

Should We be Pedantic About Reasoning Errors in Machine Translation?

Calvin Bao, Marine Carpuat
 Department of Computer Science
 University of Maryland
 College Park, MD 20742, USA
 csbao@umd.edu

Abstract

Across seven language pairs (English $\rightarrow$ Spanish, French, German, Mandarin, Japanese, Urdu, Cantonese), we show that large reasoning models (LRMs) exhibit reasoning errors during machine translation. We quantify these errors using an automated annotation protocol that labels reasoning steps as source-sentence misaligned, model-hypothesis misaligned, or reasoning-trace misaligned. We then intervene on candidate reasoning errors with weak-to-strong edits (hedging, removal, re-reasoning), and compare these interventions with new reasoning traces generated based on human translation references. Although reasoning errors can be detected with high precision in some language settings (e.g. Urdu), detection precision varies across languages and falls to 50% in Spanish. Correcting error candidates in reasoning traces also fails to reliably improve translation quality, providing evidence that reasoning traces are not consistently faithful to the mechanisms underlying translation in LRMs.

1 Introduction

Recent reasoning methods, including chain-of-thought prompting and large reasoning models (LRMs), have achieved strong performance across a variety of NLP tasks. In machine translation (MT), however, their benefits have been inconsistent. Prior work suggests that reasoning is most effective when it is explicitly structured around translation decisions, for example through knowledge retrieval (Briakou et al., 2024), multi-stage translation pipelines (He et al., 2024; Briakou et al., 2024; Li et al., 2025; Feng et al., 2025), or specialized agent workflows (Li et al., 2025; Wang et al., 2025). Yet recent studies show that native reasoning, an LRM’s default reasoning process without translation-specific guidance, can underperform direct translation (Li et al., 2026; Rajaei et al., 2026). This raises the question: what goes wrong in reasoning for translation?

We investigate reasoning traces as an object of study in their own right. Figure 1 illustrates why. Translating a FLORES sentence into Spanish, an LRM could state that *plumilla* is the correct Spanish translation of *a chick’s down*, however, *plumilla* in Spanish would mean a quill or pen nib, not down feathers. The model commits to this decision and propagates it into the output, producing a fluently inadequate translation. Notably, nothing in this trace is internally inconsistent, and the output faithfully reflects what the trace decided. By the standard test of whether stated reasoning influences the final answer (Lanham et al., 2023), this reasoning trace is faithful but incorrect. Cases like this arise across all seven language pairs we study (English $\rightarrow$ Spanish, French, German, Mandarin, Japanese, Urdu, Cantonese). To study these phenomena systematically, we develop a structured annotation framework for detecting reasoning errors (§4), carried out by an LLM-as-a-judge validated against bilingual human judgments (§4.4), and apply targeted interventions to the detected errors (§5) as diagnostic probes to assess whether reasoning traces faithfully reflect the translation process. Altogether, we address the following research questions:

- How frequent are reasoning errors in MT reasoning traces? (RQ1)

SOURCE (EN)	OUTPUT (ES)
The researchers suggested that, even though this is the tail of a young dinosaur, the sample shows adult plumage and not a chick's down.	Los investigadores sugirieron que, aunque esta es la cola de un dinosaurio joven, la muestra muestra plumaje adulto y no la plumilla de un polluelo.
TRACE (excerpt)	
... "the sample shows adult plumage and not a chick's down" – "la muestra muestra plumaje adulto y no la plumilla de un polluelo." "Plumilla" is the correct term for down feathers in Spanish. Putting it all together, the translation should be accurate and natural.	
ERROR a chick's down → la plumilla de un polluelo [†] [†] plumilla ≈ quill / pen nib; the correct term is plumón or plumaje juvenil.	

Figure 1: A reasoning trace from Qwen3-8B maps a chick's down to the Spanish *plumilla* (quill/pen nib) instead of *plumón* (down feathers). The error propagates to the output.

- Do candidate reasoning errors as detected by an LLM-as-a-judge causally impact the translation? (RQ2)

2 Background

Reasoning for machine translation. Reasoning has been consistently shown to yield large gains in tasks with verifiable answers, such as mathematics, code, and multi-step logical inference (Wei et al., 2022; Kojima et al., 2022; Yang et al., 2025; Liu et al., 2026), and these gains extend multilingually, though unevenly: models reason more accurately in English than in the task language (Shi et al., 2023; Barua et al., 2026). However, the benchmarks commonly used to demonstrate reasoning progress are still focused on tasks like multi-step question-answering (Shi et al., 2023; Lightman et al., 2023). How “native reasoning” may improve machine translation tasks is still under-explored.

Some work has recently explored how reasoning can be applied successfully in machine translation, finding that structured and human-like translation strategies that can be decomposed into multiple steps tends to work better than when generated natively by the model. He et al. (2024) induces the model to consider translation-related knowledge before producing the final translation. Similarly, Briakou et al. (2024) proposes a pipeline consisting of pre-drafting (knowledge retrieval), drafting, refinement, and proofreading stages. Feng et al. (2025); Li et al. (2025) introduce frameworks that consist of agents that collaborate in order to generate the best final translation, primarily consisting of distinct candidate generation and verification roles.

Meanwhile, Li et al. (2026); Rajaei et al. (2026) both find that “native reasoning” underperforms when compared to the direct translation approach. To dive deeper into why, Ye et al. (2026) consider the idea that some complex translation inputs may not require more reasoning for task success. They explore this question with TwT, a difficulty-adaptive framework that learns how to selectively reason for machine translation, and find that applying CoT to all inputs costs more tokens on “easy” translation inputs and still underperforms direct translation according to standard MT quality metrics (BLEU (Papineni et al., 2002), COMET and COMET-Kiwi (Rei et al., 2022)) in English, Chinese directions. However, in more difficult, out-of-domain settings, they find that both general and domain-specific CoT reasoning outperform even supervised fine-tuned (SFT) models in translation. Pei et al. (2026) consider reasoning trace content in extremely low-resource MT settings (Xibe and Chintang). They analyze reasoning traces generated with prompt designs that elicit linguistic and grammatical reasoning, and find that reliable (e.g., **not erroneous**) sentence-specific traces improve translation when supplied as in-context learning examples. Altogether, these results suggest that reasoning helps translation when it is applied selectively and when

its claims are reliable. However, neither is guaranteed when an LRM generates its own reasoning without carefully supplied external input, which is the setting we examine.

Faithful reasoning and trace sensitivity. Prior work on chain-of-thought faithfulness asks whether a model’s stated reasoning influences the final output (Lanham et al., 2023; Turpin et al., 2023; von Recum et al., 2026). This work typically intervenes on reasoning chains, by truncating, paraphrasing, injecting errors, and measuring robustness of model responses to these perturbations by assessing whether the model’s answer changes in the intended way. When it does not, the model is deemed unfaithful to the input reasoning, as canonically defined (Jacovi & Goldberg, 2020).

Our faithfulness tests are most closely related to work on reasoning faithfulness, which asks whether a model’s output is causally attributable to its intermediate reasoning (Lanham et al., 2023). With multiple-choice Question-Answering benchmarks, Lanham et al. (2023) directly intervenes on Chain-of-Thought reasoning by perturbing traces and measures model faithfulness by the extent to which the model’s accuracy changes. Interestingly, they find that larger models are often less faithful: the final answer persists even substantial trace perturbations, suggesting the model is not always faithful to the reasoning in all settings.

Paul et al. (2024) sharpens this intuition via causal mediation analysis, modeling the trace R as a mediator between input X and output Y . They decompose the total effect into a **direct effect** ($X \rightarrow Y$, bypassing R) and an **indirect effect** ($X \rightarrow R \rightarrow Y$), measured by intervening on both the input (swapping in a counterfactual question X_1) and the trace (swapping in reasoning R_1 generated for X_1). Across several reasoning-oriented benchmarks (commonsense & math) they find that indirect effects decrease as reasoning traces get larger. They also show that none of the tested models exhibit consistently high indirect effects, indicating that LLMs do not reliably use their stated reasoning when producing answers. More recently, Chen et al. (2025) study faithfulness in the extended thinking traces of LRMs, and find that models frequently omit mention of hints they relied on, with this type of faithfulness decreasing on harder reasoning-oriented questions like GPQA (Rein et al., 2023). Zhou et al. (2024) further shows that models are vulnerable to inaccurate noisy rationales in-context, with noisy reasoning degrading downstream accuracy. Together, these observations motivate our setting in MT: if models inconsistently condition on their own reasoning in tasks with verifiable answers, it is an open question whether this extends to open-ended NLG tasks like translation. Translation also affords a richer analysis than in tasks with verifiable answers, as a trace may commit to many separate lexical or syntactic decisions which may not surface in the output.

3 Experimental Setup

Here, we describe the data, languages, and models used to generate reasoning traces.

Data and languages. We test on 7 language directions: $\text{en} \rightarrow \{\text{es}, \text{ja}, \text{fr}, \text{de}, \text{yue}, \text{ur}, \text{zh}\}$, sourced from the FLORES dataset, a benchmark of Wikipedia-derived sentences with professional translations into over 100 languages (Goyal et al., 2022). We restrict to out-of-English directions as a controlled design choice. LLMs are generally stronger on $\text{English} \rightarrow X$ than on $X \rightarrow \text{English}$ (Zhu et al., 2024), and this also fixes the reasoning language to English, which prior work finds yields stronger task performance than reasoning in the target language (Barua et al., 2026; Shi et al., 2023). Our target language choices span high-resource and lower-resource targets (ur, yue), allowing us to examine whether reasoning errors and their detectability vary with target-language resource level. These design choices place the models in relatively favorable translation conditions, to minimize the risk that the reasoning errors we observe are related to adversarially difficult settings. We sample $n=100$ source sentences for all language directions, using the same source sentences across all languages.

Large reasoning models. We study four open-weight models with native reasoning modes, spanning two families at two scales: Qwen3-8B and Qwen3-14B (Yang et al., 2025), and Ministral-3-8B and Ministral-3-14B (Liu et al., 2026). All four are used from Huggingface in

	Output Valid (O✓)	Output Invalid (O✗)
Reasoning Valid (R✓)	Potentially aligned: trace has no detected issues and the translation has no detected issues.	Silent translation error: translation is wrong, but the trace does not contain detectable issues.
Reasoning Invalid (R✗)	Right answer, wrong reason: translation is correct, but the trace contains a detected issue. This may cause reduced trust in good translations.	Wrong answer, wrong reason: translation is incorrect, and trace has semantic divergences. Highlighting this may help users reject imperfect translations.

Table 1: A 2×2 matrix framing the validity of reasoning and translation outputs. In this work, we analyze the detection of invalid outputs, conditioned on invalid reasoning R✗.

thinking mode with no translation-specific prompting, the setting prior work identifies as underperforming direct translation (Li et al., 2026; Rajaei et al., 2026).

4 Detecting Reasoning Errors

In this section, we focus on answering **RQ1 (How frequent are reasoning errors in MT traces?)** by describing how we define reasoning errors, how we detect them, and how we validate their nature with annotators that are expert language specialists.

4.1 Defining Reasoning Errors

We propose 3 consistency categories connected to distinct threads of prior work in evaluation for NLG, reasoning verification, and CoT faithfulness, which are referred to in Table 2. Inspired by the evaluation axes of the general reasoning trace evaluation taxonomy proposed by Lee & Hockenmaier (2025), we operationalize this for machine translation. INPUT-REASONING and REASONING-INTERNAL instantiate the *factuality* and *validity/coherence* categories. The closest analogy to REASONING-OUTPUT relates to both the *utility* axis of Lee & Hockenmaier (2025) and to chain-of-thought faithfulness (Lanham et al., 2023).

4.2 Detection Protocol

We describe an automatic annotation protocol: we start by decomposing each trace into individual decision steps by sentence tokenization. We then check the consistency of each reasoning step with the prefix reasoning trace, the source input and the model output. We choose to annotate at the reasoning step-level for both practical reasons and inspired by work in Chain-of-Thought faithfulness, which frequently operates at the step-level to support targeted interventions (Prasad et al., 2023; Jacovi et al., 2024). An issue is marked **ERROR** if the incorrect decision persists through the trace. An issue is marked **PERSISTENT** if the incorrect decision persists through the trace, and **FIXED-LATER** if the trace later corrects it. We discard **FIXED-LATER** issues, which separates decisions that persist to the output from exploratory reasoning steps the model may self-correct within-trace. We refer to the surviving flags as **error candidates**.

Implementation We implement the protocol with an LLM-as-a-judge (GPT-5.2). The judge receives three inputs: the source sentence, the sentence-tokenized reasoning trace, and the final translation. For each issue it identifies, it returns a structured annotation with: (i) the category (INPUT-REASONING, REASONING-OUTPUT, or REASONING-INTERNAL); (ii) the index into the trace sentence; and (iii) a short rationale.

We sample $k = 5$ judgments per trace with temperature set to 0.4. For each independently sampled run, we then aggregate by category and reasoning step index and keep only the issues with at least 3 judgments referring to it (majority vote, $\geq 3/5$).

Category	Question	Example	Connection to Lee & Hockenmaier (2025)
INPUT-REASONING	Does the trace make claims supported by the source sentence?	Source: "The bug crashed the whole system." Step 3: "The word bug here means a biological organism; I'll translate it as <i>bacteria</i> ." Explanation: The trace selects the wrong sense of <i>bug</i> ; the source context (software) does not support this interpretation.	In the taxonomy of Lee & Hockenmaier (2025), this corresponds most directly to factuality , especially groundedness with respect to the source. Analogous to groundedness in NLG evaluation (Honovich et al., 2022; Maynez et al., 2020): just as a summary can hallucinate content absent from a source document, a trace can hallucinate translation equivalences.
REASONING-OUTPUT	Does the output reflect the decisions stated in the trace?	Step 8: "The phrase 'pretending to listen' maps naturally to <i>haciendo como si estuviera escuchándonos</i> ." Step 9: "I'm confident in this phrasing. Finalizing the translation." Output: "Ella estaba <i>haciéndose pasar por si estuviera escuchándonos</i> ." Explanation: The output uses a different phrase than the one committed to in Steps 8–9.	In the taxonomy of Lee & Hockenmaier (2025), this aligns most closely with utility : whether a reasoning step actually contributes to the final answer. Prior work on CoT faithfulness asks a related question: whether reasoning reflects the model's actual process (Lanham et al., 2023; Jacovi & Goldberg, 2020).
REASONING-INTERNAL	Is the trace internally consistent and coherent?	Step 5: " <i>bank</i> means <i>el banco</i> ." Step 8: "Wait—since it is next to 'river', it should be <i>la orilla</i> ." Step 9: "Alright, going with <i>el banco</i> as decided earlier." Explanation: Step 9 reverts to Step 5 without acknowledging the correction in Step 8.	In the taxonomy of Lee & Hockenmaier (2025), this aligns with both coherence and validity . Coherence is concerned with whether a reasoning step's preconditions are established consistently, while validity is concerned with whether the step contains a logical inconsistency. Draws also from detection of self-contradictory reasoning in LLM outputs (Liu et al., 2024).

Table 2: Error categories for reasoning errors in machine translation, grounded in the reasoning-trace evaluation taxonomy of Lee & Hockenmaier (2025). Highlighted spans in the example indicate the error span to focus on.

4.3 Detected Reasoning Errors

Table 3 reports error candidates detected across the four models. Error candidate rates are high at 8B scale, as the judge detects at least one candidate in 72.8% of samples for Qwen3-8B and 80.1% for Ministral-3-8B, at 3.03 and 3.55 candidates per sample respectively. Qwen3-14B flags candidate errors in 48.3% of samples at 0.95 per sample, a significant reduction over Qwen3-8B. The gap between directions persists, however: en→ur remains the worst direction at both scales, for both Qwen3 and Ministral-3.

Model	Lang. pair	# samples	# w/ errors	Err. rate	Avg. steps/trace	Avg. errors/sample
Qwen3-8B	en→es	100	56	56.0%	39.12	1.40
	en→zh	100	54	54.0%	32.69	1.86
	en→ja	100	86	86.0%	40.88	3.51
	en→fr	100	73	73.0%	42.31	2.81
	en→de	100	81	81.0%	45.64	2.94
	en→ur	100	97	97.0%	38.26	6.40
	en→yue	100	63	63.0%	36.25	2.30
	<i>All</i>	700	510	72.9%	39.31	3.03
Qwen3-14B	en→es	100	29	29.0%	32.59	0.49
	en→zh	100	31	31.0%	30.96	0.52
	en→ja	100	47	47.0%	32.38	0.88
	en→fr	100	39	39.0%	35.08	0.69
	en→de	100	56	56.0%	37.20	0.85
	en→ur	100	90	90.0%	30.55	2.20
	en→yue	100	46	46.0%	31.66	1.01
	<i>All</i>	700	338	48.3%	32.92	0.95
Ministral-3-8B	en→es	100	60	60.0%	21.59	2.24
	en→zh	100	65	65.0%	17.63	3.18
	en→ja	100	93	93.0%	22.20	3.93
	en→fr	100	74	74.0%	16.84	2.72
	en→de	100	69	69.0%	20.34	3.39
	en→ur	100	100	100.0%	16.40	5.68
	en→yue	100	100	100.0%	17.12	3.74
	<i>All</i>	700	561	80.1%	19.72	3.55
Ministral-3-14B	en→es	100	66	66.0%	82.18	1.23
	en→zh	100	75	75.0%	88.72	1.53
	en→ja	100	83	83.0%	91.68	1.81
	en→fr	100	71	71.0%	88.69	1.51
	en→de	100	73	73.0%	96.86	1.57
	en→ur	100	91	91.0%	81.22	2.98
	en→yue	100	81	81.0%	82.28	1.70
	<i>All</i>	700	432	61.7%	87.80	1.67

Table 3: Error candidates detected by the LLM-as-a-judge on FLORES devtest. Up to 100 examples were sampled per translation direction; # Samples reports traces successfully generated and judged. Aggregate averages are weighted by the number of judged traces.

4.4 Human validation

We conduct a validation with bilingual human annotators to compute the precision of the automatic LLM-as-a-judge results on error detection, using the Qwen3-8B translations for English→Spanish, and English→Urdu. For each language pair, we sample 30 FLORES devtest examples. Each devtest example yields multiple reasoning steps, and this subset produces 189 error candidates in English→Urdu and 58 in English→Spanish (Table 4) to annotate independently by 3 annotators. Each example is a triplet (x, t, y) , where x is the English source sentence, t is the model reasoning trace, and y is the final translation. Each sampled example from FLORES yields one item-level annotation for translation quality. In addition, each sample yields one or more error candidates in t , to be annotated for error validity. All items are annotated independently by three bilingual annotators per language pair. We recruit bilingual language specialists rather than general crowd workers, following conclusions that judgments involving error analysis by experts is more reliable than judgments from non-specialists (Freitag et al., 2021).

Phase 1: item-level output validity. In the first phase, annotators see only the source sentence x and the final translation y . They answer the following question: **Does the**

	English→Urdu	English→Spanish
<i>Study size</i>		
Samples annotated	30	30
Issues annotated	189	58
<i>Translation correctness (majority vote per sample)</i>		
YES (no major error)	0/30 (0.0%)	24/30 (80.0%)
NO (major error)	28/30 (93.3%)	5/30 (16.7%)
TIE	2/30 (6.7%)	1/30 (3.3%)
<i>Reasoning-error validation (majority vote per issue)</i>		
YES only	176/189 (93.1%)	29/58 (50.0%)
YES + BORDERLINE	179/189 (94.7%)	31/58 (53.4%)
NO	3/189 (1.6%)	22/58 (37.9%)
TIE	7/189 (3.7%)	5/58 (8.6%)
<i>Error reflected in output? (among validated errors only)</i>		
YES	41.1%	84.0%
NO	56.6%	16.1%
UNSURE	2.3%	0.0%
<i>Annotator confidence</i>		
Mean confidence	90.5%	91.4%
On YES judgments	94.0%	88.9%
On NO judgments	72.8%	97.3%

Table 4: Human validation for Qwen3-8B in English→Urdu and English→Spanish directions. Translation correctness and reasoning-error validation are reported at the sample and issue level respectively, using majority vote across three annotators. Error reflection rates are computed over annotations labeled YES or BORDERLINE.

translation preserve the important meaning of the source sentence? Annotators choose one of three labels: OK: no critical meaning error is present; minor fluency or stylistic differences are acceptable; NOT OK: the translation contains at least one meaning error that could mislead a reader; UNSURE: I cannot confidently judge the item.

Phase 2: issue-level validation of detector-flagged trace spans. In the second phase, annotators validate individual issues proposed by the detector. For each error candidate, annotators see the source sentence x , the final translation y , and the sentence-tokenized reasoning trace t , with the highlighted reasoning error candidate in-context. Specific questions for the annotators can be found in Appendix A.2.

First, we report **overall detector precision**, defined as the proportion of flagged issues is validated as a reasoning error by the corresponding majority human judgment. Second, we compare reasoning-level validation with Phase 1 item-level OK/NOT OK judgments.

English→Urdu. Across a sample of 30 English→Urdu (86 sample-level annotations and 541 issue-level annotations), at the instance-level, 28/30 items receive a majority translation fidelity error with a translation NOT OK judgment. In terms of what the detector flagged as reasoning errors, they are validated by humans. Precision is 93.1% averaged across 3 annotators, with high confidence (94.0%). 41.1% of detected reasoning errors are considered reflected in the final output, which shows that some reasoning traces impact downstream output. Source misinterpretations are also more likely to propagate to the final Urdu translation: among error annotations, the reflection-YES rate is 44.9% for source misinterpretation, compared to 20.2% for internal contradiction. This suggests that the detector is useful when it detects these errors, and that these errors propagate to poor English → Urdu outputs. All together, we consistently show that the English→Urdu translations traces are

not semantically faithful to the reasoning, and bilingual annotators agree with the resulting errors.

English→Spanish. The English→Spanish results (Table 4) contrast sharply. 80.0% of samples receive a majority YES on translation correctness: the Spanish translations are largely faithful to the English source. The judge flags 58 reasoning errors in these samples, but annotators confirm only 50.0% of them, with 37.9% rejected as non-errors. Crucially, annotators are more confident when rejecting the detected reasoning errors (97.3% confidence on NO) than when confirming them (88.9% on YES). Among the free-text responses, annotators note cases where the detector objects to a reasonable translation choice, flags an ambiguity the source itself does not resolve (e.g., whether “garage” means a workshop or a literal garage), or claims a grammatical rule that may be pedantically stricter than what a native speaker would expect. Among the errors annotators do confirm, the output propagation rate is high: 84.0% are decided to have been reflected in the final output, compared to 41.1% in Urdu. This suggests that the detector’s threshold in Spanish is high enough that only the consequential errors persist.

Summary. Detector precision varies with the target language: 93.1% of error candidates are confirmed in English→Urdu, compared to 50.0% in English→Spanish. The two directions also differ in output quality (0/30 Urdu outputs judged acceptable, 24/30 Spanish), so precision is higher where translations are worse. Consequentiality shows the reverse pattern: 84.0% of confirmed errors are reflected in the Spanish output, compared to 41.1% in Urdu. Spanish thus falls mainly in the *right answer, wrong reason* cell of Table 1, and Urdu in *wrong answer, wrong reason*.

5 Intervening on Reasoning Traces

In this section, we focus on answering **RQ2 (Do detected reasoning errors causally impact the translation?)** by describing our set of interventions to the reasoning trace to probe if they correct the detected error and if they make a difference in translation quality.

We follow the perspective of causal mediation analysis (Paul et al., 2024): if modifying a trace step changes the output, that step has causal influence on the translation. To port it to machine translation, we consider inputs (x, t, y) where t and y are generated in a model’s thinking mode, and we intervene on the trace by modifying $t \mapsto t'$ and injecting t' as context to the model with thinking disabled to obtain a translation y' .

5.1 Types of Interventions

We study six interventions with increasing expected strength. We construct controlled transformations $t \mapsto t'$ with just a modification on the detected reasoning error for the first three, and we consider the last three “oracle”-level interventions, since they construct t' guided by the reference, instead of the detected issue. We summarize what these interventions do in the list below, with full implementation details in Appendix A.4.

- **Hedging:** prepend hedging language to the reasoning error (e.g., “*I’m not sure, but...*”).
- **Removal:** remove the detected reasoning error from the trace.
- **Re-reasoning after removal:** remove the error span and regenerate the remainder of the trace from that point onward, allowing the model to re-derive downstream reasoning.
- **Hindsight:** use the model to generate a reasoning trace to translate from the source to the reference, then re-play with that reasoning trace.
- **Oracle-K:** K phrase-level hints covering the full source–reference alignment.
- **Oracle-1:** a single hint mapping the error-relevant source phrase to its reference translation.

5.2 Evaluation Setup

Resolution rate. We compute resolution by re-running the same LLM-as-a-judge detector (§4.2) on the intervened triplet (x, t', y') under identical settings ($k=5$, $T=0.4$, majority vote $\geq 3/5$), and checking whether the reasoning step is flagged again. We consider it resolved if it is not flagged again.

Translation evaluation. For the intervention experiments, we evaluate translation quality deltas with COMET (Rei et al., 2022), a reference-based neural metric that correlates well with human judgments of MT quality, using the Unbabel/wmt22-comet-da checkpoint against FLORES references. We report segment-level scores averaged within each direction. Δ COMET is measured by subtracting the averaged segment-level scores of the intervention from the average segment-level scores of the baseline replay.

5.3 Intervention results

Table 5 reports error resolution rates and COMET deltas across seven language pairs and two models. We see three patterns emerge.

Lightweight edits are largely ineffective for Qwen3-8B but not for Ministral-3-8B. For Qwen3-8B, **hedging** and **removal** resolve fewer than 7% of detected issues across all language pairs. This suggests that Qwen3-8B does not condition strongly on individual trace steps when generating under replay. On the other hand, under our replay setup, Ministral is more sensitive to trace edits; the same lightweight edits resolve 41–60% of issues.

Re-reasoning resolves issues but degrades translation quality. **Re-reasoning after removal** achieves resolution rates of 50–63% for Qwen3-8B, second only to hindsight. However, this comes at a cost: COMET scores drop substantially (up to -0.16 for en→ja), so while the specific detected error might disappear, the regenerated trace suffix introduces new problems that degrade translation quality. The pattern is consistent across both models and most language pairs.

Hindsight is the strongest intervention. **Hindsight**, or replaying with a new trace generated mapping source to the reference, achieves the highest resolution rates (72–84% for Qwen3-8B, 83–88% for Ministral-3-8B) and the most consistently positive COMET deltas when compared to the other intervention types.

Sensitivity to lightweight edits with model scale. Both 14B models resolve more candidates under two local corrections (hedging and removal interventions) than 8B models, but the gaps differ. Qwen goes from $\leq 6.7\%$ at 8B to 26.5 to 45.5% at 14B, whereas Ministral-3-8B and 14B both remain sensitive to the interventions. However, we see that re-reasoning continues to resolve candidates while lowering COMET (3.4 to 11.3 for Qwen3-14B).

6 Discussion & Conclusion

We summarize and discuss implications of reasoning errors in machine translation.

Frequency of errors (RQ1) Our human validation reveals a different story with Urdu and with Spanish. For English→Urdu, annotators confirm 93.3% of the translations as not being faithful semantically to the source (Table 4), and 93.1% of reasoning errors being valid errors. Here, the errors the LLM-as-a-judge detects are real, and annotators understand why they are errors with high confidence, and the translations themselves are poor. For English→Spanish, annotators confirm only 50% of the detected errors, and rate up to 80% of the translations as correct. We interpret this setting involving LLM-as-a-judge’s pedantry, flagging reasoning steps with imprecise but pedantic grammatical rules, objecting to reasonable interpretations of ambiguous phrases, or potentially diving into ungrounded stylistic preferences as errors. Notably, annotators are more confident when rejecting

detected errors (97.3%), than when confirming them (88.9%), suggesting that non-errors are perceived unambiguously as non-errors.

Additionally, we observe that among confirmed errors, 84.0% are consequential errors for Spanish versus only 41.1% for Urdu. This suggests that the LLM-as-a-judge’s standards are high enough that only consequential errors tend to propagate. In Urdu, many confirmed errors exist in the reasoning, but are either partially corrected downstream, or may manifest as different surface errors in the output, making the mapping between reasoning step to output more difficult to track.

Implications on reasoning error detection Where models qualitatively produce worse translations, like in English→Urdu, the LLM-as-a-judge reliably identifies genuine reasoning errors that a human auditor would agree with. For translation directions like in English→Spanish, where the model translates more effectively, the same judge is overly pedantic – roughly half of the detected errors do not align with what expert humans would consider reasonable reasoning errors. This means the signal is noisy, and any downstream use of the judge’s output (for example, for filtering training data, or self-correcting potential errors), should account for this in practice.

Faithfulness of MT reasoning (RQ2) The intervention results (§5.3) show that targeted trace edits guided by detected errors rarely improve translation quality. Only interventions that discard or replace large portions of the trace (re-reasoning, hindsight, oracle) achieve more substantial resolution rates, but with slight degradation of COMET. The human evaluation helps explain this pattern, and the explanation differs by language pair. For English→Spanish, where annotators reject roughly half the judge’s flags, the failure of lightweight interventions is unsurprising: you cannot improve a translation by correcting an error that was not consequential in the first place. For English→Urdu, however, the errors are more genuinely reasonable and the translations are poor, yet targeted edits still fail.

Towards better reasoning for translation. From prior research, we see evidence that native reasoning for MT often underperforms direct translation for general machine translation tasks (Li et al., 2026; Rajaei et al., 2026). We take a deeper dive and explore whether there are errors that might emerge from inconsistencies in the reasoning, finding that their properties differ across language pairs. We find qualitatively for machine translation that reasoning can manifest as free-form and over-reasoned, and results suggest that trace decisions can be decoupled from decisions made to produce the translation output, leading to faithfulness issues.

Acknowledgments

We thank the members of the UMD CLIP lab for their valuable advice and feedback on earlier iterations of this work, especially Sarah Wiegrefe, Dayeon Ki, and Hyojung Han.

Limitations

Scope of models and languages. Our experiments are conducted on four open-weight reasoning models (Qwen3-8B, Qwen3-14B, Ministral-3-8B, Ministral-3-14B). It is unclear whether the patterns we observe hold for other reasoning models. We test seven language pairs, all with English as the source. Translation from non-English sources, or between two non-English languages, may surface qualitatively different reasoning errors.

Limited human validation coverage. We conduct bilingual human validation for only two of our seven language pairs (English→Urdu and English→Spanish), which were chosen to contrast a lower-resource and a high-resource setting. The remaining five pairs lack direct human validation of the judge’s precision, for cost reasons. Given that we observe different precision rates between Urdu (93.1%) and Spanish (50.0%), it is likely that precision

varies across the unvalidated pairs as well, and we cannot determine in which direction without additional annotation. Moreover, human validation only covered outputs from Qwen3-8B, as it was a LRM that exhibited many reasoning errors. It remains to be seen how the reasoning traces from the 14B models, which exhibit fewer candidate reasoning errors, would be considered by our annotators.

Ethics Statement

Intended use. The work presented here aims to characterize where and how reasoning traces may fail in machine translation, not as a proposal for improving translations in production. Our intervention experiments show that targeted trace edits can resolve some detected errors but do not reliably improve overall translation quality. We do not recommend the interventions described here as a post-editing pipeline without further quality assurance, as they may introduce new errors not present in the original translation.

LLM disclosure. LLMs were used in several different ways in the preparation of this research. For paper writing, Anthropic Claude was used for drafting, refining language (grammar and style), and creating LaTeX tables and figures. For coding, Claude was used to refine scripts for model generation and data analysis. Finally, for research ideation, [Elicit \(2023\)](#) was used to guide and help summarize literature review. All scientific claims, ultimate experimental design decisions, and interpretations are the authors' own.

References

- Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. Long chain-of-thought reasoning across languages, 2026. URL <https://arxiv.org/abs/2508.14828>.
- Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1301–1317, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.123. URL <https://aclanthology.org/2024.wmt-1.123/>.
- Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think, 2025. URL <https://arxiv.org/abs/2505.05410>.
- Elicit. Elicit: The ai research assistant, 2023. URL <https://elicit.com>.
- Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. TEaR: Improving LLM-based machine translation with systematic self-refinement. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3922–3938, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.218. URL <https://aclanthology.org/2025.findings-naacl.218/>.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021. doi: 10.1162/tac1_a_00437. URL <https://aclanthology.org/2021.tac1-1.87/>.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tac1_a_00474. URL <https://aclanthology.org/2022.tac1-1.30/>.

- Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246, 2024. doi: 10.1162/tacl_a_00642. URL <https://aclanthology.org/2024.tacl-1.13/>.
- Or Honovich, Roei Aharoni, Jonathan Herzig, et al. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 3905–3920, 2022.
- Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL <https://aclanthology.org/2020.acl-main.386/>.
- Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roei Aharoni, and Mor Geva. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4615–4634, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.254. URL <https://aclanthology.org/2024.acl-long.254/>.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiuūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL <https://arxiv.org/abs/2307.13702>.
- Jinu Lee and Julia Hockenmaier. Evaluating step-by-step reasoning traces: A survey. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1789–1814, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.94. URL <https://aclanthology.org/2025.findings-emnlp.94/>.
- Weiya Li, Junjie Chen, Bei Li, Boyang Liu, Zichen Wen, Nuanqiao Shan, Xiaoqian Liu, Anping Liu, Huajie Liu, Hu Song, and Linfeng Zhang. Tactic: Translation agents with cognitive-theoretic interactive collaboration, 2025. URL <https://arxiv.org/abs/2506.08403>.
- Zihao Li, Shaoxiong Ji, and Jörg Tiedemann. Test-time scaling of reasoning models for machine translation. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2902–2917, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.133. URL <https://aclanthology.org/2026.eacl-long.133/>.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.
- Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan,

- Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yasmine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. *Minstral 3*, 2026. URL <https://arxiv.org/abs/2601.08584>.
- Ziyi Liu, Soumya Sanyal, Isabelle Lee, Yongkang Du, Rahul Gupta, Yang Liu, and Jieyu Zhao. Self-contradictory reasoning evaluation and detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3725–3742, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.213. URL <https://aclanthology.org/2024.findings-emnlp.213/>.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040/>.
- Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.882. URL <https://aclanthology.org/2024.findings-emnlp.882/>.
- Renhao Pei, Yihong Liu, Sampo Pyysalo, Hinrich Schütze, and Shaoxiong Ji. Reasoning over grammar: Can synthetic linguistic reasoning traces enhance low-resource machine translation?, 2026. URL <https://arxiv.org/abs/2606.03782>.
- Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. ReCEval: Evaluating reasoning chains via correctness and informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10066–10086, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.622. URL <https://aclanthology.org/2023.emnlp-main.622/>.
- Sara Rajae, Sebastian Vincent, Alexandre Berard, Marzieh Fadaee, Kelly Marchisio, and Tom Kocmi. Unlocking reasoning capability on machine translation in large language models, 2026. URL <https://arxiv.org/abs/2602.14763>.

- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névoul, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wmt-1.52. URL <https://aclanthology.org/2022.wmt-1.52/>.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof qa benchmark, 2023. URL <https://arxiv.org/abs/2311.12022>.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=fr3wGCK-IXp>.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don’t always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS ’23, Red Hook, NY, USA, 2023. Curran Associates Inc.
- Alexander von Recum, Leander Gierbach, and Zeynep Akata. Are reasoning llms robust to interventions on their chain-of-thought?, 2026. URL <https://arxiv.org/abs/2602.07470>.
- Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. DRT: Deep reasoning translation via long chain-of-thought. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6770–6782, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.351. URL <https://aclanthology.org/2025.findings-acl.351/>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- Yongshi Ye, Biao Fu, Chongxuan Huang, Yidong Chen, and Xiaodong Shi. Translation with thought: Difficulty-adaptive reasoning via reinforcement learning for multi-domain machine translation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30339–30370, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1400. URL <https://aclanthology.org/2026.acl-long.1400/>.

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales?, 2024. URL <https://arxiv.org/abs/2410.23856>.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2765–2781, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.176. URL <https://aclanthology.org/2024.findings-naacl.176/>.

A Appendix

A.1 Detecting reasoning errors

As discussed in the main paper, we detect reasoning errors in three categories. We show prompt templates.

```
You are a bilingual auditor for machine-translation reasoning traces.

You will analyze a SOURCE sentence, the model's TRACE (reasoning while translating), and
the OUTPUT (final translation).

Your task is to detect reasoning errors in three categories:
1. INPUT_TRACE: Trace statements not supported by SOURCE, or proposing incorrect
   translation semantics (e.g., hallucinated facts, wrong word meanings)
2. TRACE_OUTPUT: Trace decisions that don't match the OUTPUT (e.g., trace says "X" but
   output has "Y")
3. TRACE_INTERNAL: Contradictions, circular reasoning, or incoherent statements within
   the trace itself

IMPORTANT RULES:
- The trace will be sentence-tokenized. Reference issues by sentence index (0-indexed).
- All quotes must be EXACT substrings (copy-paste) from the provided text.
- Be strict but fair - minor rephrasing or stylistic choices are not errors.

Output ONLY valid JSON matching this schema:
{
  "has_issues": bool,
  "summary": str, // One sentence summary of trace quality
  "issues": [
    {
      "category": "INPUT_TRACE" | "TRACE_OUTPUT" | "TRACE_INTERNAL",
      "trace_sentence_idx": int,
      "trace_quote": str, // Exact substring from trace
      "source_quote": str | null, // Relevant source quote if applicable
      "output_quote": str | null, // Relevant output quote if applicable
      "rationale": str // 1-2 sentence explanation
    }
  ]
}

SOURCE:
{source}

TRACE (sentence-indexed):
[0] {trace_sentence_0}
[1] {trace_sentence_1}
...

OUTPUT:
{output}
```

A.2 Annotation interface for bilingual annotators

See Figure 2 for a flow of the annotation protocol.

Annotator questions

1. Does the highlighted reasoning span contain an error?

Response options:

- YES – the highlighted reasoning contains a clear mistake (wrong meaning, wrong claim, a contradiction, etc.),
- NO – The highlighted reasoning is correct or a reasonable interpretation,
- BORDERLINE – the reasoning is imprecise or debatable but not clearly wrong.

2. How confident are you in your judgment above?

Response options:

- CONFIDENT — I am sure of my answer.
- SOMEWHAT CONFIDENT — I think my answer is right but it requires specialized knowledge I’m not fully certain about.
- NOT CONFIDENT — This requires detailed language knowledge I’m not fully certain about.

3. Is this reasoning step reflected in the final translation?

Response options:

- YES – I see the decision or claim reflected in the final translation,
- NO – The final translation seems to not have the decision or claim reflected in it,
- NOT APPLICABLE – I cannot determine the connection.

4. Which of the following categories apply to this highlighted reasoning step? (multi-select)

Response options:

- SOURCE MISINTERPRETATION – The reasoning step is making a wrong claim or decision about the source. The reasoning misunderstands or misrepresents something in the English source (e.g., wrong word meaning, hallucinated information not in the source),
- INTERNAL CONTRADICTION – The reasoning step contradicts its context. contradicts itself or contains circular/incoherent logic,
- NO ISSUE – This reasoning step makes sense to me in this context.
- OTHER / UNSURE.

A.3 Qualitative examples of consequential reasoning errors

Detected reasoning error	Reasoning err.?	Conseq.?	Source	Translation
CONSEQUENTIAL ERROR IN SPANISH				
“Have refused to repatriate” is “han rechazado la repatriación.” (agent of the refusal inverted)	Yes (2/3)	Yes (3/3)	... which, the statement said, European nations have refused to repatriate.	... los cuales, indicó el comunicado, han rechazado la repatriación por parte de las naciones europeas.
The source is: “ Daniuš said, “Right now we are doing nothing.” (name corrupted while restating the source)	Yes (2/3)	Yes (3/3)	Daniuš said, “Right now we are doing nothing...”	Daniuš dijo: “En este momento no estamos haciendo nada...”
I’ll use “a través de los años 70 tardíos” for a more natural flow. (calque of “the late 1970s”; idiomatically a finales de los años 70)	Yes (2/3)	Yes (3/3)	... normalizing US-China relations through the late 1970s ...	... la normalización de las relaciones entre EE.UU. y China a través de los años 70 tardíos ...

continued on next page

Table 6 (continued)

Detected reasoning error	Reasoning error?	Conseq.? Source	Translation
CONSEQUENTIAL ERROR IN URDU			
Wait, "mice" is "موچے" in Urdu. (string never appears in the output)	Yes (3/3)	Yes (3/3)	"We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.
Also, "added" is translated as "اورہ کرتے ہیں" which is correct. (string never appears in the output)	Yes (3/3)	Yes (3/3)	..., "he added.
"USA Gymnastics supports the United States Olympic Committee's letter" – "امریکی اولمپک کمیٹی کا خط سپورٹ کرتا ہے" (subject dropped)	Yes (3/3)	Yes (2/3)	USA Gymnastics supports the United States Olympic Committee's letter ...
... which in Urdu would be "... کی تقریر سے اتفاق کرتے ہیں کہ USOC ہم" (تقریر means speech)	Yes (2/3)	Yes (2/3)	We agree with the USOC's statement that the interests of our athletes ...
The term "selections" could be translated as "منتخابات" (elections) or "منتخب کرنا" (to select). (unsupported gloss)	Yes (3/3)	Borderline (2/3)	During the 1976 selections he advised Carter on foreign policy ...

Table 6: Detected reasoning errors on Qwen3-8B reasoning traces, along with their human annotations from the study. "Reasoning error?" is the majority vote on Q1; "Consequential?" means the decision is reflected in the final translation (Q2).

A.4 Intervention Implementation

System message (default). You are a careful machine translation assistant.

Task instruction (default). If source and target language codes are known, the default is: Translate the following {source language} text into {target language}. Return only the translation. (with fallbacks when only the target or neither is known).

Follow-trace instruction (default). Use the reasoning trace when deciding on the translation.

User message template. Interventions that replay with an (possibly edited) trace all wrap the model input in this pattern (then passed through the model chat template if available):

```
{task_instruction}

Source:
{source}

Reasoning trace:
{edited_trace}
{optional_additional_notes_block}

{follow_trace_note}
Return only the final translation.
```

If extra_notes are supplied (Oracle methods), the block is:

```
Additional notes:
- {note_1}
- {note_2}
...
```

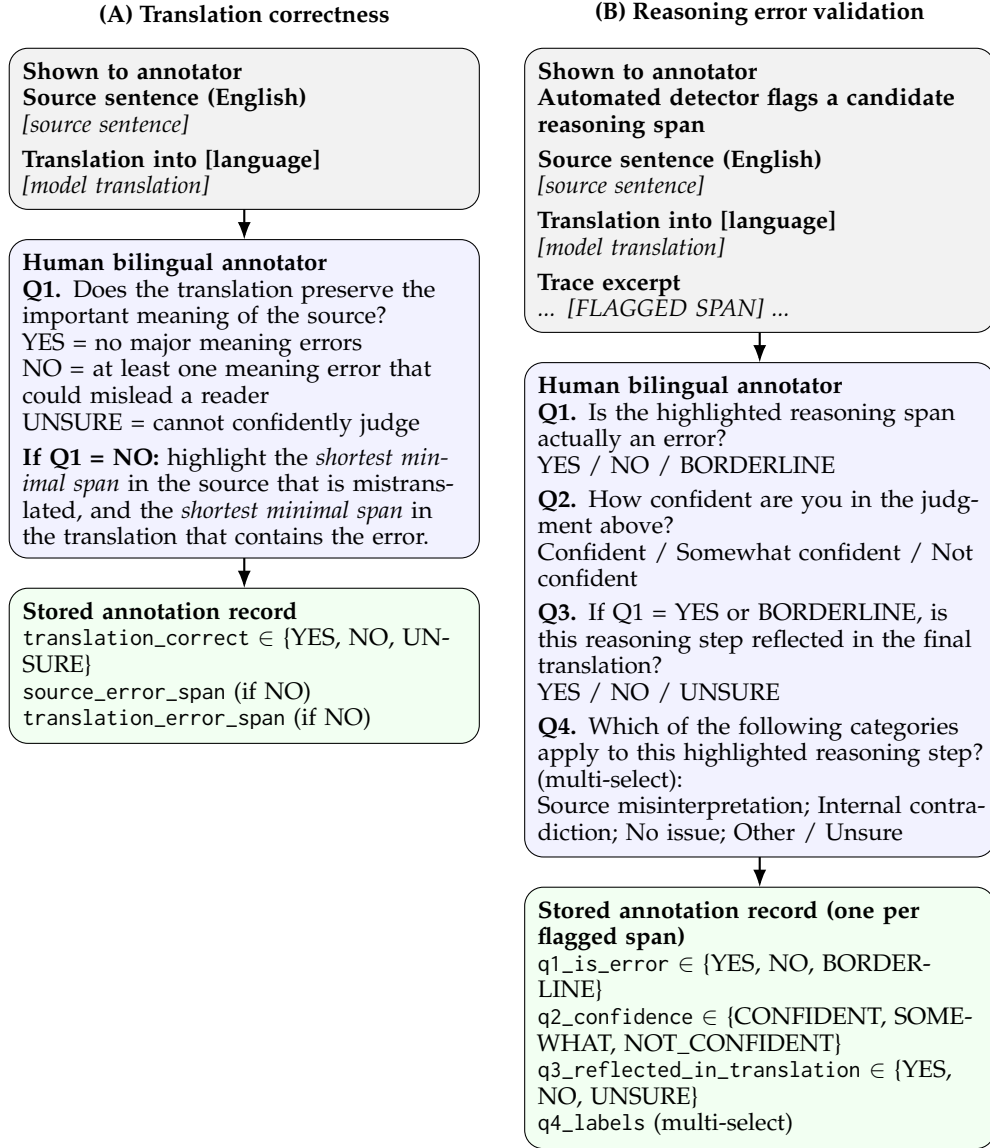


Figure 2: Bilingual human annotation protocol used to validate final translation correctness and detected reasoning errors. Part (A) is applied once per translation. Part (B) is applied once per detected reasoning error. Q4 is multi-select. For cost reasons, we run this protocol on English → Spanish and English → Urdu.

Replay generation uses greedy decoding (do_sample=False, beam width 1). The trace is *not* placed in native “thinking” channels for these replays; only the re-reasoning intervention forces a continuation inside the model’s reasoning format (see below).

A.4.1 Localizing the edited span

For hedging, removal, and re-reasoning, we find a character span for the issue in the original trace:

- 1: if issue has non-empty trace_quote then
- 2: find trace_quote in trace (fuzzy / normalized match)
- 3: else
- 4: use sentence boundaries and trace_sentence_idx

5: **end if**

6: expand to the **minimal whole sentence** that contains that span (locate_issue_edit_span), so edits remove or hedge full sentences when possible

If no span can be located, hedging, removal, and re-reasoning are skipped for that issue.

A.4.2 Hedging

Let s be the sentence text at the edit span. If s is non-empty and does not already begin (case-insensitive) with maybe, possibly, perhaps, or it may be, replace s by:

Possibly, but this should be verified against the source: $\langle s \rangle$

Otherwise s is unchanged. The modified trace is replayed with the shared template (Section A.4).

A.4.3 Removal

The sentence-level span from Section A.4.1 is deleted from the trace; consecutive blank lines are collapsed. The shortened trace is replayed with the same user template as the baseline (no extra notes).

A.4.4 Re-reasoning after removal

Let t be the original trace and $[i, j)$ the edit span. The *prefix* is $p = t_{:i}$ (text strictly before the removed span), with blank lines collapsed.

The user-visible instruction (before chat templating) is:

```
{task_instruction}

Source:
{source}

A problematic reasoning step was removed here. Reconsider the source carefully from this
point onward and do not rely on the removed unsupported step. Target issue: {issue.
rationale}
You MUST continue reasoning internally from the provided starting point, but return only
the final translation.
```

A.4.5 Hindsight (reference-guided trace, then replay)

Hindsight uses the same replay model with native reasoning enabled.

Step 1 – synthesize a trace toward the reference. Prompt:

```
{task_instruction}

Source: {source}
Reference translation: {reference}

Think step-by-step about how to translate the source to match the reference.
Analyze key phrases, idioms, and grammatical structures.
Then produce the final translation.
```

Step 2 – replay. The extracted trace t' is substituted into Reasoning trace: in the shared replay template (Section A.4) with no extra notes. One replay is stored per record (not per issue); fix-judging, if enabled, is still evaluated per targeted issue against that single output.

A.4.6 Direct oracles (Oracle-1 and Oracle-K)

Oracle interventions keep the **original** model trace and add natural-language hints derived from the judge/issue object (not from an automatic word aligner). Each issue yields one multi-line note `build_oracle_note(issue)` with bullet lines, in order when the fields are present:

- Problematic trace snippet: {trace_quote}
- Relevant source quote: {source_quote}
- Original output quote: {output_quote}
- Why it is problematic: {rationale}
- Use the source sentence to avoid carrying this error into the final translation.

- 1: **Oracle-1:** for each targeted issue, call `from_trace` with `record.trace` and a single extra note:
- 2: Oracle correction for one identified issue: + newline + `build_oracle_note(issue)`
- 3: **Oracle-K:** once per record (shared across all targeted issues on that record), call `from_trace` with `record.trace` and extra notes:
- 4: Oracle corrections for all identified issues:
- 5: followed by `build_oracle_note(·)` for every targeted issue on that record, in order

Both use the shared replay user template; generation is the same as for baseline replay (Section A.4).

	Qwen3-8B		Ministral-3-8B		Qwen3-14B		Ministral-3-14B	
Intervention	Rate $\uparrow$	ΔC $\uparrow$	Rate $\uparrow$	ΔC $\uparrow$	Rate $\uparrow$	ΔC $\uparrow$	Rate $\uparrow$	ΔC $\uparrow$
en-de	<i>candidates: 149 / 116 / 85 / 146</i>							
hedging	1.3	0.00	41.4	-0.10	32.9	+0.01	45.9	+0.01
hindsight	83.9	+0.88	88.8	+0.01	91.8	+6.74	93.0	+6.60
oracle-1	31.5	+0.38	67.2	+0.06	62.4	+0.30	61.0	+0.14
oracle-k	30.2	+0.25	66.4	+0.27	70.6	+0.47	63.0	+0.08
removal	3.4	+0.08	50.9	+0.06	36.5	-0.01	47.9	0.00
re-reason	50.3	-0.24	68.9	-7.15	63.1	-6.29	65.8	-6.28
en-es	<i>candidates: 82 / 101 / 49 / 117</i>							
hedging	0.0	0.00	54.5	+0.03	28.6	-0.01	51.3	+0.10
hindsight	81.7	-0.39	87.1	-0.13	81.6	+8.37	85.8	+8.59
oracle-1	20.7	-0.09	82.2	+0.23	65.3	-0.04	79.5	+0.32
oracle-k	23.2	-0.06	82.2	+0.67	65.3	-0.13	83.8	-0.04
removal	6.1	-0.10	61.4	+0.08	26.5	+0.03	49.6	+0.10
re-reason	51.2	-1.21	70.3	-7.77	56.2	-3.37	54.9	-6.24
en-fr	<i>candidates: 149 / 137 / 69 / 128</i>							
hedging	6.0	-0.12	48.9	+0.01	34.8	+0.05	49.2	-0.05
hindsight	76.5	+1.19	88.3	+0.55	91.3	+6.41	90.8	+6.43
oracle-1	32.2	+0.02	75.2	+0.10	65.2	+0.38	63.3	+0.20
oracle-k	38.3	+0.12	74.5	+0.30	63.8	+0.39	63.3	+0.36
removal	6.7	-0.13	55.5	-0.07	39.1	+0.02	50.0	0.00
re-reason	62.4	-0.90	65.0	-11.56	60.9	-3.87	64.1	-7.21
en-ja	<i>candidates: 200 / 214 / 88 / 134</i>							
hedging	4.5	+0.02	43.9	+0.64	39.8	0.00	53.5*	0.00*
hindsight	73.5	+1.10	83.6	+2.54	93.2	+6.85	95.9	+6.97
oracle-1	28.0	+0.20	63.6	-0.67	62.5	+0.19	65.2	+0.02
oracle-k	27.0	+0.11	60.3	-0.11	61.4	+0.41	68.7*	+0.27
removal	6.5	-0.05	47.2	+0.26	42.0	-0.05	53.2	-0.06
re-reason	50.5	-0.03	51.9	-16.07	62.5	-5.40	49.2	-9.14
en-ur	<i>candidates: 399 / 210 / 216 / 146</i>							
hedging	0.5	-0.09	20.0	+0.30	38.9	+0.18	44.5	+0.21
hindsight	0.3	+10.66	35.2	+8.92	97.2	+16.03	97.0	+14.63
oracle-1	0.0	+0.48	32.4	+0.08	64.8	+1.26	61.0	+0.10
oracle-k	0.0	+0.66	29.5	+0.04	70.4	+1.70	62.3	+0.39
removal	0.3	-0.02	41.4	+1.77	38.4	+0.26	44.5	+0.07
re-reason	0.3	-1.24	47.1	-19.75	70.5	-5.28	73.9	-11.81
en-yue	<i>candidates: 135 / 130 / 101 / 134</i>							
hedging	0.0	0.00	9.2	+0.04	43.6	+0.16	59.0	+0.13
hindsight	0.0	+1.31	13.1	+1.94	91.1	+7.79	95.3	+8.42
oracle-1	2.2	+0.41	6.9	-0.39	58.4	+0.64	73.9	+0.32
oracle-k	0.7	+0.04	6.9	-0.41	62.4	+0.59	79.9	+0.34
removal	0.0	+0.02	5.4	+0.19	45.5	+0.16	57.5	+0.13
re-reason	0.0	-1.56	29.2	-9.38	50.2	-8.36	67.3	-11.43
en-zh	<i>candidates: 94 / 123 / 52 / 141</i>							
hedging	1.1	-0.01	59.5	-0.70	42.3	0.00	53.2	-0.01
hindsight	72.3	-0.12	86.2	+0.89	76.9	+8.07	90.9	+8.49
oracle-1	26.6	-0.12	70.7	-0.33	69.2	-0.35	75.9	-0.02
oracle-k	38.3	+0.04	64.2	-1.32	67.3	-0.22	82.3	0.00
removal	6.4	-0.01	59.3	+0.26	44.2	-0.04	52.5	-0.04
re-reason	50.0	-1.02	66.7	-9.58	67.3	-8.82	68.2	-8.31

Table 5: Intervention results grouped by language pair. *Rate* is the percentage of error candidates resolved, denominators are the candidate counts given in each block header, in the order (Qwen3-8B / Ministral-3-8B / Qwen3-14B). ΔC is $\Delta \text{COMET} \times 100$ against the unmodified-trace replay (so, +10.66 is an increase in 10.66 points on a scale of 0 to 100). Bold marks the highest resolution rate and the highest ΔC within each language block and model.