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

Anonymous ACL submission

Abstract

In LLM refinement, feedback carries information, but it also acts as an interface that shapes how that information is used: the same evaluator judgment can lead to different revisions when it is organized, compressed, or overloaded with detail. We introduce a controlled protocol for machine translation that fixes the evaluator score, main issue, and recommendation while varying the feedback's form and exposed detail. Across seven models and six translation directions, information load is the most reliable factor: compact feedback outperforms detail-heavy feedback in all 42 translator-direction settings on xCOMET, with the same low-load versus high-load direction supported on a paired MetricX-24 subset. Structured and neutral rewrites show positive effects under xCOMET, but the ranking remains metric-dependent. On a matched two-round subset, raw feedback drifts sharply while compact and near-identity feedback stay stable. The strongest evidence points to compact feedback as a stabilization tool 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: Can controlling feedback form or density reduce multi-round refinement drift? We use evaluator identity as a baseline, with feedback form and density as the main variables of interest.

Contributions. We make three contributions. First, we introduce a controlled MT refinement evaluation protocol and dataset artifacts that separate critic identity from feedback form and exposed detail across seven models and six directions. Second, we show that information load is the most reliable factor in one-round refinement: compact feedback consistently outperforms overloaded

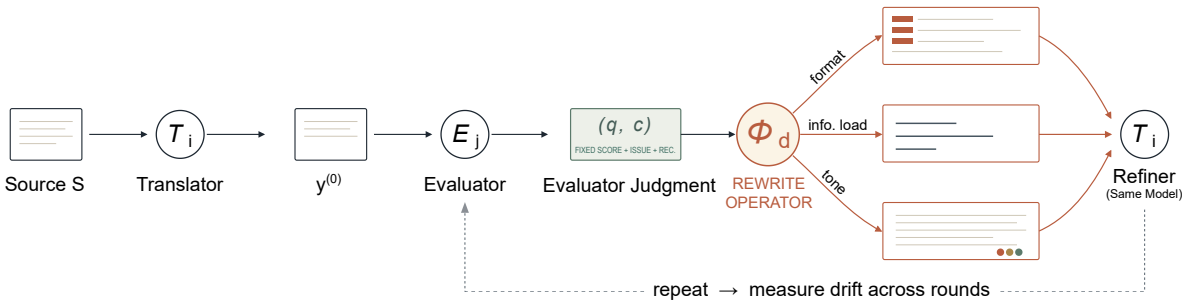

Figure 1: Controlled comparison for feedback presentation. The score, main issue, and recommendation are fixed, while the feedback is rewritten by tone, format, or information load before revision by the same translator.

feedback, with support from both xCOMET and a MetricX-24 subset check.

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). 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?

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.

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 matched 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

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 then scores the translation on a 0 to 100 scale and provides free-form feedback using prompts adapted from GEMBA[1] (Kocmi and Federmann, 2023). The evaluator sees only the source sentence and the candidate translation; no human reference is provided. The original translator receives that score and explanation 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 feedback before it is sent back to the translator; the evaluator score and core evaluative content, namely the main issue and recommendation, are kept fixed.

The matched comparison is simple: for the same source, translator, evaluator, score, main issue, and recommendation, 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 does not preserve all supporting evidence, it preserves the score, main issue, and recommendation. The original translator then revises from the same initial translation under each feedback variant.

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 three dimensions in the map above appear in the final aggregated results. All rewrites are produced by GPT-OSS-20B under prompts that keep scores, named entities, the main issue, and the main recommendation unchanged. Because this could introduce paraphraser-specific artifacts, the results include three validation checks (detailed in Section 5.5): a direct-control condition where the evaluator itself produces the target format, a near-identity rewrite that keeps wording close to the original, and a second-paraphraser rewrite using a different model. The downstream comparison asks whether different formulations of the same core critique change metric-predicted refinement quality.

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× the original character length, while high-load rewrites remain near parity (Appendix D). The intended construct is the amount of detail made available to the revising model, and the results

show that this dimension produces the most robust effects.

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 curate 1,000 English and German sentences from the Hansard corpus[2] and the official records of the Deutscher Bundestag[3], respectively. These sources

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

[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

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: Matched 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.

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. Controlled rewriting uses a matched 29,400-comment subset: 100 source items for each translator–evaluator–direction cell ($7\times7\times6$). All rewrite conditions use this same subset.

Upon acceptance, 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: 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.

general website content remains subject to the Bundestag legal notice at https://www.bundestag.de/en/service/legal_notice.

[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.

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. Throughout the paper, we report xCOMET multiplied by 100 for readability. Fine-grained one-round rankings may be xCOMET-specific. We treat them as robust only when the MetricX-24 spot check or multi-round analyses point in the same direction. Our independent metric check uses MetricX-24-Hybrid in reference-free mode (Juraska et al., 2024).

4.4 Controlled Feedback Rewriting

The base-refinement condition uses the evaluators' original free-form feedback. 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 5.5 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]

In every condition, the original translator performs the final revision, keeping the refiner fixed within each matched 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.

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

[6] The rewrite logs cover the 29,400-comment subset: 88,200 tone rewrites and 58,800 rewrites for each two-value family.

Implementation details. We use deterministic local generation for translation, evaluation, rewriting, and refinement, and run xCOMET through the COMET checkpoint interface. Appendix A reports token limits, batch sizes, software settings, the bootstrap seed, and release artifacts. These implementation choices support reproducibility and are held outside the experimental variables.

5 Results

We first show that raw feedback is an unstable baseline (§5.1). We then test the main control, information load, and check whether the result survives MetricX, length adjustment, and edit-size analysis (§5.2). We next report the secondary surface-form effects (§5.3), then evaluate repeated-critique stability on the English → German / German → Chinese two-round subset (§5.4), and finally separate rewrite-mechanism artifacts from the main pattern (§5.5).

5.1 Raw Feedback Is an Unstable Interface

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; Appendix Table 7 reports the full base matrix.

Table 3 summarizes the default refinement loop, where evaluators provide their original free-form critiques. 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 oracle best cross-model critic, a diagnostic upper bound selected retrospectively for each translator, 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 oracle best cross 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?

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 3: Mean refinement quality (xCOMET$\times$100) across the six language directions. Self feedback uses the same model as translator and evaluator; oracle best cross-critic selects the strongest non-identical evaluator for each translator after averaging over directions. Bolded values exceed the original translation baseline.

5.2 Information Load Is the Most Reliable One-Round Factor

With evaluator judgment held fixed, different feedback variants lead to different revisions. Table 4 reports the conditions by manipulation family.

The central one-round result is information load. Low-load feedback beats high-load feedback in all 42 settings with a mean gap of 2.63 xCOMET$\times$100 (95% paired-bootstrap CI [2.26, 3.05]). High-load rewrites are the worst condition overall (80.67). The pattern points to a penalty for overloading. The optimal style varies across settings.

The independent MetricX check points in the same direction. On a stratified 1,008-sentence subset, sampled as 24 source items from each of the 42 translator-direction settings and kept paired across conditions, low-load feedback exceeds high-load feedback by 0.54 points (95% paired-bootstrap CI [0.32, 0.76]).[7]

Length and edit size leave the load result intact. Information load captures the amount of supporting points and qualifiers exposed to the reviser. The length-adjusted regression confirms that the penalty persists after controlling for character count. Within each load condition, feedback length has negligible association with quality ($|\rho| < 0.03$), and controlling for feedback length leaves a large high-load penalty ($\beta = -4.02$, 95% CI [-4.41, -3.62], $p < 0.001$). A revision-size check on the available English→German treatment logs gives the same interpretation: mean normalized character edit distance from the initial translation is 0.23 for low-load feedback, 0.22 for a detail-heavy high-load realization, and 0.30 for raw feedback (Appendix Table 9). This weakens a pure fewer-edits explanation, since the compact condition edits about as much as a detail-heavy condition on this subset. Edit-distance logs are currently available for this English → German direction; extending the check to all directions remains future work.

Dimension	Condition	Avg.	Δ vs. Raw
Core comparison: information load			
Load	**Low load**	**83.30**	**+1.16**
	High load	**80.67**	**-1.47**
	Raw refinement	82.13	0.00
Secondary xCOMET summaries: surface form			
Format	Structured	83.69	+1.56
	Free-text	83.43	+1.30
	Raw refinement	82.13	0.00
Tone	Neutral	83.50	+1.37
	Positive tone	81.78	-0.35
	Negative tone	81.69	-0.45
	Raw refinement	82.13	0.00
Reference	Original translation	83.46	+1.33

Table 4: Mean controlled-rewrite quality (xCOMET$\times$100) on the matched subset ($n = 29{,}400$ per condition). The load block is the core comparison; format and tone are secondary xCOMET summaries. Deltas are relative to raw refinement within each family.

5.3 Surface-Form Effects Are Secondary

Format and tone improve over raw refinement under xCOMET, but the gains are smaller than the load contrast and not confirmed by MetricX.

Structured feedback improves over raw refinement by +1.56 xCOMET$\times$100 (38 of 42 settings; 95% paired-bootstrap CI [+1.19, +1.88]), and neutral improves by +1.37 (36 of 42 settings; [+1.03, +1.70]). However, the MetricX subset leaves structured feedback close to raw ($\Delta = -0.04$, 95% CI [-0.15, 0.07]), so we treat these gains as positive under xCOMET but not yet confirmed across metrics. Figure 2 summarizes setting-level deltas relative to

[7] We report the negated MetricX raw prediction so that higher values are better; the sign flip leaves paired differences unchanged.

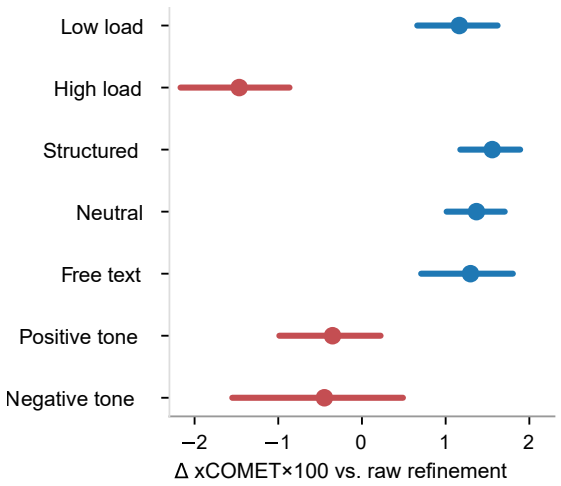

Figure 2: Setting-level effects relative to raw refinement on the matched subset. Points show mean Δ xCOMET×100 over 42 translator–direction settings; bars are 95% paired-bootstrap CIs. MetricX supports only the low-load versus high-load contrast.

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.

Harder settings show larger best-case rewrite gains, suggesting that feedback rewriting is most useful where the raw baseline is already fragile.

5.4 Repeated Critique Stability

We test whether feedback form affects stability across rounds. We construct a matched 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 starts higher in round one (81.95) but drops most sharply by round two (78.79), with an aggregate drift of -3.16 (95% CI $[-3.99, -2.37]$). Near-identity feedback drifts by only -0.54; low-load feedback drifts by -0.94. By round two the three conditions converge: the final differences are small (+0.11 for near-identity vs. raw, -0.30 for low-load vs. raw).

The result supports a stabilization claim on this subset: controlled feedback reduces multi-round drift, even when it does not improve single-round quality. Table 5 gives the translator-level breakdown and Figure 3 visualizes the aggregate pattern.

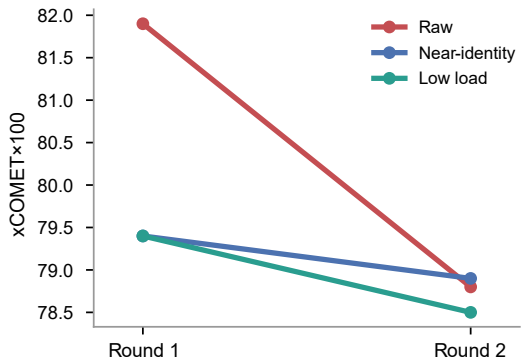

Figure 3: Aggregate two-round stability on the matched English→German/German→Chinese subset. Raw feedback starts higher in round one but drifts most sharply by round two; near-identity and low-load presentations stay more stable. Table 5 gives the translator-level breakdown.

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 5: Two-round stability on the matched slice (xCOMET×100). Cells show R1→R2 with drift in parentheses; aggregate drift reduction vs. raw is +2.62 for near-identity and +2.21 for low load.

5.5 Validation Checks for the Rewrite Mechanism

Three concerns could undermine the main findings: (a) the paraphraser itself may create the observed effects, (b) any rewrite may generically improve feedback, and (c) rewrites may fail to preserve the intended critique. We address each on a validation subset of 5,400 examples per condition (three translators, three directions, two evaluators, 300 sentences each).

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. Concern: any rewrite might improve feedback by smoothing the text. Check: a near-identity rewrite keeps the feedback close to its original wording. Result: near-identity falls 1.87 points below matched raw refinement (95% CI $[-2.27, -1.45]$), so generic rewriting alone does not improve one-round quality. This makes near-identity useful as a drift control (§5.4): it reduces two-round drift despite weak round-one quality, separating stabilization from generic rewriting gains.

Rewrite validity. Concern: controlled rewrites may fail to preserve the intended issue and recommendation. Check: Appendix E reports an adjudicated stratified 84-pair sample. Result: rewrites preserve the original issue and recommendation in 81/84 cases each and introduce no new unsupported issue in 80/84 cases; weaker cases concentrate on negative-tone rewrites.

Because this validation is sampled rather than exhaustive, we interpret all controlled-rewrite results as downstream behavioral effects under approximate preservation, not as effects under perfect semantic invariance.

6 Discussion

Feedback as an interface. The results suggest that feedback acts as an interface for evaluator judgment, not a neutral channel. 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.

Why compact feedback is safer. Extra detail may create competing revision targets, especially when supporting explanations include caveats or secondary issues. Detail-heavy feedback may also encourage over-editing by giving the reviser more surface cues to act on. A third possibility is distributional: compact feedback resembles direct instruction more closely than long evaluator prose, which may better match instruction-tuned models. The revision-size data in Appendix Table 9 weaken a pure over-editing account, since low-load and high-load yield similar edit sizes. Disentangling these mechanisms requires targeted baselines such as length-matched variants or score-only feedback.

Practical implications. Multi-agent and self-refinement systems typically pass feedback through as if its form were neutral. Our results suggest adding a lightweight compression layer between evaluator and reviser that preserves the score, main issue, and recommendation while suppressing unnecessary detail. This is especially valuable for fragile settings and repeated critique, where near-identity rewriting contributes little to one-round quality but substantially reduces multi-round drift.

7 Conclusion

Feedback presentation is an interface around evaluator judgment. Raw one-round refinement is brittle, with critic identity serving as an important baseline. The more actionable finding is that feedback form and exposed detail change downstream behavior even when evaluator judgment and core critique are fixed. The most robust one-round factor is compactness versus overload; structured and neutral rewrites are positive under xCOMET but metric-dependent.

The clearest practical result comes from the two-round setting. On the evaluated matched two-direction subset, raw feedback degrades by 3.16 xCOMET$\times$100, whereas near-identity rewrite and low-load controls are much less volatile. In the evaluated high-resource MT settings under automatic metrics, compact feedback is the safest default for fragile refinement. Together with the rewrite validation checks, these results support fixed-judgment rewriting as a protocol for studying feedback form and exposed detail as controllable variables in LLM refinement.

Limitations

Claim scope. The paper supports a controlled behavioral claim: feedback form and exposed detail can change LLM revision under fixed

evaluator judgment and preserved core critique. The strongest one-round claim is the load/overload contrast; structured, neutral, and free-text rankings are secondary and metric-dependent. The drift-reduction claim applies to the evaluated English→German/German→Chinese matched two-round subset.

Load construct. Information load is implemented by changing exposed supporting detail while preserving the score, main issue, and recommendation. Length-adjusted and edit-distance checks reduce simple confound explanations, yet information load, length, and distributional match to instruction-style prompts remain intertwined. The current experiments also lack length-matched low/high-load variants and score-only, issue-only, or 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 used as a spot check. 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. A venue-specific checklist and any required LLM-use disclosure will be submitted with the final anonymized package.

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.

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.

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

The appendix follows the same order as the main paper. Appendix B summarizes checkpoint metadata, Appendix C gives the active prompt families, Appendix D records the controlled dimensions, Appendix E reports rewrite validation checks and illustrative cases, and Appendix F visualizes the full translator and evaluator refinement grid.

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, 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. The anonymized release will include the executable scripts, prompts, package versions, and cluster job logs; the exact GPU allocation varied across jobs and is treated as a release artifact outside the experimental variables.

B Model Selection

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

Table 7 reports the base translation quality matrix summarized in Section 5.1.

C Prompt Design

This appendix reports only the prompt families that remain active in the current experiment suite: translation, evaluation, refinement, and controlled feedback rewriting. Prompt families not used in the current reported results are omitted.

D Controlled Dimensions

Table 8 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:

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.

D.2 Prompts for Controlled Feedback Rewriting

The reported rewriting experiments manipulate one dimension at a time: tone, response format, or information load. In every case, the prompt instructs

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.

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 7: Base translation quality (xCOMET$\times$100) by translator and language pair.

the paraphraser to preserve the score, named entities, main issue, and main recommendation while changing only the requested presentation variable.

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:

E Validation Checks for Controlled Rewrites

We run two supporting checks over the logged rewrite outputs to see whether the paraphrasing stage behaves in the way the main experiments assume. These checks are intended as diagnostic and prioritization tools, without serving as proof that every rewrite perfectly preserves the original critique.

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

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 8: 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.

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.

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 are consistent with the paper's broader conclusion that feedback organization and information load are powerful control levers, while the rewrites remain controlled approximations with imperfect meaning preservation.

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 caveat, 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× 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.

Revision-size check. Table 9 reports a simple edit-distance check on the available English→German treatment logs. Low-load revisions are smaller than raw revisions, but a detail-heavy high-load realization is similarly conservative in edit size. This pattern weakens the explanation that low-load succeeds solely by making fewer edits.

Condition	Subset	Mean	Median
Raw feedback	En→De logs	0.305	0.205
Low load	En→De logs	0.227	0.112
High load (detail-heavy)	En→De logs	0.225	0.097

Table 9: Normalized character edit distance between the initial translation and revised translation on available English→German treatment logs (n = 10,000 paired outputs per condition). Lower values indicate smaller revisions.

F Results of Base Refinements

Figures 4 and 6 report mean Δ xCOMET×100 relative to the original translation for each translator and evaluator pair. The diverging color scale is centered at zero, so positive cells indicate helpful refinement and negative cells indicate degradation.

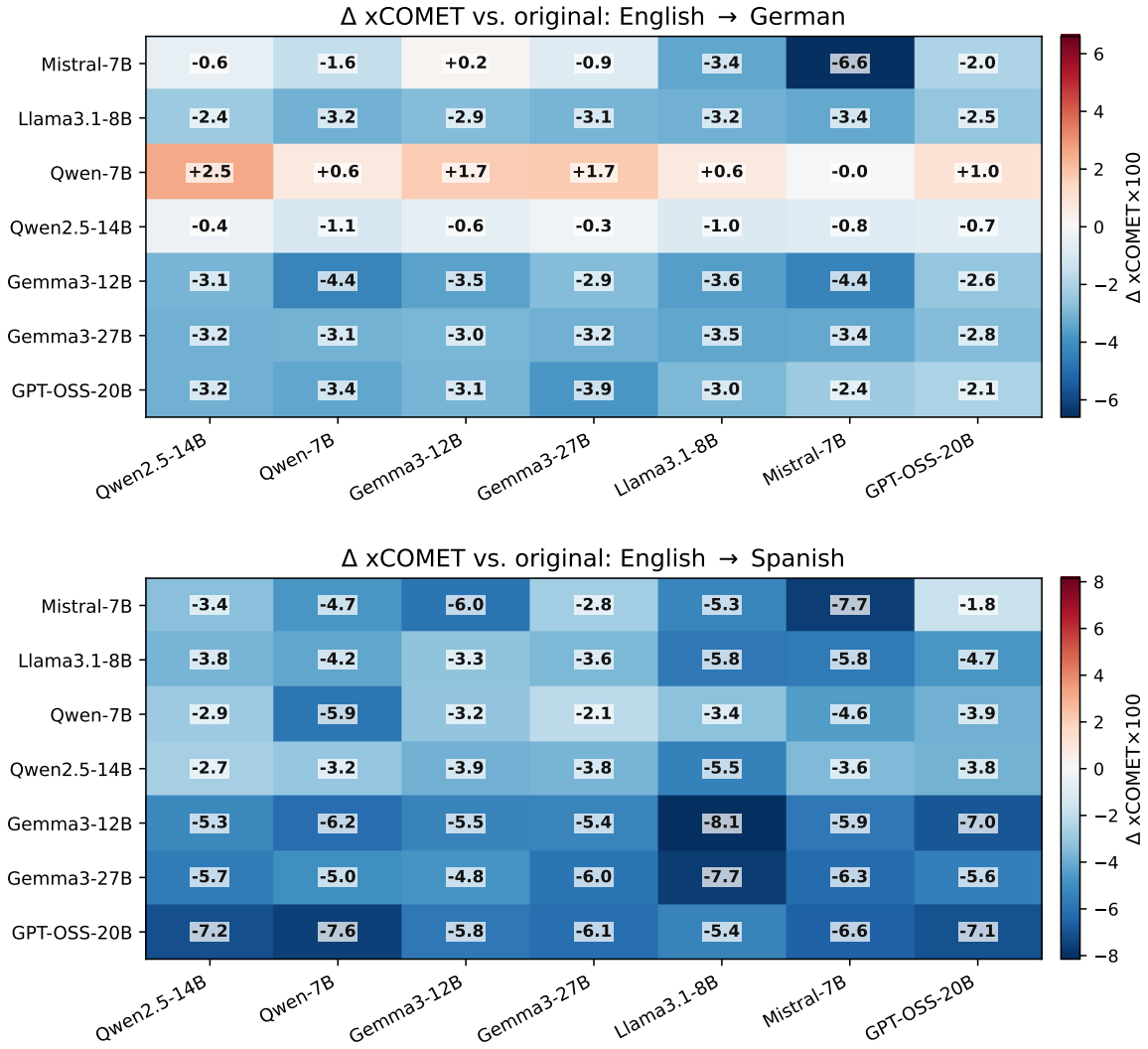

Figure 4: Appendix heatmaps for English-source directions. The top panel shows English→German and the bottom panel shows English→Spanish. Positive cells indicate that refinement helps relative to the original translation; negative cells indicate degradation.

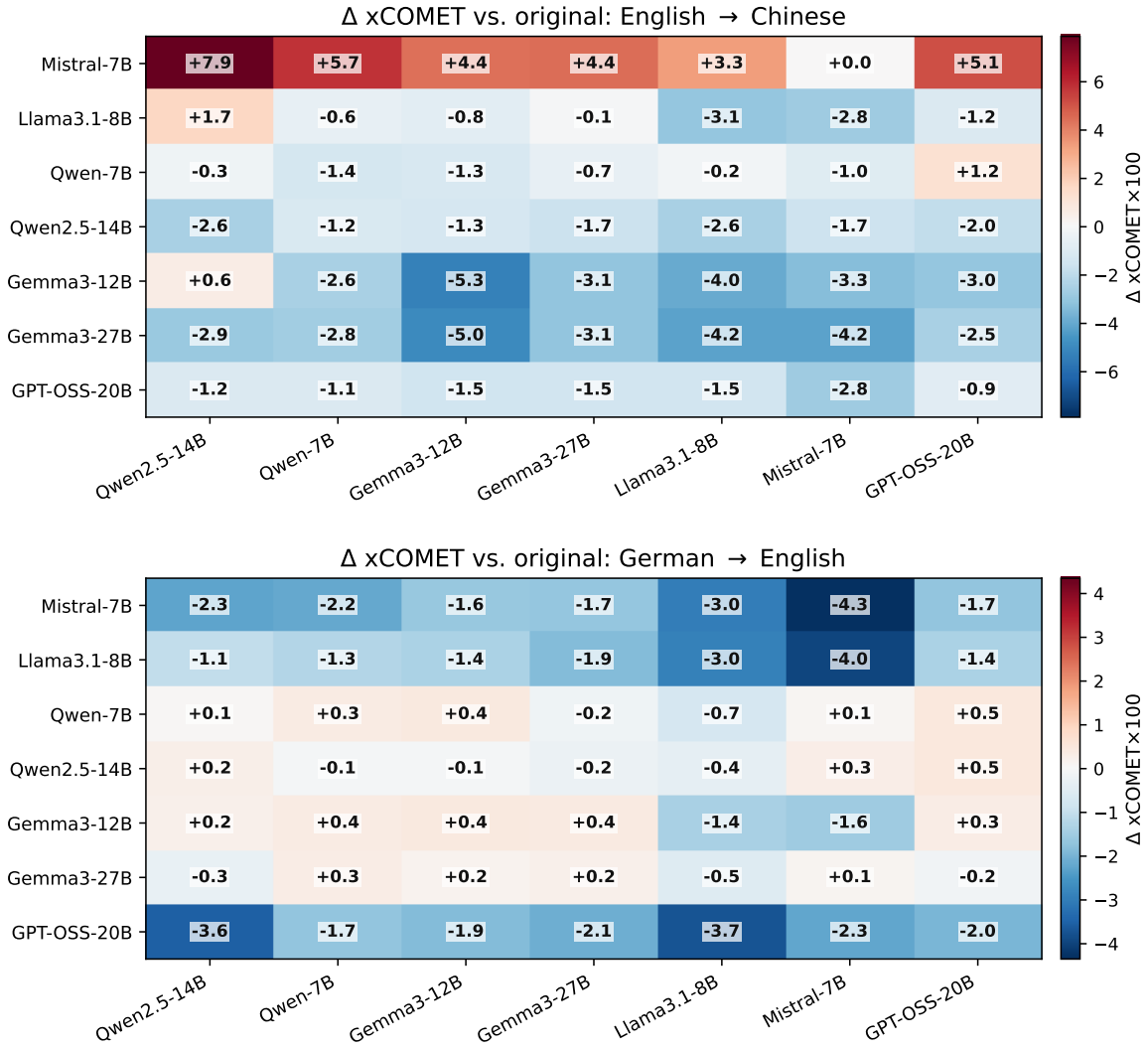

Figure 5: Appendix heatmaps for English→Chinese (top) and German→English (bottom). These panels contrast one harder Chinese-target setting with one easier English-target setting under the same translator and evaluator grid.

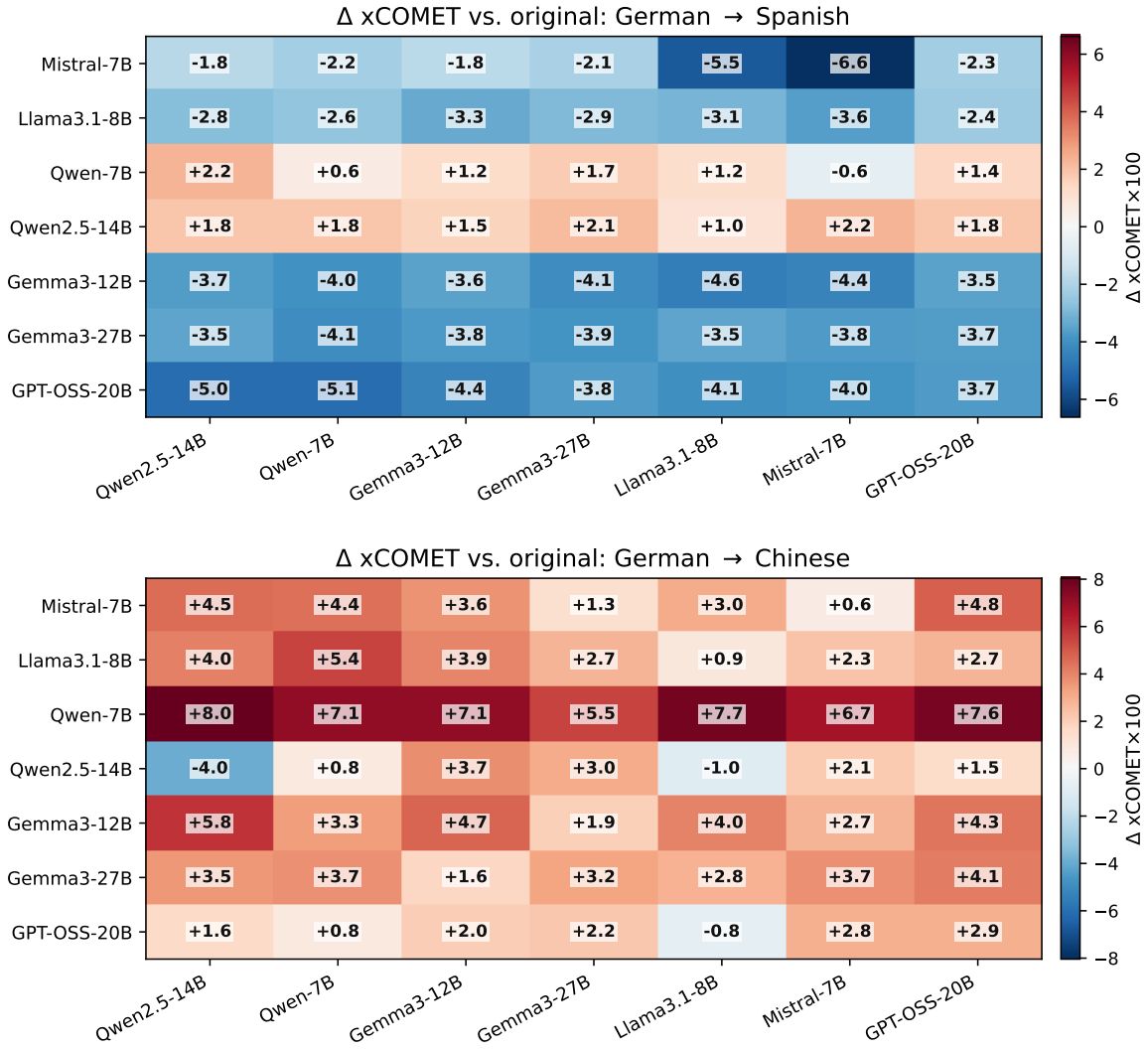

Figure 6: Appendix heatmaps for the remaining German-source directions: German→Spanish (top) and German→Chinese (bottom). Across all appendix panels, warmer colors indicate mean improvement and cooler colors indicate mean degradation relative to the original translation.