concrete example from the rewrite logs.

3.1 Translator and Evaluator Refinement Framework

As shown in Figure 1, our setup has two roles: a translator and an evaluator, which may be the same or different underlying models. We use machine translation (MT) because small lexical or structural revisions can materially change quality, and sentence-level automatic metrics make large

Factor	Values and change
Tone	Positive, neutral, negative. Changes evaluative wording.
Response format	Structured, free-text. Changes organization.
Information load	Low load, high load. Changes supporting detail.

Table 1: Controlled feedback dimensions. Tone and response format are surface-form manipulations; information load changes the amount of supporting detail shown to the reviser.

controlled comparisons feasible. MT is also a common testbed for iterative LLM refinement (Madaan et al., 2023; Pan et al., 2024; Kocmi and Federmann, 2023).

In each round, the translator receives a source sentence and produces a translation. An evaluator is prompted to score the translation on a 0 to 100 scale using a prompt adapted from GEMBA[1] (Kocmi and Federmann, 2023). The evaluator sees only the source sentence and the candidate translation; no human reference is provided. We retain the full generated evaluator output as evaluation text and parse the numerical score separately. When the output contains explanatory text in addition to the requested score, that text is retained as feedback. The original translator receives the parsed score and retained evaluation text and produces a single revised translation. The evaluator may be identical to the translator (self feedback) or a different model (cross-model feedback). In controlled-rewriting conditions, a separate model rewrites the evaluator's evaluation text before it is sent back to the translator; the evaluator score and core evaluative content, namely the main issue and recommendation, are kept fixed.

For each comparison, the source, translator, evaluator, score, main issue, and recommendation remain fixed, while the displayed feedback variant changes. Surface-form variants change tone or organization. Information-load variants preserve the core judgment while changing how much supporting detail the reviser sees. Low-load feedback preserves the score, main issue, and recommendation while reducing supporting detail. The original translator then revises from the same initial translation under each feedback variant.

[1]Prompts for translator and evaluator roles are provided in the Appendix.

3.2 Controlled Feedback Dimensions

The experiment suite directly controls three observable properties of the evaluator comment. The target is preservation of *core evaluative content*: the score, main issue, and main recommendation. Tone and format are surface-form manipulations. Information load is a content-density manipulation: low-load feedback deliberately changes how much supporting detail is exposed to the reviser.

The main rewriting experiment uses GPT-OSS-20B, with prompts that instruct it to preserve the score, named entities, main issue, and recommendation. Additional experiments with direct structured or neutral feedback, near-identity rewriting, and Llama-3.1-8B as a second paraphraser are reported in Section 5.4. Section 5.3 evaluates direct low-load and high-load feedback generation. We assess content preservation in Section 5.4.

Tone. Tone prompts make the same core critique sound more positive, neutral, or negative while preserving the core issue and recommendation. We test these rewrites to measure whether sentiment-laden language helps or distracts the translator during revision (Van-Dijk and Kluger, 2004).

Response Format. Format controls whether the same feedback is presented as a structured template or as a single free-text paragraph. This isolates the contribution of explicit organization without changing the underlying criticism.

Information Load. This control varies information load: low-load rewrites keep the score, core issue, and direct recommendation while removing secondary detail and qualifiers where possible, whereas high-load rewrites preserve the fuller detail and qualification structure of the original comment (Paas et al., 2003; Centre for Education Statistics and Evaluation, NSW Department of Education, 2017). Length changes as a surface consequence: low-load rewrites average roughly $0.45\times$ the original character length, while high-load rewrites remain near parity (Appendix E).

The intended construct is the amount of supporting detail made available to the revising model.

4 Experimental Setup

4.1 Data Collection and Curation

We use parliamentary text to obtain grammatical, naturally occurring sentences while reducing obvious overlap with standard MT benchmarks. We

Variant	Feedback shown to the reviser
Raw	The translation captures the main idea of the German sentence but has issues with accuracy and naturalness. The message about heart disease as a leading cause of death is conveyed. However, "more than any other single factor" misrenders "als alles andere zusammen," which means "than all others combined." The last sentence is also off: the German says heart disease is preventable for almost all people, while "almost preventable in everyone" changes the emphasis. A score around 75 reflects the general idea with inaccuracies and awkward phrasing.
Low load	Score: 75/100. Main issue: two meanings are inaccurate, "than all others combined" and "preventable for almost all people." Recommendation: revise those phrases for accuracy and naturalness.
High load	The translation captures the main idea, but several accuracy and naturalness issues remain. The opening correctly conveys that heart disease is a leading cause of death worldwide. The comparative phrase weakens the source meaning: "als alles andere zusammen" means "than all others combined"; the translation gives "more than any other single factor." The final sentence also changes the emphasis. The German says heart disease is preventable for almost all people; "almost preventable in everyone" sounds awkward and suggests a different meaning. A score around 75 is appropriate because the message is understandable, but these two places need revision.
Structured	**Overall:** acceptable with notable inaccuracies. **Key reasons:** the core message is correct; "more than any other single factor" weakens "than all others combined"; "almost preventable in everyone" changes the final sentence. **Recommendation:** revise the comparative phrase and final sentence. **Rating:** 75/100.

Table 2: Feedback example for one German-to-English item with a 75/100 evaluator score. The table lightly shortens logged feedback for space. All rows keep the same score, main issue, and recommendation, while varying how the critique is organized and how much supporting detail is included. Bold labels in the structured variant reflect the labeled-field format; other variants are shown without added formatting.

curate 1,000 English and German sentences from the Hansard corpus[2] and the official records of the Deutscher Bundestag[3], respectively. These sources provide formal but naturally occurring text suitable for translation evaluation.

We sample 1,000 English sentences and 1,000 German sentences after sentence segmentation, language filtering, removal of obvious boilerplate, and exact-duplicate filtering in the final sentence list. We then run each translator on six directions: English→German/Spanish/Chinese and German→English/Spanish/Chinese. These model outputs serve as refinement objects and are distinct from human references. Unlike widely used MT benchmarks such as WMT, these parliamentary materials were curated from official record pages accessed in November 2024 and kept separate from standard MT benchmark releases. This avoids standard benchmark test-set reuse; exposure to public parliamentary records in model pretraining remains possible. We avoid standard benchmark test sets but do not claim source-text novelty. The resulting benchmark-style artifact contains 42,000 initial translations (7 translator models $\times$ 6 directions $\times$ 1,000 sentences) and 294,000 evaluator feedback instances. The rewriting experiment uses a stratified subset of 29,400 comments, with 100 source items for each combination of translator, evaluator, and translation direction ($7\times7\times6$). All controlled rewrite families are generated from this same 29,400-comment base set. This gives equal coverage of all combinations while keeping rewriting, refinement, and metric evaluation computationally feasible. Applying the same design to all 294,000 evaluator comments would increase this workload by approximately a factor of ten.

We will release the sentence lists or source identifiers, generation scripts, prompts, translations, evaluator outputs, paraphrased feedback variants, and scoring scripts where redistribution is permitted; for any source text with unresolved redistribution terms, we will release URLs/IDs and derived artifacts, while restricted text stays unreleased.

4.2 Model Selection

We use seven open or openly available instruction-tuned models spanning 7B to 27B-class systems:

[2] https://hansard.parliament.uk/; UK Parliament content is available under the Open Parliament Licence v3.0 at https://www.parliament.uk/site-information/copyright-parliament/open-parliament-licence/.

[3] https://www.bundestag.de/dokumente/protokolle/amtlicheprotokolle; the Bundestag provides plenary records through its Open Data service at https://www.bundestag.de/services/opendata, while general website content remains subject to the Bundestag legal notice at https://www.bundestag.de/en/service/legal_notice.

LLaMA, Qwen, Mistral, Gemma, and GPT-OSS-20B.[4] This gives a heterogeneous translator/evaluator matrix for multi-model comparison. The main matrix excludes frontier API-only evaluators such as GPT-4 or Claude because the goal here is a repeatable, large-scale benchmark. Full checkpoint names are provided in Table 6.

4.3 Meta-Evaluation with xCOMET

We evaluate metric-predicted refinement quality using xCOMET-XL,[5] an automatic MT quality metric where higher scores indicate better predicted translation quality (Guerreiro et al., 2024). We run xCOMET in reference-free mode, using only the source sentence and the candidate translation. xCOMET builds on the COMET family of metrics, which has demonstrated high correlation with expert human MQM annotations across recent benchmark evaluations (Kocmi et al., 2021). It is a useful automatic proxy for sentence-level translation quality, but it remains one metric family. Prior work also shows that language-model-based evaluation can exhibit systematic bias, particularly in reference-free settings (Liu et al., 2024). Throughout the paper, we report xCOMET multiplied by 100 for readability. The reported findings on tone, response format, and two-round score changes are specific to xCOMET. The low-load versus high-load comparison is additionally evaluated with MetricX-24. We use MetricX-24-Hybrid in reference-free mode (Juraska et al., 2024). Both xCOMET and MetricX-24 are learned metrics, so agreement between them provides additional automatic-evaluation evidence.

4.4 Controlled Feedback Rewriting

The base-refinement condition uses the evaluators' original evaluation text. We paraphrase this 29,400-comment subset with GPT-OSS-20B while preserving each comment's score, main issue, and main recommendation. The paraphraser only rewrites the evaluator's existing comment and leaves the score unchanged. To ensure the validity of the rewrite, the evaluation text is checked through the 84-pair rewrite validity sample in Section 5.4 and Appendix E.

The final analysis reports three prompt families: tone (positive, neutral, negative), response format (structured, free-text), and information load (low/high-load variants).[6]

Rewriting prompt. The rewriting model receives the evaluator output and a requested presentation value while being instructed to preserve the score, factual content, named entities, main issue, and recommendation. The exact prompt used in the main rewriting experiment is shown below.

Prompt - Controlled Feedback Rewriting

Rewrite the evaluator feedback below to match the requested presentation style while preserving the underlying critique.

Requirements. Preserve the original score, factual content, named entities, language, main issue, and recommendation. Do not add new facts or remove key conclusions. Adjust only the requested presentation dimension. Output only the rewritten evaluation.

Input feedback. "{*evaluation_text*}"

Target dimension values. Tone: positive / neutral / negative. Response format: structured / free_text. Information load: low / high.

Style guidance. Structured uses labeled fields; free-text uses one cohesive paragraph. Low load keeps only decision-critical points; high load keeps relevant details and qualifiers. Tone changes wording emphasis while preserving the same core critique.

Rewritten evaluation:

In every condition, the original translator performs the final revision, keeping the refiner fixed within each comparison. The remaining variation comes from evaluator choice and the feedback shown to the translator. For load conditions, this includes the amount of supporting detail; for all rewrite conditions, paraphraser artifacts remain a possible confound. The results therefore include direct-control, near-identity, and second-paraphraser checks.

Implementation details. The local generation scripts use the HuggingFace `transformers` text-generation pipeline with bfloat16 weights, `device_map=auto`, and `return_full_text=False`. Generation is deterministic under the recorded scripts because no sampling parameters are enabled; translation,

[4] In preliminary experiments, we also evaluated smaller models but found that models with approximately 3B parameters often struggled to follow refinement instructions reliably, introducing substantial noise that could confound the analysis. We therefore focus on models with at least 7B parameters to ensure sufficient instruction-following and refinement capability.

[5] https://huggingface.co/Unbabel/XCOMET-XL

[6] The main rewriting subset contains 29,400 evaluator comments. Tone has three rewrite values, while response format and information load have two each.

evaluator-feedback, refinement, and score-extraction calls use maximum new-token budgets of 400, 1000, 600, and 512 respectively. Batch sizes are 50 to 100 except for Gemma-family jobs, which use batch size 1. xCOMET is run through the COMET checkpoint interface with batch sizes of 50 to 100 on one GPU. Paired-bootstrap summaries use a fixed analysis seed of 23.

5 Results

We first report raw-feedback refinement and a no-feedback second pass (§5.1), followed by the information-load comparison (§5.2), direct evaluator prompting (§5.3), and feedback preservation (§5.4). Tone, response format, and two-round results are reported in Appendices G and H.

5.1 Raw Feedback and a No-Feedback Second Pass

Base quality varies substantially by model and direction. This matters because weak translations leave room for correction while strong translations can be over-edited. German→English is easiest on average, whereas Chinese-target directions are markedly harder; Table 3 reports the full base matrix.

Table 4 summarizes the default refinement loop using the evaluators' original outputs. The overall mean falls from 83.46 for the original translations to 82.13 after one refinement round. Self feedback is the weakest one-round setting (81.40), while the cross-model critic selected retrospectively from the evaluated outputs, reaches 83.32.

Brittleness varies by translator and direction. Weaker translators sometimes benefit from suitable cross-model critics, while stronger translators are often over-edited: even the retrospectively selected cross-model critic stays below the original score for Gemma3-12B, Gemma3-27B, Llama3.1-8B, and GPT-OSS-20B. Raw refinement helps German→Chinese (+3.29 over the original average), hurts English→Spanish (-5.01) and English→German (-2.04). This conditional pattern motivates the central question: if critique frequently hurts, can controlling its form and density reduce the risk?

No-feedback second pass. The translator receives the source sentence and its original translation and produces a second version without evaluator feedback. On 2,700 examples, this condition improves xCOMET×100 by 0.39 over the original translations (95% CI [0.05, 0.70]).

5.2 Low-Load Feedback Limits Quality Degradation

With evaluator judgment held fixed, different feedback variants lead to different revisions. Table 5 reports low-load and high-load results on the rewriting subset alongside full-matrix reference means for raw refinement and the original translations.

Low-load feedback outperforms high-load feedback in all 42 translator-direction settings, with a mean difference of 2.63 xCOMET×100 (95% CI [2.26, 3.05]). On the 29,400-record rewriting subset, the mean scores are 83.30 for low-load feedback and 80.67 for high-load feedback. For reference, the full-matrix means are 83.46 for the original translations and 82.13 for raw refinement. Low-load feedback therefore limits the quality loss relative to high-load feedback.

The MetricX-24 subset evaluation shows the same low-load advantage. On a stratified 1,008-example subset, we sample 24 feedback instances from each of the 42 translator-direction settings, with evaluator identities represented within each setting; low-load and high-load feedback use the same 1,008 common keys. After converting MetricX-24 predictions to higher-is-better scores, low-load feedback exceeds high-load feedback by 0.54 points (95% bootstrap CI [0.32, 0.76]).[7]

5.3 Direct Evaluator Prompting

We prompt evaluators to generate low-load or high-load feedback directly, without a separate rewriting stage or an injected score line. On a 5,400-example subset, the mean xCOMET×100 scores are 68.42 for low-load feedback and 67.27 for high-load feedback, a difference of 1.15 (95% CI [0.43, 1.85]). This subset has a different model and direction composition from the main rewriting experiment, so we interpret the difference within this subset. The evaluator can produce different judgments under the two prompts; this experiment evaluates direct feedback generation.

5.4 Feedback Preservation and Validation

Faithfulness triage. We run a heuristic pass over 205,800 rewrite rows and flag cases where explicit

[7] We report 25 minus the MetricX-24 raw prediction so that higher values indicate better predicted quality. This transformation reverses the direction of the original scale while preserving pairwise difference magnitudes.

Translator \ LP	En→De	En→Es	En→Zh	De→En	De→Es	De→Zh	Avg.
Mistral-7B	87.51	83.83	58.78	93.21	85.86	66.38	79.26
Llama3.1-8B	91.07	86.65	68.76	94.34	88.58	71.95	83.56
Qwen2.5-7B	86.66	85.87	72.06	92.74	83.95	68.65	81.66
Qwen2.5-14B	90.39	88.02	73.64	93.10	84.55	72.86	83.76
Gemma3-12B	93.36	89.97	73.78	92.60	89.47	69.88	84.84
Gemma3-27B	93.91	90.16	75.40	92.70	89.49	71.91	85.60
GPT-OSS-20B	92.78	88.91	72.41	94.90	89.68	74.62	85.55
Avg.	90.81	87.63	70.69	93.37	87.37	70.89	–

Table 3: Base translation quality (xCOMET×100) by translator and language pair.

Translator	**Original**	**Self Feedback**	**Best Cross-Critic**	**Avg. over Critics**
Mistral-7B	79.26	75.16	**80.53**	78.47
Llama3.1-8B	83.56	80.70	83.21	81.86
Qwen2.5-7B	81.66	**81.89**	**83.70**	**82.52**
Qwen2.5-14B	83.76	82.49	**84.06**	83.15
Gemma3-12B	84.84	82.72	84.08	82.68
Gemma3-27B	85.60	83.47	84.06	83.41
GPT-OSS-20B	85.55	83.39	83.61	82.85
Avg.	83.46	81.40	83.32	82.13

Table 4: Mean refinement quality (xCOMET×100) across six translation directions. Self feedback uses the same model as translator and evaluator. For each translator and translation direction, Best Cross-Critic selects the non-identical evaluator with the highest mean refinement score, then averages the selected scores across directions. Selection uses the evaluated outputs. Bold values exceed the original translation baseline.

Condition	**xCOMET**×100
Original translation (full matrix)	83.46
Raw refinement (full matrix)	82.13
Low-load feedback	83.30
High-load feedback	80.67

Table 5: Mean xCOMET×100 scores. Original translation and raw refinement are full-matrix reference means; low-load and high-load feedback are means on the 29,400-record rewriting subset.

scores, numeric spans, or quoted fragments appear to have changed or disappeared. Because the raw paraphraser logs often contain wrapper text and formatting artifacts, this stage is deliberately conservative and is used to prioritize follow-up inspection while avoiding a precise preservation-rate claim. It produces a stratified 84-pair annotation sheet covering all reported rewrite families. We then run a judge-based check over that full sample with Qwen2.5-14B and Gemma3-27B as independent judges, followed by third-judge adjudication on disagreement cases. The label set records whether the rewrite preserves the original issue, severity, and recommendation, whether it introduces a new unsupported issue, and whether it removes important detail. Adjudicated majority labels indicate that the rewrite preserves the original issue in 81/84 cases, preserves the recommendation in 81/84 cases, and introduces no new unsupported issue in 80/84 cases; a conservative strict-preservation criterion holds in 70/84 cases. By value, the weakest family remains negative tone (7/12 strict-preserved), while structured and low-load rewrites more often preserve the core issue than exact secondary detail. The scaled check therefore strengthens the methodological claim without licensing a statement of perfect semantic invariance.

Paraphraser artifacts. Concern: the paraphraser might create the apparent presentation effect. Check: we prompt the evaluator itself to produce structured or neutral feedback (direct control), then rerun the same subset with Llama-3.1-8B as a second paraphraser. Result: the direct-control condition partially reproduces the same ordering (structured 76.88, neutral 76.30, gap +0.57; 95% CI [0.05, 1.08]), and the second paraphraser preserves the same qualitative picture. A single-paraphraser

explanation becomes less likely.

Generic rewriting effects. A near-identity rewrite keeps feedback close to its original wording. Its mean xCOMET$\times 100$ score is 1.87 points below raw refinement on the same examples (95% CI $[-2.27, -1.45]$). Appendix H reports its scores across two refinement rounds.

Illustrative validation patterns. The failure modes in the scaled check are also informative. Negative-tone failures rarely invent a completely unrelated critique; they more often compress a largely positive or mixed assessment into language that sounds globally harsher, shifting severity and occasionally implying a new unsupported issue. Structured and low-load rewrites show a different pattern: they usually retain the main recommendation, but sometimes compress away supporting evidence, quoted phrases, or numeric detail that justified the original score. These patterns show that the requested presentation changes are usually realized, although severity and supporting detail can change during rewriting.

Representative cases. The sampled cases reflect both sides of this pattern. One preserved example comes from a Qwen→Chinese rewrite where a free-text paraphrase keeps the original score, core issue, and wording adjustments almost verbatim while removing only superficial redundancy. One risky negative-tone example comes from a German→English case where the original comment assigns 95/100 with only a minor stylistic qualification, then the rewrite compresses that assessment into a more globally unfavorable judgment. One risky structured example comes from an English→German case where the main recommendation survives, although the rewrite drops quoted evidence and numeric detail that justified the score. These cases match the aggregate validation result: the main issue is usually preserved, while severity and supporting detail are the first things to drift.

Style-compliance check. We also measure whether requested surface-form controls were actually realized. Response-format control is strongest: structured rewrites almost always exhibit list-like or labeled organization, while free-text rewrites remain single-paragraph prose. Information-load control is also visible on the surface, with low-load rewrites averaging roughly $0.45\times$ the original character length and high-load rewrites staying near parity. Tone control is noticeably noisier under simple lexical heuristics, especially for negative and neutral prompts. The clearest evidence is therefore for steerable structure and information load, while tone should be interpreted more cautiously unless stronger validation is added.

Feedback length. Within each load condition, feedback length has little association with xCOMET scores ($|\rho| < 0.03$). A mixed-effects regression over 58,800 low-load and high-load rows models xCOMET as a function of feedback condition and log feedback character count, with translator random intercepts and language-pair variance components. The high-load coefficient is $\beta = -4.02$ (95% Wald CI $[-4.42, -3.62]$, $p < 0.001$).

6 Discussion

Feedback as an interface. The results show that feedback presentation affects how LLMs use evaluator judgments during revision. Preserving the score and core recommendation still permits different revision behavior depending on how the same critique is organized or compressed. A refinement pipeline should therefore treat critic identity, critique content, and presentation form as separate design choices.

Connecting to feedback-design work. The older feedback-design vocabulary remains useful as an interpretation layer, although the controlled experiments here measure form and exposed detail without measuring psychological constructs. Compact feedback increases directness by removing secondary explanations; structured feedback changes organization; neutral tone changes framing but with weaker and less robust evidence.

Possible explanations. Additional supporting detail may direct attention to several revision targets. Compact feedback may also resemble the direct instructions used to train instruction-following models. Distinguishing these explanations requires targeted experiments, such as equal-length feedback variants or score-only, issue-only, and recommendation-only baselines.

Practical implications. In pipelines that already use evaluator-guided refinement, low-load feedback can be generated directly when evaluator prompts can be changed. A separate rewriting stage remains useful when evaluator feedback is fixed or supplied externally. The direct-prompt experiment supports this application on its evaluation subset.

7 Conclusion

Across the evaluated models and translation directions, low-load feedback produces less quality degradation than high-load feedback under xCOMET and shows the same comparative advantage under MetricX-24. Its subset mean is close to the full-matrix original-translation reference mean. The no-feedback second-pass experiment provides a baseline for revision without evaluator feedback, and direct evaluator prompting reproduces the low-load advantage without a separate rewriting stage. These findings characterize feedback use in the evaluated high-resource MT settings.

Limitations

Claim scope. The main finding concerns the relative effects of low-load and high-load feedback in the evaluated models and translation directions. Low-load feedback yields higher scores than high-load feedback, and its subset mean is close to the full-matrix original-translation reference mean. Tone, response format, and two-round results are specific to xCOMET. The two-round comparison describes score changes from different round-one starting points on the English→German and German→Chinese subset.

Load construct. Information load is implemented by changing exposed supporting detail while preserving the score, main issue, and recommendation. The feedback-length analysis leaves the contributions of supporting detail, feedback length, and resemblance to instruction-style prompts unresolved. The experiments also lack low-load and high-load variants of equal length, as well as score-only, issue-only, and recommendation-only baselines. Multi-turn interaction and richer conversational feedback strategies remain open avenues.

Rewrite validity. Most rewrite experiments use GPT-OSS-20B as the paraphraser, and GPT-OSS-20B also appears in the translator/evaluator matrix. Direct-control, near-identity, and second-paraphraser checks reduce this concern, yet paraphraser-specific behavior remains possible. The faithfulness and style-compliance checks are sampled, so the strongest claims concern downstream behavior under controlled approximations; perfect semantic invariance is outside the supported scope.

Task and evaluation scope. The study is confined to high-resource machine translation directions and primarily to automated xCOMET evaluation, with MetricX-24 evaluated on a stratified subset. Human MQM-style judgments, additional independent metrics, lower-resource directions, and non-translation tasks would be needed before making broader quality or transfer claims. The parliamentary source text avoids standard MT benchmark releases, while exposure through public-record pretraining remains possible.

Ethical Considerations

The study uses public parliamentary proceedings and model-generated translations, critiques, and revisions. The data may include names and political statements by public officials, so released artifacts will preserve source attribution and avoid suggesting endorsement by the source institutions. The work is a diagnostic study of refinement behavior outside deployment as a translation or political analysis system; generated translations and critiques may contain metric-preferred but human-unverified errors.

Release is governed by source terms. Hansard material is reused under the Open Parliament Licence, whereas Bundestag website content and Open Data materials are subject to the Bundestag's published terms. We will therefore release scripts, prompts, model-output artifacts, scoring outputs, and source IDs/URLs, and will redistribute raw source text only where the applicable terms permit it.

References

Centre for Education Statistics and Evaluation, NSW Department of Education. 2017. Cognitive load theory: Research that teachers really need to understand.

Justin Chih-Yao Chen, Archiki Prasad, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. 2024a. MAgICoRe: Multi-agent, iterative, coarse-to-fine refinement for reasoning. *arXiv preprint arXiv:2409.12147*.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024b. Teaching large language models to self–debug. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*. Poster.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*,

pages 11040–11053, Suzhou, China. Association for Computational Linguistics.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

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.

Hanhua Hong, Chenghao Xiao, Yang Wang, Yiqi Liu, Wenge Rong, and Chenghua Lin. 2026. Beyond one-size-fits-all: Inversion learning for highly effective NLG evaluation prompts. *Transactions of the Association for Computational Linguistics*, 14:689–710.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899*.

Qinjin Jia, Jialin Cui, Ruijie Xi, Chengyuan Liu, Parvez Rashid, Ruochi Li, and Edward Gehringer. 2024. On assessing the faithfulness of LLM-generated feedback on student assignments. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 491–499, Atlanta, Georgia, USA. International Educational Data Mining Society.

Dongwei Jiang, Jingyu Zhang, Orion Weller, Nathaniel Weir, Benjamin Van Durme, and Daniel Khashabi. 2025. Self-[in]correct: LLMs struggle with discriminating self-generated responses. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24266–24275.

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.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can LLMs actually correct their own mistakes? a critical survey of self–correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Christoph Leiter and Steffen Eger. 2024. PrExMe! large scale prompt exploration of open source LLMs for machine translation and summarization evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11481–11506, Miami, Florida, USA. Association for Computational Linguistics.

Yiqi Liu, Nafise Moosavi, and Chenghua Lin. 2024. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701, Bangkok, Thailand. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36.

Erik Miehling, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth M. Daly, Kush R. Varshney, Eitan Farchi, Pierre Dognin, Jesus Rios, Djallel Bouneffouf, Miao Liu, and Prasanna Sattigeri. 2025. Evaluating the prompt steerability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7874–7900, Albuquerque, New Mexico. Association for Computational Linguistics.

Sumeet Ramesh Motwani, Chandler Smith, Rocktim Jyoti Das, Rafael Rafailov, Ivan Laptev, Philip HS Torr, Fabio Pizzati, Ronald Clark, and Christian Schroeder de Witt. 2024. MALT: Improving reasoning with multi-agent LLM training. *arXiv preprint arXiv:2412.01928*.

Mahmud Omar, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. 2025. Refining LLM outputs with iterative consensus ensemble (ICE). *International Journal of Medical Informatics*, 196:110731.

Fred Paas, Alexander Renkl, and John Sweller. 2003. Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38(1):1–4.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction

strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2024. REFINER: Reasoning feedback on intermediate representations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1100–1126, St. Julian's, Malta. Association for Computational Linguistics.

Mrinal Rawat, Ambuje Gupta, Rushil Goomer, Alessandro Di Bari, Neha Gupta, and Roberto Pieraccini. 2025. Pre-act: Multi-step planning and reasoning improves acting in LLM agents. *arXiv preprint arXiv:2505.09970*.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

Vighnesh Subramaniam, Yilun Du, Joshua B. Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. 2025. Multiagent finetuning: Self improvement with diverse reasoning chains. In *The Thirteenth International Conference on Learning Representations*.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. 2024. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13894–13908, Bangkok, Thailand. Association for Computational Linguistics.

Dina Van-Dijk and Avraham N. Kluger. 2004. Feedback sign effect on motivation: Is it moderated by regulatory focus? *Applied Psychology*, 53(1):113–135.

David Wan, Justin Chen, Elias Stengel-Eskin, and Mohit Bansal. 2025. Mamm-refine: A recipe for improving faithfulness in generation with multi-agent collaboration. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Mexico. Association for Computational Linguistics.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, Sebastian Möller, and Vera Schmitt. 2025. Cross-refine: Improving natural language explanation generation by learning in tandem. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1150–1167, Abu Dhabi, UAE. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Minghao Wu and Alham Fikri Aji. 2025. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 297–312, Abu Dhabi, UAE. Association for Computational Linguistics.

Tianqiang Yan and Tiansheng Xu. 2023. Refining the responses of LLMs by themselves. *arXiv preprint arXiv:2305.04039*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Appendix Roadmap

Appendix B summarizes checkpoint metadata, Appendix C gives the active prompt families, Appendix D records the controlled dimensions, Appendix E gives representative validation cases, and Appendix F visualizes the full translator and evaluator refinement grid. Appendix G and Appendix H report tone and response-format effects and two-round xCOMET results, respectively.

B Model Selection

Table 6 summarizes the checkpoints used in the main matrix together with publicly stated cutoff information where it is available.

C Prompt Design

This appendix documents the core prompt families used in the main rewriting experiment: translation, evaluation, refinement, and controlled feedback rewriting. Exact prompts for the no-feedback second-pass and direct evaluator-prompting follow-up experiments are included in the released experiment scripts.

D Controlled Dimensions

Table 7 summarizes the controllable feedback dimensions used in the final experiments.

D.1 Prompts for Translation, Evaluation, and Refinement

The boxes below show the English→German instantiation used for illustration; the same templates are applied to the other directions by swapping language names and placeholders.

Prompt - Translation

Translate the following English sentence into German. Please provide the translation only. Please be brief and do not include any notes or discussion.

English sentence: {*sentence*}

German sentence:

Prompt - Evaluation

Score the following translation from English to German on a continuous scale from 0 to 100, where a score of zero means "no meaning preserved" and a score of one hundred means "perfect meaning and grammar".

English source: "{*source_seg*}"

German translation: "{*target_seg*}"

Score:

Evaluator-output handling. The prompt above is the executed evaluator prompt and explicitly requests a numerical score. We retain the full generated output as the evaluation text and parse the score separately; explanatory text generated alongside the score is retained as feedback for refinement.

Prompt - Refinement

You previously provided the following translation from English to German:

Source Sentence: {*source_seg*}
Your Original Translation: {*target_seg*}

The translation was evaluated by a mentor who scored it {*mentor_score*} out of 100.

Feedback:
"{*mentor_explanation*}"

Task:
Revise it to improve meaning preservation, grammatical fluency, and accuracy as needed. If you think no changes are needed (based on the feedback), feel free to leave it unchanged.

E Validation Checks for Controlled Rewrites

We report supporting analyses of rewrite generation, content preservation, style compliance, and feedback length. These analyses characterize how the rewriting stage behaves and where the requested presentation changes can alter secondary details.

F Results of Base Refinements

Figures 2 — 4 report mean Δ xCOMET$\times$100 relative to the original translation for each translator and evaluator pair. The diverging color scale is

Model Name	Name of Checkpoint	Training Data Cutoff	Before Nov. 2024?
Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct	2023-12	✓
Qwen2.5-7B	Qwen/Qwen2.5-7B-Instruct	Not publicly disclosed (before 2024-09)	✓
Qwen2.5-14B	Qwen/Qwen2.5-14B-Instruct	Not publicly disclosed (before 2024-09)	✓
Mistral-7B	mistralai/Mistral-7B-Instruct-v0.3	Not publicly disclosed (before 2024-05)	✓
Gemma3-12B	google/gemma-3-12b-it	2024-08	✓
Gemma3-27B	google/gemma-3-27b-it	2024-08	✓
GPT-OSS-20B	openai/gpt-oss-20b	2024-06	✓

Table 6: Selected models used as both translators and evaluators.

Dimension	What changes	Values	Held fixed
Tone	Surface wording and sentiment	Positive, neutral, negative	Score, main issue, recommendation
Response format	Organization of the same critique	Structured, free-text	Score, main issue, recommendation
Information load	Amount of supporting detail shown to the reviser	Low load, high load	Score, main issue, recommendation

Table 7: Controlled feedback dimensions used in the final experiments. Tone and response format manipulate surface form; information load manipulates accessible detail while keeping the core judgment fixed.

centered at zero, so positive cells indicate helpful refinement and negative cells indicate degradation.

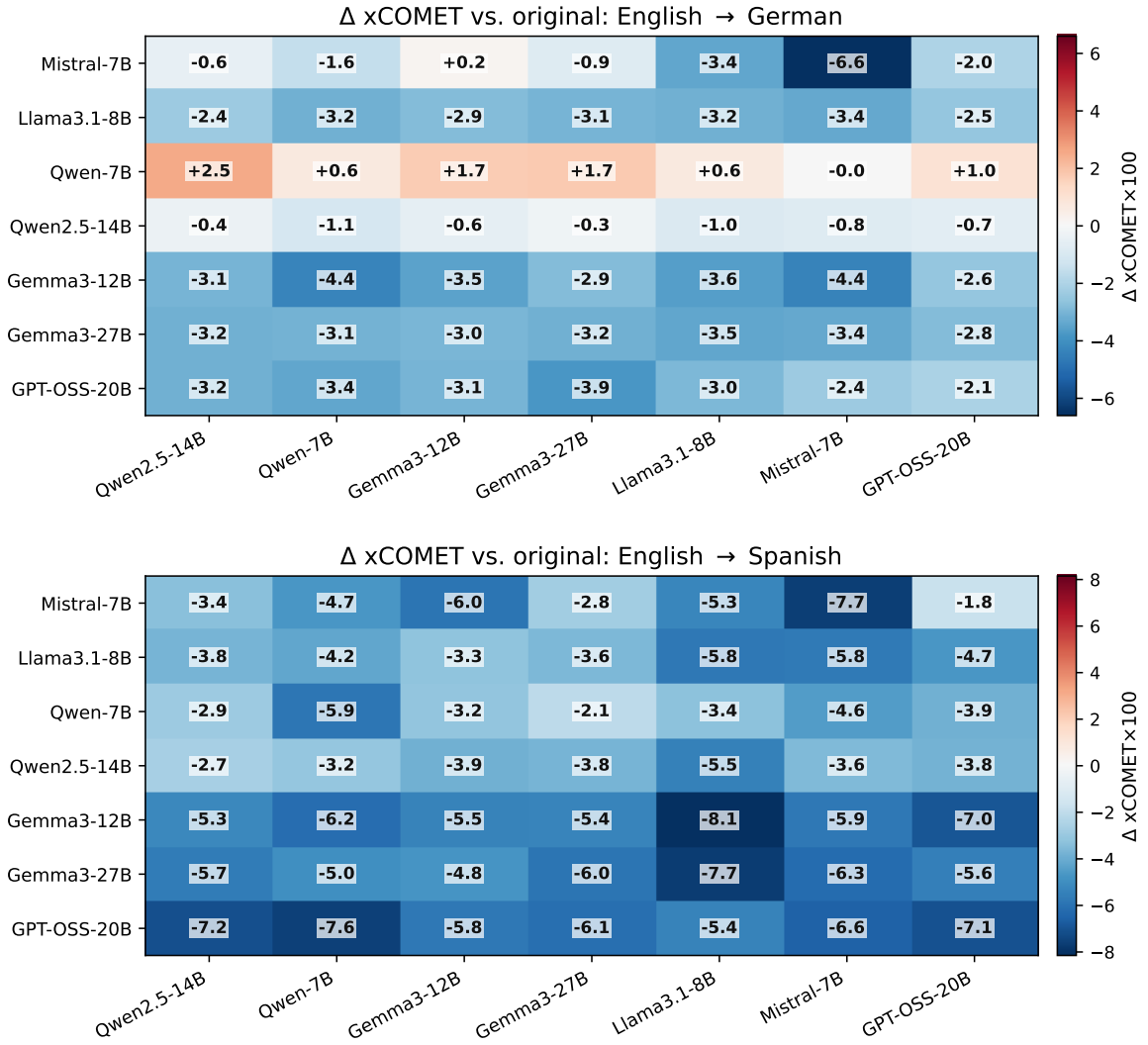


Figure 2: Appendix heatmaps for English-source directions. The top panel shows English→German and the bottom panel shows English→Spanish. Positive cells indicate that refinement helps relative to the original translation; negative cells indicate degradation.

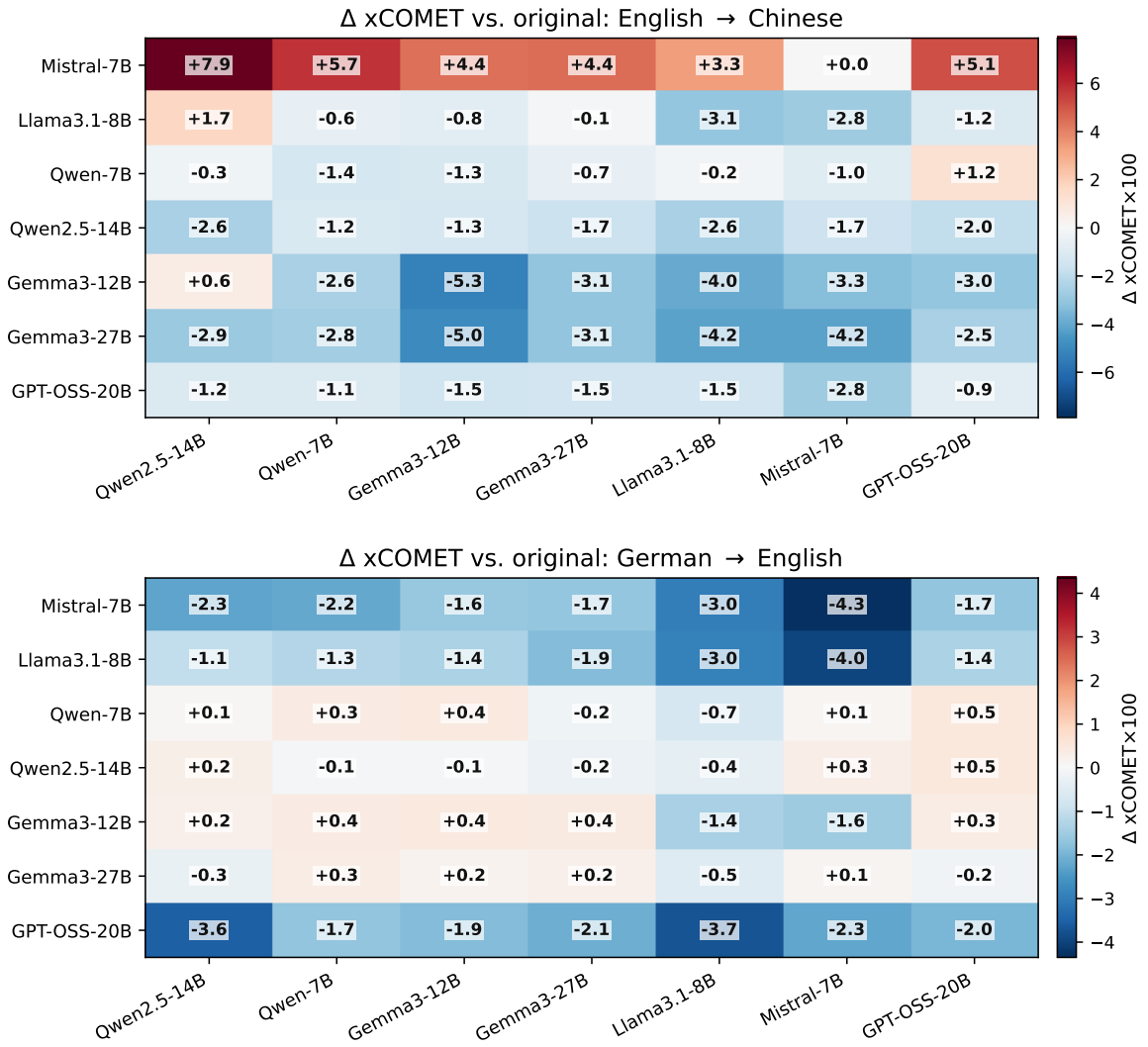


Figure 3: Appendix heatmaps for English→Chinese (top) and German→English (bottom). These panels contrast one harder Chinese-target setting with one easier English-target setting under the same translator and evaluator grid.

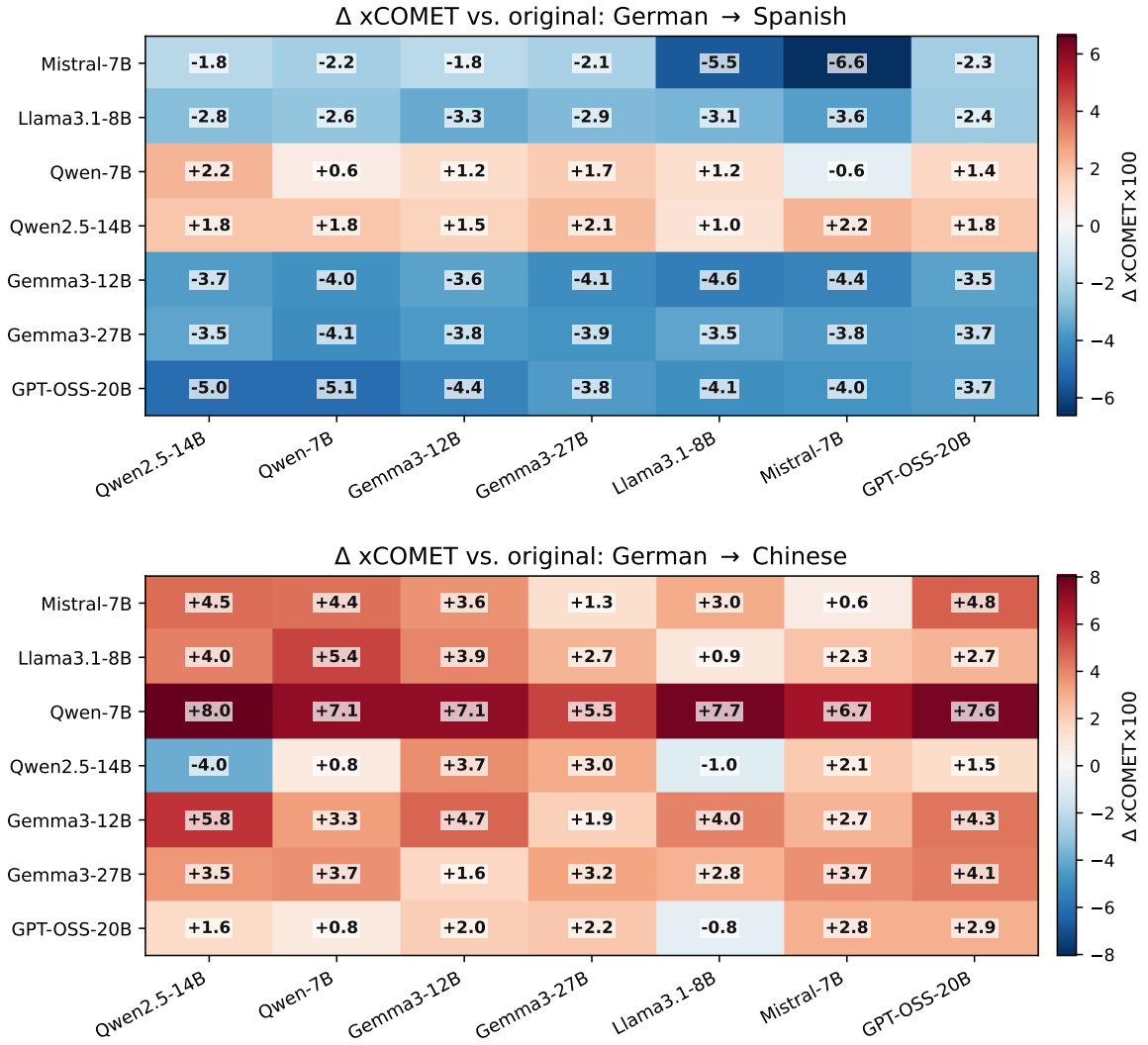


Figure 4: Appendix heatmaps for the remaining German-source directions: German→Spanish (top) and German→Chinese (bottom). Across all appendix panels, warmer colors indicate mean improvement and cooler colors indicate mean degradation relative to the original translation.

G Tone and Response Format under xCOMET

Structured, free-text, and neutral rewrites score above raw refinement under xCOMET. The reported tone and format findings remain metric-dependent. Table 8 reports these results.

Raw feedback is generated without an explicit requirement on tone or organization. Neutral and free-text prompts impose additional presentation requirements, so their outputs can differ from raw feedback. The observed differences also include the effects of paraphrasing.

Structured and neutral rewrites have xCOMET means 1.56 and 1.37 points above the full-matrix raw-refinement reference, respectively. The MetricX subset leaves structured feedback close to raw ($\Delta = -0.04$, 95% CI [-0.15, 0.07]), so these surface-form results remain metric-dependent. Figure 5 summarizes setting-level deltas relative to raw refinement.

Tone is weaker and noisier: both positive and negative rewrites underperform raw refinement, and Appendix E shows that negative rewrites have the lowest preservation rate among all rewrite families.

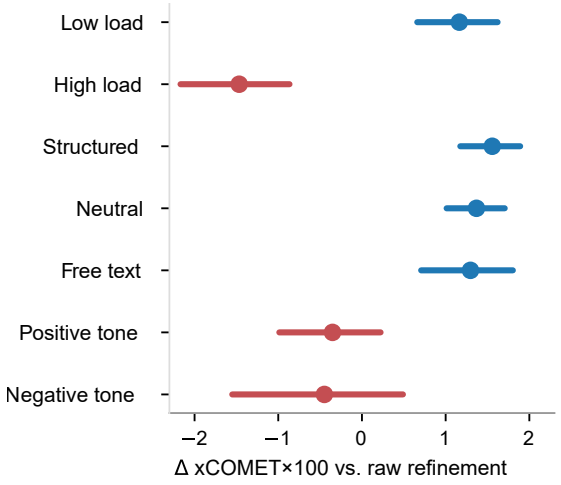


Figure 5: Setting-level xCOMET×100 differences relative to the raw-refinement reference across 42 translator-direction settings.

Condition	Mean	Δ vs. Raw
Raw refinement (reference)	82.13	0.00
Structured	83.69	+1.56
Free-text	83.43	+1.30
Neutral	83.50	+1.37
Positive tone	81.78	-0.35
Negative tone	81.69	-0.45

Table 8: Tone and response-format results under xCOMET×100 on the rewriting data. Raw refinement is shown as a full-matrix reference mean; differences are relative to that reference.

H Two-Round xCOMET Results

We examine xCOMET score changes across two refinement rounds. We use a two-round subset with three translators (Mistral-7B, Qwen2.5-14B, Gemma3-27B), two evaluators, and two directions (English→German, German→Chinese), using 150 sentences each (1,800 examples per condition). In each case, the same evaluator re-critiques the round-one revision and the translator revises again.

Raw feedback scores 81.95 in round one and 78.79 in round two, a change of -3.16 (95% CI $[-3.99, -2.37]$). The corresponding changes are -0.54 for near-identity feedback and -0.94 for low-load feedback. At round two, the differences relative to raw feedback are +0.11 for near-identity feedback and -0.30 for low-load feedback.

This comparison describes round-to-round score changes from different round-one starting points. The final scores are similar across conditions. Table 9 reports results by translator, and Figure 6 shows the aggregate scores.

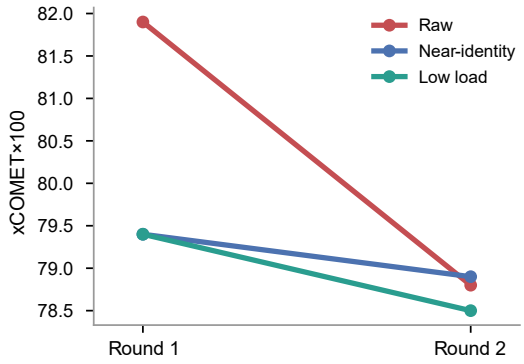


Figure 6: Mean xCOMET×100 scores across two refinement rounds on the English→German and German→Chinese subset. Raw feedback has a higher round-one score, and the conditions reach similar scores in round two. Table 9 reports results by translator.

Translator	Raw	Near-id.	Low load
Mistral-7B	79.7→74.2 (-5.5)	75.5→74.6 (-0.9)	77.1→75.4 (-1.7)
Qwen2.5-14B	82.1→78.9 (-3.2)	78.5→79.3 (+0.8)	78.1→76.9 (-1.2)
Gemma3-27B	84.0→83.2 (-0.8)	84.3→82.8 (-1.5)	83.1→83.1 (+0.1)
All translators	81.9→78.8 (-3.2)	79.4→78.9 (-0.5)	79.4→78.5 (-0.9)

Table 9: Two-round xCOMET$\times$100 results on the English→German and German→Chinese subset. Cells show round-one and round-two scores; parentheses show the change from round one to round two.