Reflective Translation: Enhancing Low-Resource Machine Translation through Self-Reflection

Lailah Denny, Agrim Sharma, Nicholas Cheng, Erin Tan

Algoverse AI Research

lailahdenny2@gmail.com, agrim400101@gmail.com, nicholascheng123456@gmail.com, erin@algoverseairesearch.org

Abstract

Low-resource languages such as isiZulu and isiXhosa face persistent challenges in machine translation (MT) due to limited parallel corpora and scarce linguistic resources. Recent work on large language models (LLMs) suggests that self-reflection—the ability of a model to critique and revise its own outputs—has been shown to enhance reasoning and factual consistency. Building on this idea, we present a framework for *Reflective Translation*, wherein an LLM internally evaluates and corrects its own translations to improve semantic fidelity by employing multi-round prompting. We apply our method using GPT-3.5 and Claude Haiku 3.5 on English–isiZulu and English–isiXhosa pairs from the OPUS-100 and NTREX-African datasets. To assess translation quality, we compute BLEU and COMET scores. We find that *Reflective Translation* yields consistent improvements in translation quality from the first to the second pass, across both isiZulu (+0.08 BLEU, +0.13 COMET) and isiXhosa (+0.07 BLEU, +0.09 COMET). We further introduce a first-of-its-kind reflection-augmented dataset built from model-generated self-critiques and corrected translations. Overall, this paper demonstrates reflection-based prompting as a promising approach for enhancing data quality and improving MT in under-resourced languages, bridging the gap between LLM reasoning research and practical translation for global linguistic inclusion.

Introduction

Machine Translation (MT) is a fundamental task for global communication, enabling users to exchange information across languages without the need for human intermediaries. The effectiveness of MT systems depends on their linguistic, factual, and logical faithfulness. Recently, large language models (LLMs) have emerged as powerful translation engines, demonstrating strong performance on task-specific benchmarks without additional fine-tuning (Brants et al. 2007; Moslem et al. 2023). Despite these advances, there remains a substantial gap in LLM performance for low-resource languages (Robinson et al. 2023; Haddow et al. 2022).

Low-resource languages suffer from limited high-quality labeled data, which constrains models in learning incomplete distributions that do not capture full linguistic and sociocultural variation (Pava et al.). Under these conditions, LLM-based translators are particularly prone to hallucinations, omissions, and culturally biased renderings (Wang and Sennrich 2020). To mitigate these challenges, researchers have developed multilingual pretraining strategies such as mBART and mRASP, which extend coverage and improve cross-lingual transfer (Pan et al. 2021). Although these methods improve robustness, they still struggle to ensure semantic faithfulness and contextual grounding in low-resource settings.

An emerging line of research explores *self-reflection*—prompting a model to critique and refine its own outputs—as a pathway to improve reasoning quality and factuality. Iterative prompting frameworks such as Reflexion (Shinn et al. 2023) enhance reasoning through verbal feedback loops, while Self-Refine (Madaan et al. 2023) achieves similar gains through multi-round self-revision. The Chain-of-Verification framework (Creswell and Shanahan 2023) further decomposes generation into explicit verification and synthesis stages, providing an interpretable reasoning scaffold. Beyond inference-time prompting, Reflection-Tuning (Li et al. 2023) and the recently proposed ReflectionLLMMT model (Wang et al. 2024) demonstrate that reflection signals can be incorporated into model training or translation pipelines to improve consistency and accuracy.

In this work, we investigate whether reflective checking can be leveraged as an explicit reasoning step to improve translation fidelity without multi-round prompting or model fine-tuning. We conceptualize translation as a form of constrained reasoning, where each target segment must faithfully represent the semantics of the source text. To operationalize this, we design a reflection-guided translation framework in which an LLM first produces an initial translation, then generates a structured self-reflection identifying characteristic failure modes (e.g., mistranslation, omission, distortion) and concise corrective guidance. Unlike prior reflection-based approaches that depend on retraining or direct synthetic data generation, our method embeds reflection directly into the prompting process, enabling test-time self-correction.

Empirical evaluation of reasoning-intensive translation benchmarks shows that the introduction of self-reflection significantly improves both the COMET and BLEU scores over standard prompting baselines, without any additional fine-tuning or training. Our results align with previous findings that reflective reasoning improves factual consistency and robustness in generation (Shinn et al. 2023; Li et al. 2023), while differing from concurrent approaches such as Wang et al. (2024), which integrate reflection into supervised translation fine-tuning rather than prompt-level control.

Concretely, our contributions are as follows:

- We propose a novel reflection-guided prompting framework for machine translation, where models generate and act upon structured self-assessments to improve translation faithfulness.
- We perform empirical evaluation on two multilingual and low-resource translation datasets (OPUS-100 and NTREX-African) across two different LLMs (GPT-3.5 and Claude Haiku 3.5).

Methods

We propose Reflective Translation, an approach in which a model is guided to self-review its own translations to produce improved outputs based on structured feedback. For each source sentence, GPT-3.5 (OpenAI 2023) and Claude Haiku 3.5 (Anthropic 2024) first generate an initial translation, which is then evaluated against predefined BLEU (Papineni et al. 2002) and COMET (Rei et al. 2020) thresholds. If the translation does not meet these criteria, the model produces a structured reflection that identifies key errors, suggests concise corrections, and highlights critical phrases or factual details to preserve.

Each reflection consists of three components: identifying key errors in the initial translation, suggesting reusable high-level corrections, and highlighting critical phrases or content that must be preserved. This reflection informs the generation of a second, refined translation, enhancing both semantic fidelity and fluency without requiring external feedback or additional parallel data.

To prevent leakage from the reflection into the second translation, key content words are extracted and masked using the Rapid Automatic Keyword Extraction (RAKE) algorithm, implemented in the NLTK library (Rose et al. 2010). Key phrases are replaced with the `<MASK>` token, ensuring that the model relies on comprehension rather than copying. Second-attempt translations are generated based on these masked reflections, guided by structured feedback. BLEU and COMET scores are computed for the improved translations to quantify the impact of reflection-informed guidance, forming the final reflection dataset used in subsequent experiments.

The framework is designed to be generalizable across low-resource languages and compatible with any large language model capable of following structured prompts. It can also incorporate supplementary guidance strategies, such as few-shot examples or threshold-based performance criteria, to further improve reliability and consistency. In this work, we apply the reflective translation method to English–isiZulu and English–isiXhosa corpora.

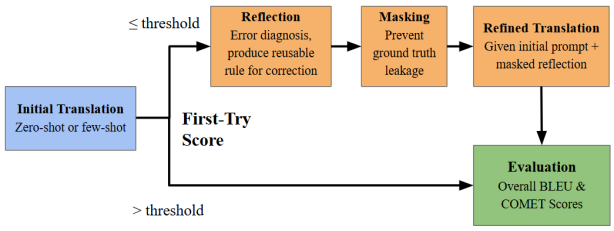

Figure 1: Overview of the reflective translation framework. The model generates an initial translation, evaluates it using structured reflection, and produces a refined translation guided by key error corrections and masked content words.

Datasets

We evaluate our approach using two datasets: *OPUS-100* (Tiedemann 2012), which provides broad multilingual coverage across diverse languages and domains, and *NTREX-African* (Nekoto et al. 2023), which contains sentence-level data spanning general topics in African languages. Our experiments focus on the low-resource languages isiZulu and isiXhosa, enabling assessment of model performance in challenging translation scenarios. Specifically, we use OPUS-100 for English–isiZulu translation and NTREX-African for English–isiXhosa translations.

OPUS-100 is a large-scale parallel corpus covering 100 language pairs, constructed primarily from web-crawled sources including government documents, subtitles, and domain-diverse online texts. The English–isiZulu subset contains approximately 250,000 parallel sentences. NTREX-African, in contrast, provides professionally curated evaluation sets for several African languages, including isiXhosa, with roughly 40,000 sentence pairs, collected from educational and media sources to ensure linguistic quality and coverage.

IsiZulu and isiXhosa are Bantu languages spoken predominantly in South Africa. IsiZulu is the most widely spoken language in the country, with roughly 12 million native speakers, while isiXhosa has around 8 million native speakers. Despite this, both languages are considered low-resource in NLP due to the limited availability of high-quality parallel corpora and computational resources. This scarcity makes it challenging to train and evaluate machine translation models, highlighting the need for methods like reflective translation that can improve performance without relying on massive datasets.

Both datasets differ in their language coverage and construction: OPUS-100 offers large parallel corpora across 100 language pairs, including English–isiZulu, derived from web-crawled and domain-diverse sources. NTREX-African, in contrast, provides a smaller but professionally curated evaluation set for several African languages, including isiXhosa. Using each dataset for the language it most robustly

supports ensures that our experiments rely on the highest-quality and most complete parallel data available for each translation direction.

Baseline

We establish baseline translation performance using GPT-3.5 and Claude 3.5 without fine-tuning. For each source sentence, the model produces an initial translation, which is then reviewed via reflection-informed prompting. This reflection identifies key errors and provides concise corrective guidance, which the model uses to generate a second, refined translation. In addition to zero-shot generation, we conduct few-shot experiments where exemplar translations are included in the prompt, allowing assessment of the impact of in-context examples on translation quality. This setup enables evaluation of both the contribution of reflection and the benefit of few-shot prompting.

Evaluation

Translation quality is assessed using two complementary metrics: BLEU and COMET. BLEU quantifies n-gram overlap between generated and reference translations, providing insight into surface-level lexical accuracy while penalizing overly short outputs. Formally, BLEU is computed as:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where p_n is the n-gram precision, w_n are the weights (usually uniform), N is the maximum n-gram order, and BP is the brevity penalty:

$$BP = \{\, 1 \; if c > re^{1-r/c} if c \leq r$$

with c as the candidate translation length and r as the reference length.

COMET is a neural-based metric defined as:

$$COMET(x, y) = f_\theta(x, y)$$

where x is the generated translation, y is the reference translation, and f_θ is a pretrained multilingual embedding-based model that predicts a human-aligned quality score.

Both first- and second-attempt translations are scored, and averaging across examples provides a robust estimate of baseline performance and the gains attributable to reflection. This dual-metric evaluation captures both literal accuracy and semantic preservation, which is critical for low-resource language settings.

OPUS-100 and NTREX-African differ in their language coverage and construction: OPUS-100 offers large parallel corpora across 100 language pairs, including English–isiZulu, derived from web-crawled and domain-diverse sources, whereas NTREX-African provides smaller, professionally curated evaluation sets for several African languages, including isiXhosa. Using each dataset for the language it most robustly supports ensures that our experiments rely on the highest-quality and most complete parallel data available for each translation direction.

Results

Across translation tasks, error thresholds, and prompting strategies, we find that self-reflection consistently improves translation quality. Tables 4 and 5 report the BLEU and COMET scores for isiXhosa and isiZulu translations, respectively. Results are reported across five different score thresholds—that is, the minimum acceptable score for first-try translations without executing our self-reflection pipeline. We observe consistent increases in second-try translations after applying Reflective Translation to incorrectly translated samples. Overall, the results show that iterative self-refinement offers a simple but effective way to boost translation quality without requiring additional training data or parameter updates.

Across all settings, reflective translation improves second-pass outputs by an average of +0.07 BLEU and +0.18 COMET, with the larger gains appearing under stricter confidence thresholds. These consistent improvements demonstrate that self-reflection reliably enhances translation quality. The plateau in second-pass BLEU at higher thresholds reflects how strict confidence requirements limit the number of sentences eligible for reflection, causing the metric to stabilize as fewer sentences undergo refinement due to the focus on certain translations.

Ablations

Error Threshold. To understand how confidence filtering interacts with reflection, we perform a threshold ablation over a range of cutoff values.

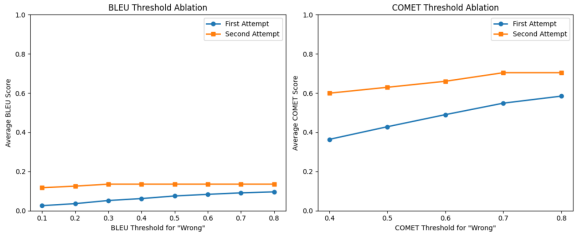

Figure 2: Threshold Ablation using Haiku 3.5. Higher thresholds produce fewer translations but lead to larger BLEU and COMET improvements, highlighting the effectiveness of self-reflection under high-confidence predictions.

Lower thresholds increase coverage, but yield smaller quality gains, indicating that reflection has limited impact when the initial translation is highly uncertain.

Taken together, these results show that self-reflection remains helpful across all thresholds, with its strongest benefits appearing under stricter filtering where refinement is most impactful.

Prompting Strategies In addition to the zero-shot strategy employed in our baseline results, we evaluate two additional prompting strategies to isolate the influence of reasoning structure, in-context signals, and self-correction mechanisms on translation quality. We employ Few-Shot (Brown et al. 2020) and Chain-of-Thought (CoT) (Wei et al. 2023) prompting. These prompting strategies are incorporated into

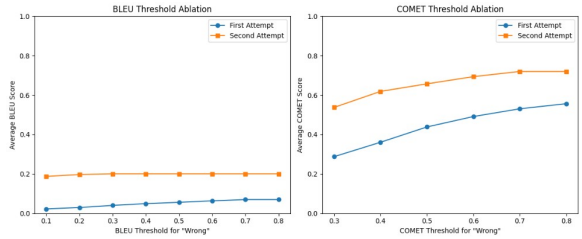

Figure 3: Threshold Ablation using ChatGPT 3.5. Lower thresholds increase coverage but produce smaller gains, highlighting the trade-off between output volume and quality.

our reflective-translation pipeline. All prompts are provided in the Appendix.

Chain-of-thought (CoT) prompting encourages the model to articulate intermediate reasoning before producing the final translation. This setup allows us to test whether making the model "think out loud" helps it navigate structural ambiguity, complex sentence constructions, or subtle semantic nuances. The few-shot prompt presents a small number of input–output examples. These demonstrations give concrete guidance on target style and sentence structure, helping the model align its translations through pattern matching and in-context learning.

Table 1: Average scores for prompting techniques using Haiku 3.5. $BLEU_1$ and $COMET_1$ correspond to the first attempt, while $BLEU_2$ and $COMET_2$ correspond to the second attempt.

English-isiXhosa				
Prompt	$BLEU_1$	$COMET_1$	$BLEU_2$	$COMET_2$
baseline	0.08414	0.57696	0.16297	0.65792
chain-of-thought	0.09109	0.56773	0.11680	0.64955
few-shot	0.08936	0.57109	0.16970	0.69555
English-isiZulu				
Prompt	$BLEU_1$	$COMET_1$	$BLEU_2$	$COMET_2$
baseline	0.16068	0.65606	0.24389	0.78464
chain-of-thought	0.15195	0.64749	0.23953	0.75492
few-shot	0.06751	0.57770	0.28726	0.79589

Discussion

Our study demonstrates a significant boost in translation quality for low-resource, morphologically rich languages via structured self-reflection. Across both *isiZulu* and *isiXhosa*, second-pass translations incorporating masked self-critiques consistently outperformed initial outputs, confirming that explicit reflection helps the model identify and correct systematic errors such as agreement mismatches and named-entity distortions. Among prompting strategies, the few-shot reflection setup achieved the most balanced and stable improvements across BLEU and COMET metrics, highlighting the value of in-context exemplars for guiding reflective behavior. Interestingly, self-reflection in the zero-shot setting still surpassed few-shot prompting *without* reflection, suggesting that structured critique provides a stronger inductive bias for faithfulness than mere exposure to exemplars. Additional experiments demonstrated a predictable scaling of performance with stricter adequacy and COMET thresholds, supporting the hypothesis that reflection can serve as a controllable lever for enhancing semantic precision.

Table 2: Average scores for prompting techniques using ChatGPT 3.5. $BLEU_1$ and $COMET_1$ correspond to the first attempt, while $BLEU_2$ and $COMET_2$ correspond to the second attempt.

English-isiXhosa				
Prompt	$BLEU_1$	$COMET_1$	$BLEU_2$	$COMET_2$
baseline	0.13351	0.63365	0.12326	0.66553
chain-of-thought	0.12307	0.62451	0.10580	0.68168
few-shot	0.11132	0.60456	0.11571	0.66560
English-isiZulu				
Prompt	$BLEU_1$	$COMET_1$	$BLEU_2$	$COMET_2$
baseline	0.16111	0.67172	0.20321	0.76672
chain-of-thought	0.17722	0.66225	0.15124	0.74844
few-shot	0.17217	0.62550	0.11886	0.71479

Comparison to Baselines

To further validate our approach, we compared Reflective Translation (RT) against two relevant baselines: Self-Refine (SR) (Madaan et al. 2023) and GEPA (Agrawal et al. 2025). Self-Refine iteratively refines translations through model self-critique, which aligns closely with the premise of Reflective Translation, while GEPA serves as an additional reference. The results, shown in Table **??**, demonstrate that RT consistently outperforms both SR and GEPA across BLEU and COMET metrics, with more robust and consistent translation quality across languages. These findings indicate that reflective prompting improves structural constancy and syntactic fidelity in low-resource settings beyond what iterative self-refinement alone can achieve.

Table 3: Comparison of Reflective Translation (RT) with Self-Refine (SR) and GEPA baselines.

BLEU Comparison			
Method	BLEU(SR)	BLEU(GEPA)	BLEU(RT)
Scores	0.00726	0.01651	0.13543
COMET Comparison			
Method	COMET(SR)	COMET(GEPA)	COMET(RT)
Scores	0.33204	0.40096	0.54861

Overall, this work presents a lightweight, model-agnostic framework for improving translation in low-resource settings through reflection-guided prompting. By embedding structured self-critique into the inference process rather than relying on fine-tuning or additional data collection, this approach offers a practical and reproducible method for enhancing translation fidelity while minimizing anno-

tation costs. Beyond performance gains, our framework also functions as a form of model-based data augmentation—producing interpretable (source, draft, critique, revision) tuples that can support future supervised training and analysis of reflective behavior.

Limitations and Future Work

While our findings highlight the potential for self-reflection to improve translation quality, there are several limitations in our preliminary work. The scope of our experiments is restricted to isiZulu and isiXhosa, which are both members of the Nguni language group, and thus share similarities in grammar structure and phonetics; extended experiments are required on a variety of low-resource languages from independent origins and may not show the ability of Reflective Translation to generalize to other low-resource or morphologically complex languages. Additionally, only two large language models (GPT-3.5 and Claude Haiku 3.5) were evaluated; further studies should include additional model architectures to assess generalizability.

Moreover, while BLEU and COMET scores effectively capture surface and semantic similarity, they may overlook sociocultural nuances, small grammatical distinctions, and contextual fidelity, particularly in underrepresented languages. Developing richer evaluation protocols that incorporate human judgment could provide deeper insights into translation quality. Future work may also explore broader typological coverage, more adaptable architectures, and expanded reflective prompting strategies to enhance scalability, fairness, and representation in low-resource machine translation.

References

Agrawal, L.; Tan, S.; Soylu, D.; Ziems, N.; Khare, R.; Opsahl-Ong, K.; Singhvi, A.; Shandilya, H.; Ryan, M.; Jiang, M.; and Potts, C. 2025. GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning.

Anthropic. 2024. Claude 3.5 Sonnet ModelReport. https://docs.anthropic.com/en/docs/legacy-model-guide\#claude-35sonnet.

Brants, T.; Popat, A.; Xu, P.; Och, F. J.; and Dean, J. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 858–867.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Creswell, A.; and Shanahan, M. 2023. Faithful Reasoning Using Chain-of-Verification. *arXiv preprint arXiv:2309.11495*.

Haddow, B.; Bawden, R.; Miceli Barone, A. V.; Helcl, J.; and Birch, A. 2022. Survey of Low-Resource Machine Translation. *Computational Linguistics*, 48(3): 673–732.

Li, M.; Chen, L.; Chen, J.; He, S.; Huang, H.; Gu, J.; and Zhou, T. 2023. Reflection-Tuning: Data Recycling Improves LLM Instruction-Tuning.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Gupta, S.; Majumder, B. P.; Hermann, K.; Welleck, S.; Yazdanbakhsh, A.; and Clark, P. 2023. Self-Refine: Iterative Refinement with Self-Feedback. *arXiv preprint*, arXiv:2303.17651.

Moslem, Y.; Haque, R.; Kelleher, J. D.; and Way, A. 2023. Adaptive Machine Translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 227–237. Tampere, Finland: European Association for Machine Translation.

Nekoto, W.; Dossou, B. F. P.; Orife, I.; Kreutzer, J.; Nabwire, L.; Sefara, T.; et al. 2023. The NTREX-128 Machine Translation Benchmark for African Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

OpenAI. 2023. GPT-3.5 Model Card. https://platform.openai.com/docs/models/gpt-3-5.

Pan, X.; Wang, M.; Wu, L.; and Li, L. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. ACL.

Pava, J. N.; Meinhardt, C.; Zaman, H. B. U.; Friedman, T.; Truong, S. T.; Zhang, D.; Marivate, V.; and Koyejo, S. ???? Mind the (Language) Gap.

Rei, R.; Stewart, C.; Farinha, A. C.; and Lavie, A. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702.

Robinson, N.; Ogayo, P.; Mortensen, D. R.; and Neubig, G. 2023. ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. In *Proceedings of the Eighth Conference on Machine Translation*, 392–418. Singapore: Association for Computational Linguistics.

Rose, S.; Engel, D.; Cramer, N.; and Cowley, W. 2010. *Automatic Keyword Extraction from Individual Documents*, 1 – 20. ISBN 9780470689646.

Shinn, N.; Cassano, F.; Berman, E.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 1–16.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, 2214–2218.

Wang, C.; and Sennrich, R. 2020. On exposure bias, hallucination and domain shift in neural machine translation. *arXiv preprint arXiv:2005.03642*.

Wang, Y.; Li, Z.; Hu, B.; Huang, S.; Zhang, M.; and Luo, W. 2024. ReflectionLLMMT: Reflection-Based Machine Translation with Large Language Models. *arXiv preprint arXiv:2406.08434*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Reference Examples and Prompt Templates for Low-Resource Languages

Baseline Translation Prompts

Baseline First-Try Prompt:

```
Source ({lang_name}): {source_text}
You are a professional translator. Translate the given text accurately into English.
Preserve the original meaning, tone, and nuance.
Output format (exact):
Translation:
<START_TRANSLATION>
<your English translation here>
<END_TRANSLATION>
Do NOT include any explanations.
```

Figure 4: Baseline translation prompt for first attempt.

Baseline Second-Try Prompt (Reflection-Based):

```
Source ({lang_name}): {source_text}
You are a professional translator. Based on the following review and reflection,
provide an improved translation.
Reflection: {reflection}
Output format (exact):
Translation:
<START_TRANSLATION>
<your improved English translation here>
<END_TRANSLATION>
Do NOT include explanations.
```

Figure 5: Baseline second attempt with reviewer reflection.

Few-Shot Translation Prompts

Few-Shot First-Try Prompt:

```
Source ({lang_name}): {source_text}
You are a professional translator. Translate the following text into English
accurately.
Here are examples for guidance:
Source (isiZulu): Ngiyabonga kakhulu.
Translation: Thank you very much.
Source (isiZulu): Unjani namhlanje?
Translation: How are you today?
Output format:
Translation:
<START_TRANSLATION>
<your English translation here>
<END_TRANSLATION>
```

Figure 6: Few-shot prompt with guiding examples.

Brief Reasoning (Chain-of-Thought) Prompts

Few-Shot Second-Try Prompt (Reflection-Based):

```
Source ({lang_name}): {source_text}
Improve the English translation of the following text using the review and
reflection: {reflection}
Here are examples for guidance:
Reflection: Incorrect verb nuance fixed.
Improved Translation: I would like to ask a question.
Output format:
Translation:
<START_TRANSLATION>
<your improved English translation here>
<END_TRANSLATION>
```

Figure 7: Few-shot second attempt with reflection and examples.

Brief Reasoning First-Try Prompt:

```
Translate the following {lang_name} text into English.
Before giving the final answer, perform brief internal reasoning. Do NOT reveal your
reasoning.
Source ({lang_name}): {source_text}
Output format:
Translation:
<START_TRANSLATION>
<your English translation here>
<END_TRANSLATION>
```

Figure 8: First attempt with brief internal reasoning.

Brief Reasoning Second-Try Prompt (Reflection-Based):

```
Improve the English translation of the following {lang_name} text.
Use the review and reflection to fix errors. Perform brief internal reasoning but do
NOT reveal it.
Source ({lang_name}): {source_text}
Review and Reflection: {reflection}
Output format:
Translation:
<START_TRANSLATION>
<your improved English translation here>
<END_TRANSLATION>
```

Figure 9: Second attempt with brief reasoning and reflection.

Tables for BLEU and COMET Scores

Table 4: isiXhosa (xh) BLEU and COMET Threshold Ablation (ChatGPT 3.5 / Haiku 3.5)

	BLEU		**COMET**	
Threshold	First-Try	***RT***	First-Try	***Reflective Translation***
0.40	0.050 / 0.062	**0.201** / 0.135	0.361 / 0.364	**0.619** / 0.599
0.50	0.056 / 0.075	**0.201** / 0.135	0.438 / 0.428	**0.657** / 0.630
0.60	0.064 / 0.084	**0.201** / 0.135	0.491 / 0.490	**0.694** / 0.660
0.70	0.070 / 0.091	**0.201** / 0.135	0.531 / 0.549	**0.720** / 0.704
0.80	0.070 / 0.096	**0.201** / 0.135	0.556 / 0.585	**0.720** / 0.704

Table 5: isiXhosa (xh) BLEU and COMET Threshold Ablation (ChatGPT 3.5 / Haiku 3.5)

	BLEU		**COMET**	
Threshold	First-Try	***RT***	First-Try	***Reflective Translation***
0.40	0.050 / 0.364	0.201 / **0.599**	0.361 / 0.364	**0.619** / 0.599
0.50	0.056 / 0.438	0.201 / **0.657**	0.438 / 0.428	**0.657** / 0.630
0.60	0.064 / 0.491	0.201 / **0.694**	0.491 / 0.490	**0.694** / 0.660
0.70	0.070 / 0.531	0.201 / **0.720**	0.531 / 0.549	**0.720** / 0.704
0.80	0.070 / 0.556	0.201 / **0.720**	0.556 / 0.585	**0.720** / 0.704