Retrieve, Refine, and Translate: LLM-Based Translation for Low-Resource Languages

Shibo Zhang[1,2,3,4], Mieradilijiang Maimaiti[1,2,3,4,*], Zhengyi Guo[1,2,3,4], Dezhi Wang[1,2,3,4], Wu Le[5], Zhuofei Xie[5], Jiawei Chen[5], Wushouer Silamu[1,2,3,4],

[1]School of Computer Science and Technology, Xinjiang University, Urumqi, China
[2]Xinjiang Laboratory of Multi-Language Information Technology, Xinjiang University, Urumqi, China
[3]Xinjiang Multilingual Information Technology Research Center, Urumqi, China
[4]International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi, China
[5]Integrated Laboratory for Space, Air, and Ground Systems
Email: {zhangshibo, guozhengyi, kyriezhizhi}@stu.xju.edu.cn, {miradeljan51, wushour}@xju.edu.cn, alaia@richnetworks.hk, zhx34@pitt.edu, syscjw@zohomail.cn

***Abstract*—Large language models (LLMs) have demonstrated promising performance in machine translation, and retrieval-augmented generation (RAG) can further improve translation quality by incorporating external examples and leveraging the in-context learning capabilities of LLMs. However, in low-resource scenarios, empirical studies indicate that generated translations still exhibit diverse and persistent errors, and effectively refining such errors remains a challenge. In this work, we present a two-stage refinement framework for low-resource machine translation. The first stage focuses on correcting major translation errors through implicit refinement that exploits retrieved parallel examples, together with auxiliary knowledge guidance. The second stage introduces quality-oriented feedback to identify and revise the remaining errors iteratively. By integrating implicit correction with explicit feedback, the proposed framework provides a structured refinement process to improve translation quality. Extensive experiments on three benchmarks (FLORES-200, NTREX-128, and TICO-19) covering 8 low-resource languages demonstrate the effectiveness of our framework, showing that our introduced approach achieves highly competitive performance against strong baseline systems with XCOMET and BLEURT. The code is publicly available.** [1]

***Index Terms*—Large Language Models, Low-Resource Languages, Machine Translation, Retrieval-Augmented Generation, Translation Refinement.**

I. Introduction

Large language models (LLMs) exhibit strong generalization in natural language processing tasks [1], including machine translation (MT). Unlike conventional neural MT that requires massive parallel data and task-specific training [2], LLMs can perform translation in zero-shot or few-shot settings by leveraging their strong in-context learning (ICL) capability [3]. Retrieval-Augmented Generation (RAG) enhances this capability by using retrieved external examples as task-specific prompts that guide the model toward more accurate outputs [4]. This paradigm allows LLMs to better leverage both examples and context during translation [5], making RAG especially appealing in low-resource settings where parallel data are limited. However, achieving high-quality translation in such scenarios remains a challenge.

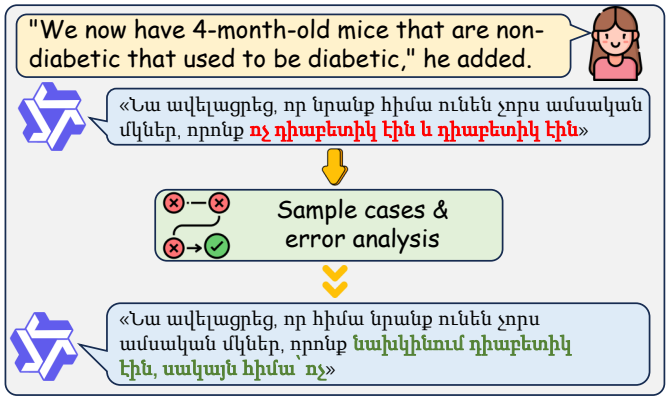

Fig. 1. An illustrative example of error correction in English to Armenian translation. The proposed framework successfully refines the semantic inconsistency found in the initial generation, demonstrating the effectiveness of our approach in low-resource scenarios.

Recent studies have explored various strategies to enhance LLM-based machine translation, such as incorporating multi-aspect linguistic knowledge to elicit translation [6], applying lexical constraints to enhance the precision of rare words through multilingual dictionary chains [7], and decomposing complex sentences to be shorter and simpler [8]. Although such methods can successfully guide LLMs toward better initial translations, the outputs generated in low-resource scenarios still contain diverse errors. Consequently, refinement-based strategies have emerged [9]–[11], which prompt LLMs to revise translations using auxiliary quality feedback and have shown that LLMs can refine imperfect outputs with additional guidance. Nevertheless, resolving a wide spectrum of translation errors simultaneously in a uniform refinement strategy remains a highly challenging task. Moreover, in low-resource settings, limited parallel data further worsen these issues, making it essential to strategically exploit existing signals,

*Corresponding author
[1]https://github.com/qaza200/2stagerefine

such as parallel examples retrieved during both generation and refinement [4].

Specifically, we observe that retrieved parallel sentence pairs can provide not only contextual examples but also valuable correction signals when their source sentences are translated by the same LLM and compared with reference translations. Such comparisons implicitly reveal how translation errors can be corrected, offering a natural basis for refining model outputs. Moreover, additional knowledge related to the source sentence and quality-related feedback on the generated translation can further complement example-based guidance by providing more explicit cues for translation decisions.

In this work, we propose a two-stage refinement framework for low-resource machine translation that builds upon this insight. The first stage performs implicit refinement by guiding the model to revise its initial translation through contextual comparison with the retrieved parallel sentence pairs. This process is further supported by auxiliary knowledge derived from the source sentence, such as topical information and salient keywords, which help constrain and guide the refinement. The second stage introduces explicit refinement using quality-oriented guidance to iteratively identify and correct residual errors that remain after the initial contextual adjustment. As depicted in Figure 1, combining these two stages, the proposed framework enables a progressive improvement of translation quality in low-resource scenarios.

Our contributions are summarized as follows:

(1) We propose a two-stage refinement framework for low-resource machine translation that integrates retrieval-augmented contextual guidance.

(2) We design an implicit refinement strategy that leverages retrieved parallel examples and auxiliary knowledge to correct initial translations, followed by an explicit refinement stage to further improve translation quality.

(3) We conduct extensive experiments on multiple benchmark datasets covering eight low-resource languages, demonstrating consistent improvements over baselines and validating the effectiveness of the proposed framework.

II. Related Work

A. In-Context Learning for LLM-Based MT

The emergence of LLMs has fundamentally shifted the paradigm of machine translation from supervised training to In-Context Learning (ICL). Seminal work [3] first introduced LLMs as few-shot learners, tackling tasks such as translation through in-context demonstrations. Following this, extensive empirical studies have validated the potential of LLMs in multilingual scenarios. For example, investigations of models such as BLOOM [12] and PaLM [1] have shown that the size of the scaling model significantly benefits translation performance. Subsequent research [13]–[15] provided comprehensive evaluations, highlighting that while LLMs can compete with commercial systems in high-resource languages through few-shot prompting, they still face challenges in low-resource settings compared to specialized NMT models such as NLLB [16].

To fully exploit the ICL capability of LLMs and overcome the limitations of static examples, Retrieval-Augmented Generation (RAG) [17] was introduced to dynamically select semantically relevant demonstrations from external datastores. By retrieving contextually similar parallel pairs [4] or optimizing retrieval strategies [18] to fetch more relevant content [19], such as error demonstrations [20], these methods significantly enhance the model's ability to handle domain-specific terminology and low-resource adaptation compared to random example selection. Furthermore, the introduction of Chain-of-Thought prompting [21] has inspired methods such as decomposed prompting [22] and step-by-step translation [23], which break down complex source sentences to improve translation accuracy for long-form texts. Additionally, dictionary-based prompting [24] injects lexical constraints into the context to address specific terminological inaccuracies. Furthermore, scaling the number of examples has been explored, with recent findings demonstrating the effectiveness of many-shot ICL [25] in further improving performance.

B. Refinement and Feedback Mechanisms

Although standard ICL focuses on generating a translation in a single pass, a growing body of literature suggests that LLMs perform better when allowed to refine their own outputs. This line of research is particularly relevant to address the hallucination and semantic errors often seen in low-resource translation [26]. Methods such as Pinpoint [10] and Ladder [11] introduce frameworks where the model receives fine-grained feedback or self-evaluates its initial output to produce a polished version. Similarly, xTower [27] was developed to explicitly explain and correct translation errors, shifting the focus from mere generation to critical evaluation.

In addition, adaptive approaches have been proposed to make these refinements more context-aware. Adaptive machine translation has been explored [28], where the model adjusts its output based on real-time feedback, while dynamic focus anchoring [29] attempts to optimize how the model attends to the relevant context during the generation process. These refinement-based strategies demonstrate that the translation quality of LLMs can be significantly improved by treating translation as an iterative problem-solving process rather than a one-time generation task.

III. Method

As illustrated in Figure 2, this section presents the proposed two-stage refinement framework for low-resource machine translation. The framework starts by retrieving relevant parallel sentence pairs and producing initial translations using a large language model. Section III-A introduces an implicit refinement stage, where retrieved examples, model-generated translations, reference translations, and auxiliary knowledge are jointly used to guide the correction. Section III-B then describes an explicit refinement stage that employs MQM-based quality analysis to iteratively review the translation and correct the remaining errors.

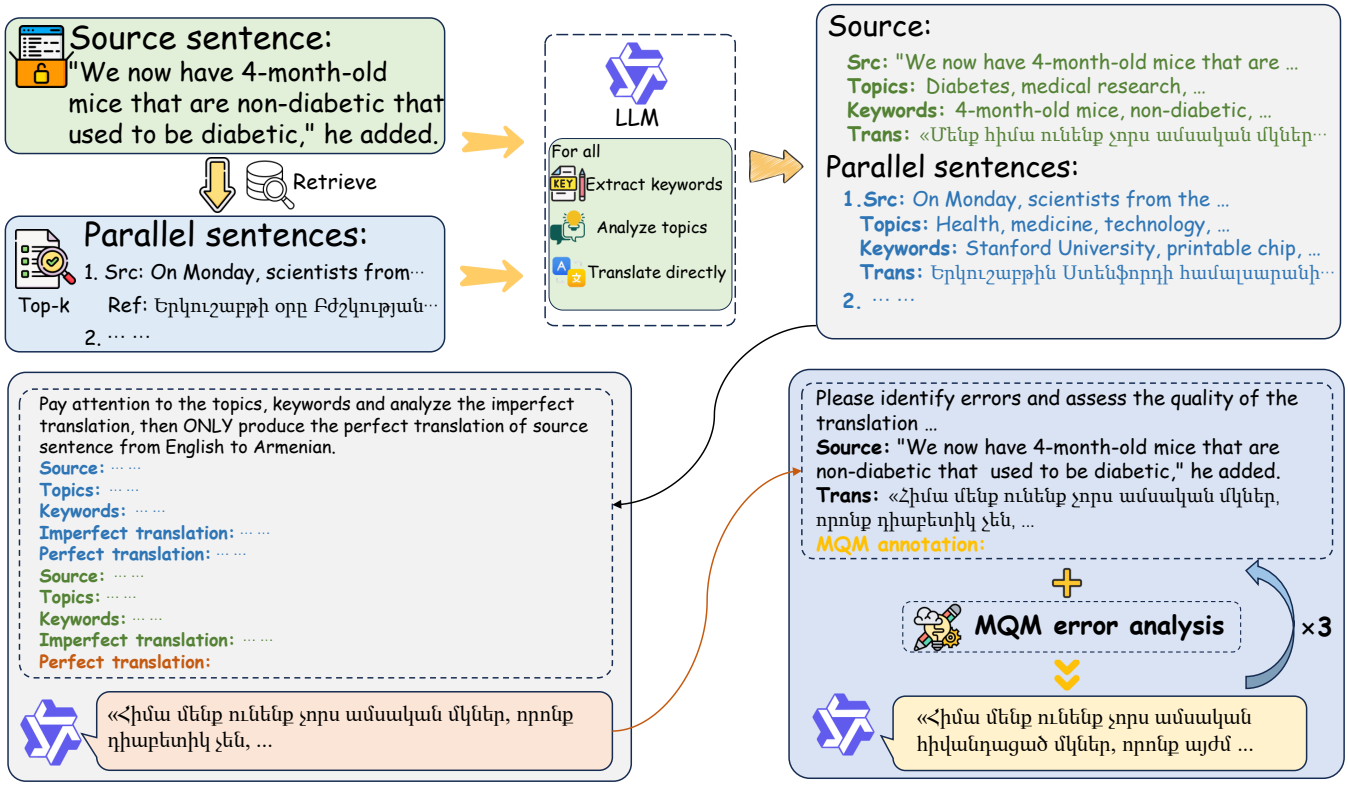

Fig. 2. Overview of the proposed framework. It integrates Implicit Refinement using retrieved knowledge and semantic constraints with Explicit Refinement via iterative ($T = 3$) MQM-based self-correction.

A. Implicit Refinement with Retrieved Parallel Examples

To effectively improve translation quality, the proposed framework performs implicit refinement by exploiting retrieved parallel examples to guide the correction of an initial translation. Formally, given a source sentence x, we first retrieve a small set of k relevant parallel pairs $\mathcal{R} = \{(x_i, y_i)\}_{i=1}^{k}$ from an external bilingual corpus, where x_i and y_i denote the source and reference retrieved, respectively. The LLM then generates an imperfect zero-shot translation $\hat{y}$ for the input x, and similarly produces a translation $\hat{y}_i$ for each retrieved source x_i using the same LLM. Simultaneously, auxiliary knowledge signals $\mathcal{K}(x), \mathcal{K}(x_i)$ are extracted from both input x and the retrieved source x_i, including topical information and salient keywords. This ensures a consistent information structure across the retrieval examples and the input query.

Subsequently, we construct a structured refinement context C_{imp} as shown in Figure 2 to facilitate implicit correction. For each retrieved example, the model-generated translation and its corresponding reference are integrated with the auxiliary knowledge to formulate contrastive demonstrations:

$$C_{imp} = \{(x_i, \mathcal{K}(x_i), \hat{y}_i, y_i)\}_{i=1}^{k} \tag{1}$$

By presenting multiple instances of model output $\hat{y}_i$ alongside high-quality references y_i under similar contextual conditions, the LLM captures error patterns and correction strategies within C_{imp} to refine its own imperfect translation $\hat{y}$ into $y_{imp} = \{C_{imp}, \mathcal{K}(x), x, \hat{y}\}$ through in-context learning without requiring explicit supervision or task-specific training.

B. Explicit Refinement with MQM-Guided Iteration

Although the implicit refinement stage effectively corrects the main translation errors, the resulting translation y_{imp} may still contain subtle issues. The second stage introduces an explicit, quality-oriented refinement step that encourages the model to examine its own translation and make targeted corrections.

Following prior work showing that LLM can produce human-like MQM (Multidimensional Quality Metrics) annotations, we prompt the LLM to identify potential error types and spans, as well as to provide brief explanations according to established MQM guidelines. The prompt format builds upon existing MQM-based frameworks [9] and incorporates example annotations from earlier studies [30]. Formally, given the source sentence x and the initial refined translation y_{imp}, the model produces a MQM analysis f:

$$f = \text{LLM}_{\text{MQM}}(x, y_{imp}) \tag{2}$$

The resulting MQM-style analysis f offers explicit fine-grained feedback on translation errors, and these identified issues are subsequently used as revision cues to guide the model in improving translation.

Although this explicit refinement stage yields noticeable improvements by correcting many prominent errors, we argue that a single round is often insufficient to resolve all residual

issues, as professional translation workflows typically involve multiple rounds of checking and revising rather than a single pass. Therefore, we apply the MQM-guided refinement process iteratively. Formally, we initialize the sequence with $y_0 = y_{imp}$. At each iteration step $t \in \{1, \ldots, T\}$, the LLM utilizes feedback f_t to update the translation from the previous step y_{t-1}:

$$y_t = \text{LLM}_{\text{refine}}(x, y_{t-1}, f_t) \tag{3}$$

By feeding the updated translation back into the explicit refinement cycle for a total of three iterations, we yield the final translation.

IV. Experiments

A. Setup

Datasets We use the following dataset, and the detailed statistics of the test sets are summarized in Table I.

- FLORES-200 [31] provides professionally translated sentences in 204 languages. We use the dev split as the retrieval pool and evaluate on the devtest split. The experiments are conducted in eight low-resource languages in the English $\rightarrow$ X direction.
- NTREX-128 [32] consists of 1997 parallel news-domain sentences. Following prior work [8], we use the first 1,000 pairs for evaluation and the remaining 997 pairs as a retrieval pool. Translation is performed from English into the same set of low-resource languages.
- TICO-19 [33] contains COVID-19 related texts in 35 languages. We use the dev set (971 pairs) as the retrieval pool and test in four languages in the English and Chinese $\rightarrow$ X direction.

TABLE I
STATISTICS OF THE EVALUATION DATASETS.

Dataset	Domain	Test Size (sentences)
FLORES-200	Multi-domain	1,012
NTREX-128	News	1,000
TICO-19	Medical	2,100

Baselines We compare our method with the following LLM-based translation baselines:

- COD [7] integrates multilingual dictionary lookups into the prompt to construct a semantic chain that bridges lexical gaps for rare words in low-resource languages.
- MAPS [6] prompts LLMs with multiple types of knowledge and dynamically selects the most effective to guide the translation process.
- TEaR [9] employs a self-refinement framework where the LLM rectifies the initial translations.
- CompTra [8] decomposes complex source sentences into simpler ones to retrieve relevant examples, then reconstructs the final translation from the sub-translations.

Implementation Details We use Qwen3-30B-A3B-Instruct [34] as the base language model for all experiments, including baselines. During inference, we employ greedy decoding by setting the temperature to 0.0 to ensure the reproducibility of our results. Detailed hyperparameters and implementation settings are summarized in Table II. Following prior findings that hybrid retrieval has better performance [35] [18], we adopt a mixed retrieval strategy that retrieves three parallel sentences for each input, comprising two using dense retrieval with Qwen3-Embedding-4B [36] and one using sparse retrieval with BM25s [37]. Translation quality is evaluated using two reference-based metrics, XCOMET-XL [38] and BLEURT-20 [39] [40], which measure semantic adequacy and overall fluency from complementary perspectives.

TABLE II
HYPER-PARAMETERS AND IMPLEMENTATION DETAILS.

Category	Value
Base LLM	Qwen3-30B-A3B-Instruct
Embedding Model	Qwen3-Embedding-4B
Data Type	bfloat16
Temperature	0.0
Top-p	1.0
Max Tokens	512

B. Main results

Overview Performance Table III – VI presents the comparative results on three diverse benchmarks: FLORES-200, NTREX-128, and TICO-19. By evaluating across multi-domain, news, and specialized medical contexts, we provide a comprehensive assessment of our framework's adaptability. In general, our method consistently achieves the strongest or highly competitive performance among a range of LLM-based baselines. In particular, it delivers superior XCOMET scores, demonstrating robust semantic fidelity and generalization capabilities regardless of domain complexity or source language variations from English and Chinese to low-resource targets.

English to Low-resource Results As shown in Tables III to V, our method consistently exceeds the 0-shot baseline and competitive approaches, ranging from dictionary-prompting (COD) and self-refinement (TEaR) to retrieval-augmented methods (CompTra). Across multi-domain (FLORES-200), news-domain (NTREX-128), and specialized medical-domain (TICO-19) scenarios, we achieve the best performance in the vast majority of metrics. Specifically, we observe substantial XCOMET gains of 8.0 points for Armenian on FLORES-200 and 9.1 points for Hebrew on NTREX-128 compared to the 0-shot baseline, while in the medical domain, our framework maintains superior adaptability, yielding the highest XCOMET scores across all four target languages despite the high density of specialized terminology. Although MAPS remains competitive, occasionally leading in specific metrics, our refinement strategy dominates in XCOMET, ensuring superior semantic fidelity and robustness across various professional domains and complex technical content.

Domain and Language Generalization Table VI assesses generalization in the medical TICO-19 dataset, characterized

TABLE III
RESULTS ON THE FLORES-200 DATASET (ENGLISH → XX) EVALUATED BY XCOMET AND BLEURT

Methods	Armenian		Azerbaijani		Hebrew		Lao	
	XCOMET	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**
0-shot	68.57	72.56	64.92	62.45	70.72	67.34	49.49	62.06
Vanilla RAG	71.47	74.90	68.63	64.31	71.50	68.06	55.55	67.35
COD	70.83	73.57	66.64	63.04	70.44	66.99	52.22	63.80
MAPS	75.53	76.53	71.31	65.20	76.33	71.27	57.05	**68.00**
TEaR	71.19	73.66	67.66	63.29	73.42	68.94	51.36	63.61
CompTra	61.11	62.71	67.32	63.32	70.74	67.53	49.99	52.54
Ours	**76.56**	**76.87**	**72.15**	**65.58**	**78.57**	**72.07**	**57.65**	67.64
Methods	Khmer		Tamil		Urdu		Bengali	
	XCOMET	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**
0-shot	50.98	57.35	53.01	74.83	66.84	56.38	67.33	73.98
vanilla rag	55.28	60.51	55.10	76.77	68.15	56.66	68.48	74.87
COD	51.24	57.49	53.25	74.60	63.65	55.81	66.22	74.01
MAPS	57.11	61.87	**57.87**	77.51	71.04	**57.56**	71.31	75.81
TEaR	52.93	58.88	55.23	76.42	67.86	56.65	68.44	74.32
CompTra	44.95	44.42	42.03	59.77	64.45	56.23	54.68	63.02
Ours	**57.74**	**61.97**	57.38	**77.75**	**71.52**	56.95	**71.45**	**75.94**

TABLE IV
RESULTS ON THE NTREX-128 DATASET (ENGLISH → XX) EVALUATED BY XCOMET AND BLEURT

Methods	Armenian		Azerbaijani		Hebrew		Lao	
	XCOMET	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**
0-shot	59.85	67.48	58.09	60.55	62.30	59.99	46.02	57.33
Vanilla RAG	64.25	70.19	61.49	62.06	65.71	62.17	52.86	**65.23**
COD	64.14	68.43	60.67	61.60	64.47	61.36	48.56	57.96
MAPS	67.88	**72.08**	64.41	64.53	69.03	64.45	52.92	63.77
TEaR	63.32	69.27	60.86	61.36	66.00	62.28	48.35	59.88
CompTra	55.37	58.29	59.86	61.29	65.26	62.01	48.95	52.27
Ours	**68.32**	71.80	**65.44**	**64.68**	**71.45**	**65.83**	**54.01**	64.44
Methods	Khmer		Tamil		Urdu		Bengali	
	XCOMET	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**
0-shot	49.51	52.69	49.78	70.25	62.37	55.22	63.01	69.48
Vanilla RAG	55.40	**58.19**	51.94	72.32	64.22	55.88	66.08	71.43
COD	50.39	52.01	48.26	67.96	58.44	54.01	62.63	69.46
MAPS	55.59	57.39	53.43	72.87	67.03	56.80	68.19	72.30
TEaR	51.50	53.59	51.25	71.53	63.93	55.27	65.72	70.63
CompTra	48.47	43.73	43.05	55.45	59.87	54.61	52.95	58.87
Ours	**56.99**	58.14	**53.74**	**73.76**	**67.75**	**56.95**	**68.36**	**72.39**

by a challenging shift to the Chinese source language. Despite these domain and linguistic hurdles, our method demonstrates superior adaptability, achieving the highest XCOMET scores in all four target languages. In particular, Vanilla RAG emerges as the most competitive baseline in this setting, surpassing other methods like MAPS. This performance spike is likely due to the high density of specialized terminology in medical texts, which retrieval mechanisms effectively supply. However, our method consistently exceeds it across the majority of metrics. This suggests that while retrieval provides the necessary domain terms, our refinement ensures their correct syntactic and semantic integration. Collectively, these results highlight the overall robustness of our framework, validating its ability to generate high-quality, semantically coherent translations even under the dual challenges of linguistic changes and specialized technical content.

C. Analysis

Impact of Knowledge Guidance To validate the contribution of knowledge guidance within the implicit refinement stage, as presented in Table VIII, we conducted an ablation study of FLORES-200 using the output of the first stage. Removing knowledge guidance leads to consistent degradation in both metrics. This indicates that while the retrieved paral-

TABLE V
RESULTS ON THE TICO DATASET (**ENGLISH → XX**) EVALUATED BY XCOMET AND BLEURT

Methods	Bengali		Khmer		Tamil		Urdu	
	XCOMET	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**
0-shot	62.22	72.37	56.70	59.60	48.54	70.87	63.44	52.09
Vanilla RAG	68.20	76.00	65.77	**67.75**	53.71	77.21	67.12	**54.88**
COD	61.43	72.38	57.65	60.59	49.24	71.87	59.70	52.23
MAPS	68.25	75.89	64.06	66.14	53.99	76.23	67.55	52.91
TEaR	65.75	74.66	60.66	63.62	51.61	74.91	65.63	52.53
CompTra	52.02	61.48	52.89	50.71	42.18	59.21	62.25	54.78
Ours	**69.30**	**76.43**	**65.97**	67.43	**55.09**	**77.54**	**68.70**	53.95

TABLE VI
RESULTS ON THE TICO-19 DATASET (**CHINESE → XX**) EVALUATED BY XCOMET AND BLEURT

Methods	Bengali		Khmer		Tamil		Urdu	
	XCOMET	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**
0-shot	58.18	69.30	53.42	57.84	47.86	70.15	59.70	46.87
Vanilla RAG	62.96	73.04	60.59	65.59	51.71	76.53	63.10	**48.64**
COD	57.83	69.32	54.96	58.89	48.43	70.73	58.20	46.82
MAPS	62.85	73.01	59.41	64.60	52.11	75.56	63.23	47.24
TEaR	60.54	72.00	55.23	61.37	50.01	74.34	61.49	47.44
CompTra	49.52	57.76	47.12	45.00	40.49	55.79	56.25	47.32
Ours	**63.82**	**73.80**	**60.61**	**65.73**	**52.50**	**77.02**	**64.44**	48.33

TABLE VII
IMPACT OF EXPLICIT REFINEMENT ON FLORES-200 (ENGLISH → XX) EVALUATED BY XCOMET AND BLEURT

Methods	Armenian		Azerbaijani		Hebrew		Lao	
	XCOMET	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**
Implicit Ref.	74.41	75.87	69.80	64.94	74.34	69.44	55.19	66.84
+ Explicit Ref.	**76.56**	**76.87**	**72.15**	**65.58**	**78.57**	**72.07**	**57.65**	**67.64**
Methods	Khmer		Tamil		Urdu		Bengali	
	XCOMET	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**
Implicit Ref.	55.97	61.31	56.53	77.31	69.57	**56.96**	69.86	75.14
+ Explicit Ref.	**57.74**	**61.97**	**57.38**	**77.75**	**71.52**	56.95	**71.45**	**75.94**

lel examples provide a solid reference, supplementing them with explicit semantic constraints further enhances the fidelity of translation. This additional context effectively guides the model toward more accurate and domain-consistent translations.

TABLE VIII
IMPACT OF KNOWLEDGE INTEGRATION. AVERAGE XCOMET AND BLEURT SCORES ACROSS ALL 8 LANGUAGES ON FLORES-200.

Settings	**Average (8 langs)**	
	XCOMET	**BLEURT**
w/o	65.36	68.25
w/	**65.71**	**68.48**

Impact of Explicit Refinement As given in Table VII, to verify the necessity of the second stage, we evaluate the performance gain achieved by introducing the explicit refinement mechanism. We compare the results of the initial implicit refinement (the first stage only, denoted as *Implicit Ref.*) against the full framework equipped with MQM-guided explicit feedback (the second stage, denoted as *+Explicit Ref.*). As observed, while the first stage provides a strong foundation, the addition of explicit refinement yields consistent improvements across all eight languages. For example, in Hebrew, XCOMET scores increase from 74.34 to 78.57, and significant gains are also evident in Armenian and Azerbaijani. This shows that while contextual demonstrations in the first stage effectively resolve major translation errors, the quality-oriented feedback in the second stage is critical to correct subtle residual errors and further polish semantic adequacy and fluency.

Effect of Refinement Iterations As depicted in Figure 3,

TABLE IX
A CASE STUDY ON FLORES-200 (ENGLISH → HEBREW). THE RED TEXT HIGHLIGHTS TRANSLATION ERRORS, WHILE THE BLUE TEXT INDICATES THE CORRECTIONS MADE BY OUR FRAMEWORK.

	Sentence
Source	**USA Gymnastics and the USOC have the same goal — making the sport of gymnastics, and others, as safe as possible for athletes to follow their dreams in a safe, positive and empowered environment.**
Vanilla RAG	להפוך את ספורט הגימנסטיקה, וכן ספורט אחרים, ליותר – יש להם את אותו מטרה américain גימנסטיקה של ארצות הברית והוועד האולימפי כדי שיתאמצו למשהו שמחפשים בסביבה בטוחה, חיובית ומעודדת, letesאת בטוח ככל האפשר עבור. (Error: Hallucinated French word "américain" & Mixed-script"letes")
Implicit Refinement	להפוך את ספורט הגימנסטיקה, וכן ספורט אחרים, ליותר בטוח – מחפשים את אותו מטרה האמריקאי גימנסטיקה ארה"ב והוועד האולימפי כדי שיתאמצו לרצונותיהם בסביבה בטוחה, חיובית ומעודדת, letesאת ככל האפשר עבור. (Correction: Fixed hallucination "américain", but mixed-script "letes" persists)
Explicit Refinement	להפוך את ספורט הגימנסטיקה, וכן ספורט אחרים, ליותר בטוח – גימנסטיקה ארה"ב והוועד האולימפי האמריקאי מחפשים את אותה מטרה בסביבה בטוחה, חיובית ומאפשרת חלומותיהם אחרי כדי שיתאימו, אתלטים ככל האפשר עבור. (Corrected terminology "athletes" & improved idiomatic phrasing)
Reference	מועצת ההתעמלות של ארהב והוועד האולימפי שותפים לאותה מטרה - להפוך את ענף ההתעמלות, וענפים אחרים, לכמה שיותר בטוחים לספורטאים לשאוף להגשים את חלומותיהם בסביבה בטוחה, חיובית ומועצמת.

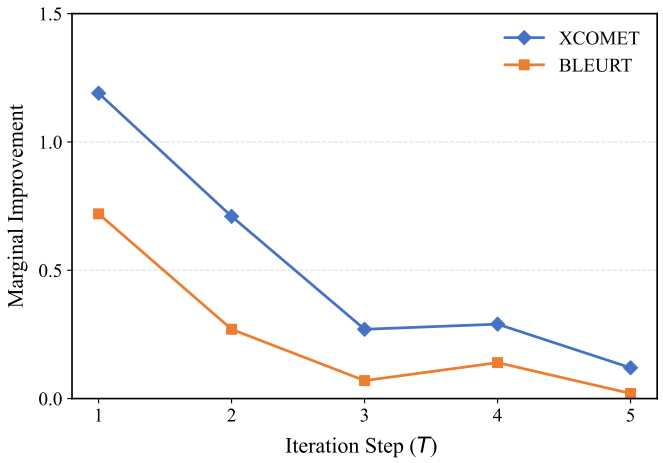

Fig. 3. Effect of the refinements with different iteration steps. We report the average marginal improvement of XCOMET and BLEURT across 8 languages on FLORES-200. The improvement converges after the third iteration, suggesting that $T = 3$ is an optimal trade-off between quality and efficiency.

we further investigated the optimal number of feedback iterations in the second stage by analyzing the average marginal performance gains at each iteration step in all eight languages on the FLORES-200 dataset. The initial iteration ($T = 1$) delivers substantial quality improvements. Crucially, extending the process yields additional consistent gains, confirming the effectiveness of iterative feedback in further polishing the translation. However, beyond the third iteration, the marginal benefits diminish significantly as performance stabilizes. Therefore, to balance translation quality with inference cost, we identify $T = 3$ as the optimal trade-off point.

Case Study Table IX presents an English to Hebrew example. Vanilla RAG exhibits severe errors, including a French hallucination ("*américain*") and a mixed-script term ("*ha*-letes"). Although implicit refinement corrects the named entity, the morphological error persists. The second stage finally resolves this by generating the correct Hebrew "*athletim*" and optimizing the idiom "follow their dreams." This proves explicit feedback is essential for fixing errors, yet the stylistic disparity with the reference underscores the difficulty of fully replicating human nuance in low-resource scenarios.

V. CONCLUSION

In this paper, we propose a two-stage refinement framework designed to address the challenges of low-resource machine translation. By synergizing retrieval-augmented generation with self-correction, our approach effectively integrates external contextual guidance with the model's intrinsic reasoning capabilities. Specifically, the first stage uses retrieved examples for implicit refinement, while the second stage uses iterative feedback to correct residual errors. Extensive experiments on FLORES-200, NTREX-128, and TICO-19 demonstrate that our method consistently outperforms competitive baselines, highlighting its robustness even in specialized medical domains. However, given the computational overhead of iterative refinement, future work will prioritize optimizing efficiency, exploring adaptive strategies to balance high translation quality with practical inference cost.

ACKNOWLEDGMENT

We sincerely thank our collaborators for their valuable support and anonymous reviewers for their constructive comments and suggestions on this work. This work was supported by the National Natural Science Foundation of China (Grant

No.62406316 and 201404041254), the Xinjiang 'Tianchi Talent' Recruitment and Introduction Program (awarded to Mieradilijiang Maimaiti).

REFERENCES

[1] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[2] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*, T. Luong, A. Birch, G. Neubig, and A. Finch, Eds. Vancouver: Association for Computational Linguistics, Aug. 2017, pp. 28–39. [Online]. Available: https://aclanthology.org/W17-3204/

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[4] U. Khandelwal, A. Fan, D. Jurafsky, L. Zettlemoyer, and M. Lewis, "Nearest neighbor machine translation," *arXiv preprint arXiv:2010.00710*, 2020.

[5] C.-C. Chang, C.-F. Li, C.-H. Lee, and H.-S. Lee, "Enhancing low-resource minority language translation with llms and retrieval-augmented generation for cultural nuances," in *Intelligent Systems Conference*. Springer, 2025, pp. 190–204.

[6] Z. He, T. Liang, W. Jiao, Z. Zhang, Y. Yang, R. Wang, Z. Tu, S. Shi, and X. Wang, "Exploring human-like translation strategy with large language models," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 229–246, 2024.

[7] H. Lu, H. Yang, H. Huang, D. Zhang, W. Lam, and F. Wei, "Chain-of-dictionary prompting elicits translation in large language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 958–976.

[8] A. Zebaze, B. Sagot, and R. Bawden, "Compositional translation: A novel llm-based approach for low-resource machine translation," *arXiv preprint arXiv:2503.04554*, 2025.

[9] Z. Feng, Y. Zhang, H. Li, B. Wu, J. Liao, W. Liu, J. Lang, Y. Feng, J. Wu, and Z. Liu, "Tear: Improving llm-based machine translation with systematic self-refinement," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 3922–3938.

[10] W. Xu, D. Deutsch, M. Finkelstein, J. Juraska, B. Zhang, Z. Liu, W. Y. Wang, L. Li, and M. Freitag, "Pinpoint, not criticize: Refining large language models via fine-grained actionable feedback," *CoRR*, 2023.

[11] Z. Feng, R. Chen, Y. Zhang, Z. Meng, and Z. Liu, "Ladder: A model-agnostic framework boosting llm-based machine translation to the next level," *arXiv preprint arXiv:2406.15741*, 2024.

[12] R. Bawden and F. Yvon, "Investigating the translation performance of a large multilingual language model: the case of bloom," *arXiv preprint arXiv:2303.01911*, 2023.

[13] D. Vilar, M. Freitag, C. Cherry, J. Luo, V. Ratnakar, and G. Foster, "Prompting palm for translation: Assessing strategies and performance," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 15 406–15 427.

[14] B. Zhang, B. Haddow, and A. Birch, "Prompting large language model for machine translation: A case study," in *International Conference on Machine Learning*. PMLR, 2023, pp. 41 092–41 110.

[15] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li, "Multilingual machine translation with large language models: Empirical results and analysis," in *Findings of the association for computational linguistics: NAACL 2024*, 2024, pp. 2765–2781.

[16] M. R. Costa-Jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard *et al.*, "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.

[17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[18] L. Tang, J. Qin, W. Ye, H. Tan, and Z. Yang, "Adaptive few-shot prompting for machine translation with pre-trained language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 24, 2025, pp. 25 255–25 263.

[19] J. Wang, F. Meng, Y. Zhang, and J. Zhou, "Retrieval-augmented machine translation with unstructured knowledge," *arXiv preprint arXiv:2412.04342*, 2024.

[20] Y. C. Liao, C.-J. Yu, C.-Y. Lin, H.-F. Yun, Y.-H. Wang, H.-M. Li, and Y.-C. Fan, "Learning-from-mistakes prompting for indigenous language translation," in *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, 2024, pp. 146–158.

[21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[22] T. Khot, H. Trivedi, M. Finlayson, Y. Fu, K. Richardson, P. Clark, and A. Sabharwal, "Decomposed prompting: A modular approach for solving complex tasks," *arXiv preprint arXiv:2210.02406*, 2022.

[23] E. Briakou, J. Luo, C. Cherry, and M. Freitag, "Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts," *arXiv preprint arXiv:2409.06790*, 2024.

[24] M. Ghazvininejad, H. Gonen, and L. Zettlemoyer, "Dictionary-based phrase-level prompting of large language models for machine translation," *arXiv preprint arXiv:2302.07856*, 2023.

[25] R. Agarwal, A. Singh, L. Zhang, B. Bohnet, L. Rosias, S. Chan, B. Zhang, A. Anand, Z. Abbas, A. Nova *et al.*, "Many-shot in-context learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 76 930–76 966, 2024.

[26] M. Enis and M. Hopkins, "From llm to nmt: Advancing low-resource machine translation with claude," *arXiv preprint arXiv:2404.13813*, 2024.

[27] M. V. Treviso, N. M. Guerreiro, S. Agrawal, R. Rei, J. Pombal, T. Vaz, H. Wu, B. Silva, D. Van Stigt, and A. F. Martins, "xtower: A multilingual llm for explaining and correcting translation errors," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 15 222–15 239.

[28] Y. Moslem, R. Haque, J. Kelleher, and A. Way, "Adaptive machine translation with large language models," in *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 2023, pp. 227–237.

[29] Q. Ding, Z. Cao, H. Cao, and T. Zhao, "Enhancing large language models' machine translation via dynamic focus anchoring," *arXiv preprint arXiv:2505.23140*, 2025.

[30] T. Kocmi and C. Federmann, "Gemba-mqm: Detecting translation quality error spans with gpt-4," in *Proceedings of the Eighth Conference on Machine Translation*, 2023, pp. 768–775.

[31] N. Team, M. R. Costa-jussà, J. Cross, C. Onur, and ..., "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.

[32] C. Federmann, T. Kocmi, and Y. Xin, "Ntrex-128–news test references for mt evaluation of 128 languages," in *Proceedings of the first workshop on scaling up multilingual evaluation*, 2022, pp. 21–24.

[33] A. Anastasopoulos, A. Cattelan, Z.-Y. Dou, M. Federico, C. Federmann, D. Genzel, F. Guzmán, J. Hu, M. Hughes, P. Koehn *et al.*, "Tico-19: the translation initiative for covid-19," in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.

[34] Q. Team, "Qwen3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2505.09388

[35] Y. Luan, J. Eisenstein, K. Toutanova, and M. Collins, "Sparse, dense, and attentional representations for text retrieval," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 329–345, 2021.

[36] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, and J. Zhou, "Qwen3 embedding: Advancing text embedding and reranking through foundation models," *arXiv preprint arXiv:2506.05176*, 2025.

[37] X. H. Lù, "Bm25s: Orders of magnitude faster lexical search via eager sparse scoring," *arXiv preprint arXiv:2407.03618*, 2024.

[38] N. M. Guerreiro, R. Rei, D. v. Stigt, L. Coheur, P. Colombo, and A. F. Martins, "xcomet: Transparent machine translation evaluation through fine-grained error detection," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 979–995, 2024.

[39] T. Sellam, D. Das, and A. Parikh, "Bleurt: Learning robust metrics for text generation," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 7881–7892.

[40] A. Pu, H. W. Chung, A. P. Parikh, S. Gehrmann, and T. Sellam, "Learning compact metrics for mt," in *Proceedings of EMNLP*, 2021.