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

A Two-Stage Refinement Framework integrating Implicit Context and Explicit Feedback

Shibo Zhang[1234], Mieradilijiang Maimaiti[1234*], Zhengyi Guo[1234], Dezhi Wang[1234], Wu Le[5], Zhuofei Xie[5], Jiawei Chen[5], Wushouer Silamu[1234]

[1] School of Computer Science and Technology, Xinjiang University;
[2] Xinjiang Laboratory of Multi-Language Information Technology;
[3] Xinjiang Multilingual Information Technology Research Center;
[4] Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi, China
[5] Integrated Laboratory for Space, Air, and Ground Systems;

* Corresponding author: miradeljan51@xju.edu.cn

CATALOG

01 **Background and Motivation**
Why it matters

02 **Research Goal and Contributions**
What we build

03 **Methodology**
How we build it

04 **Experimental Setup**
How we evaluate it

05 **Results**
What we find

06 **Conclusion and Q&A**
Takeaways

Background and Motivation

Research gap

LLMs excel in In-Context Learning (ICL) but face persistent errors and hallucinations in low-resource machine translation (MT).

Challenges

While RAG improves quality, generated translations still exhibit diverse errors. Effectively refining these errors without massive retraining remains a critical hurdle for low-resource languages.

Our goal

Introduce a two-stage refinement framework that strategically exploits retrieved contextual signals and explicit quality feedback to improve low-resource MT.

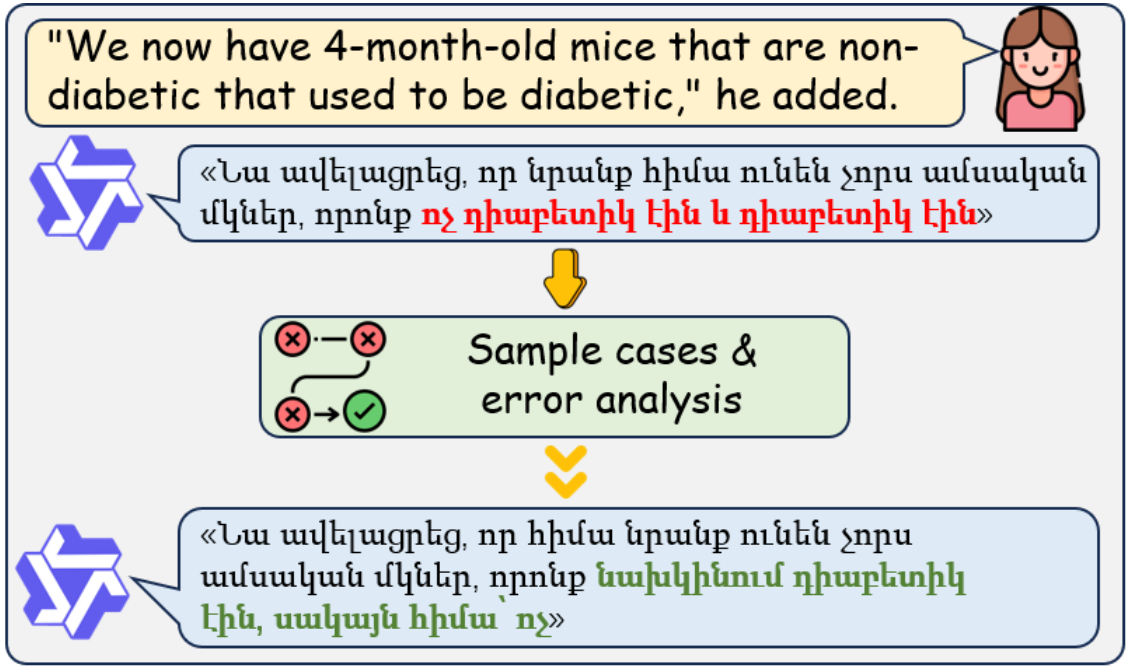


Fig. 1. Standard LLM Generation vs. Two-Stage Refinement Framework(English to Armenian).

Research Goal and Contributions

Goal **Improve LLM-based translation quality in low-resource settings through a structured, retrieval-augmented refinement process.**

01 **Two-Stage Framework**

Present a 2-stage refinement framework integrating retrieval-augmented contextual guidance.

02 **Implicit Refinement**

Leverage retrieved parallel examples and auxiliary knowledge (topics/keywords) to correct major initial translation errors.

03 **Explicit Refinement**

Introduce an MQM-based quality feedback mechanism to iteratively identify and revise residual errors.

04 **Benchmark validation**

Validate the framework across 8 low-resource languages on 3 benchmarks, demonstrating consistent improvements.

Methodology

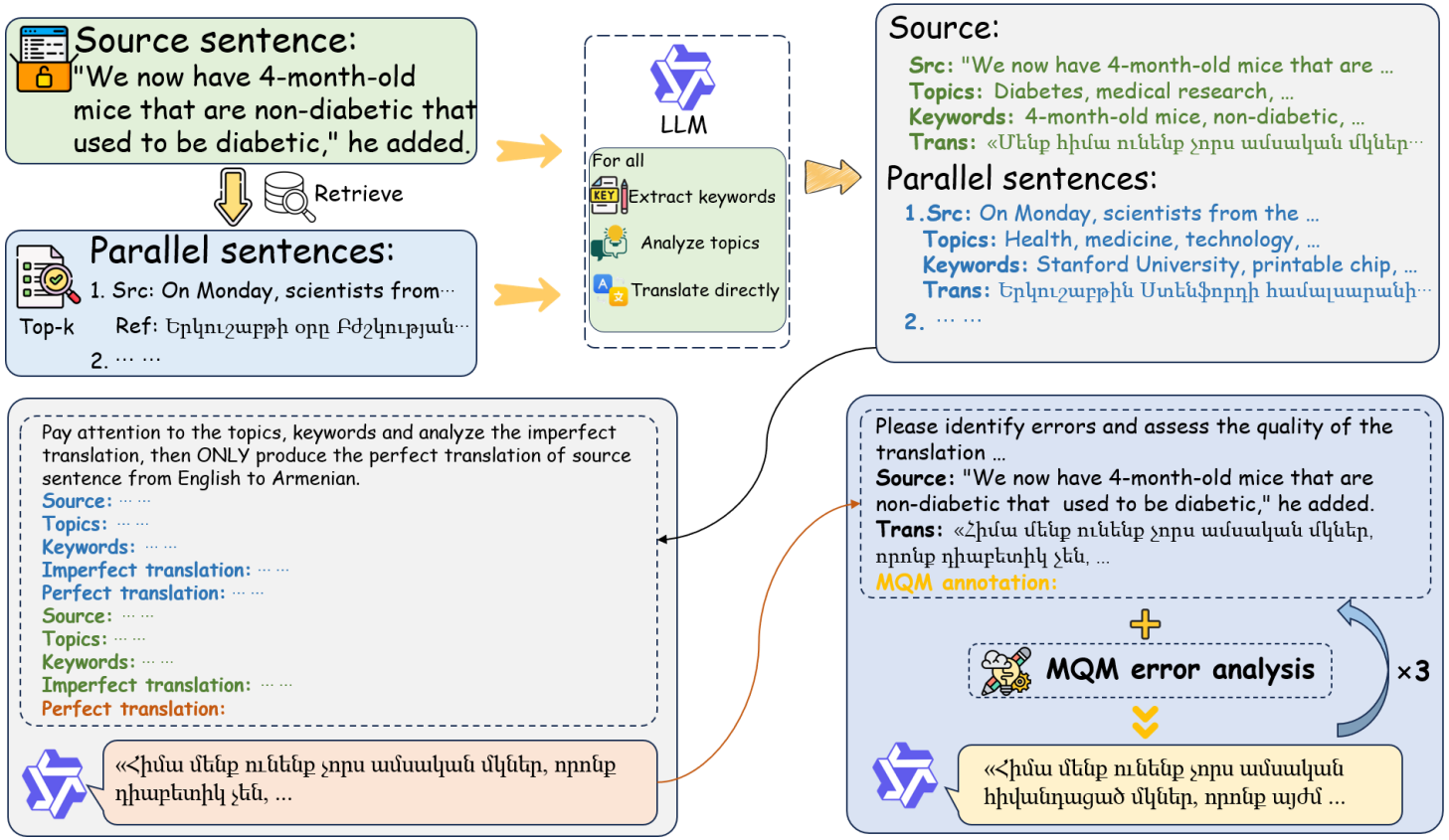


Fig. 2. Overview of the pipeline.

Three-stage workflow

1 Initial Generation & Retrieval

Retrieve top-k parallel examples and extract auxiliary knowledge from all the sentences while translating them directly.

2 Implicit Refinement

Formulate contrastive demonstrations to contextually correct the initial imperfect translation.

3 Explicit Refinement

Apply MQM-based error analysis iteratively (T=3) to polish semantic adequacy and fluency.

Experimental Setup

Benchmarks

FLORES-200

Use the dev split as the retrieval pool and evaluate on the devtest split.

NTREX-128

Use the first 1,000 pairs for evaluation and the remaining 997 pairs as a retrieval pool.

TICO-19

Specialized medical domain. Use the dev set (971 pairs) as the retrieval pool and test in four languages on testset.

Target languages

Armenian, Azerbaijani, Hebrew, Lao, Khmer, Tamil, Urdu, Bengali

Metrics

XCOMET-XL: Semantic Adequacy
BLEURT: Semantic Equivalence

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

Implementation Details.

Experiments: Main Results

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**

Results on FLORES-200(Above) and NTREX-128(Below) from English to X.

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**

Key findings

Our framework consistently achieves the competitive performance among LLM-based baselines.

Significant gains observed: +8.0 XCOMET for Armenian (FLORES-200) and +9.1 for Hebrew (NTREX-128) compared to the 0-shot baseline.

Takeaway

The two-stage refinement approach provides robust and consistent gains across multi-domain and news scenarios in extremely low-resource settings.

Experiments: Generalization

Medical Domain Adaptability

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

Results on TICO-19 (English to X)

Results

Despite the high density of specialized medical terminology, our framework maintains superior adaptability, yielding the highest XCOMET scores across all four target languages.

Cross-Lingual Shift (Chinese to X)

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

Results on TICO-19 (Chinese to X)

Results

Successfully handles the challenging shift to the Chinese source language, proving the refinement ensures correct syntactic and semantic integration.

Key findings

The framework generates high-quality, semantically coherent translations even under the dual challenges of linguistic changes and specialized technical content.

Analysis and Evidence

Ablation Study

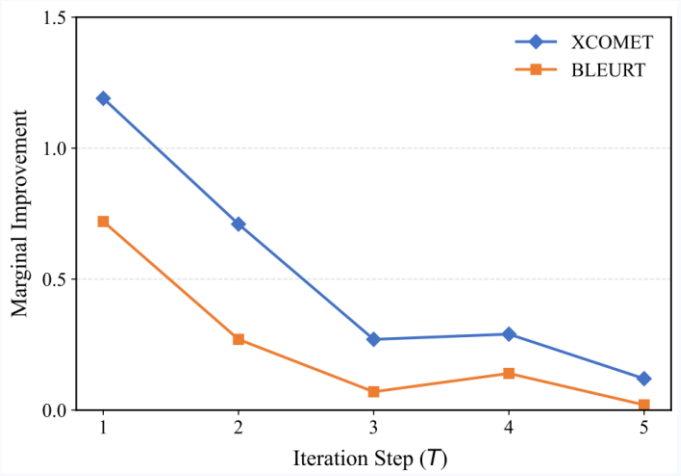
Settings	Average (8 langs)	
	XCOMET	BLEURT
w/o	65.36	68.25
w/	**65.71**	**68.48**

Knowledge Integration: Removing explicit semantic constraints leads to consistent degradation in both metrics.

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**

Explicit Refinement Impact: The addition of Stage 2 yields massive improvements.

Iteration Effect



Performance gains converge at T=3 iterations, finding the optimal trade-off.

Case study

	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.**
Reference	מועצת ההתעמלות של ארהב וההועד האולימפי שותפים לאותה מטרה - להפוך את ענף ההתעמלות, וענפים אחרים, לבמה שיותר בטוחים לספורטאים לשאוף להגשים את חלומותיהם בסביבה בטוחה, חיובית ומועצמת.
Vanilla RAG	להפוך את ספורט הגימנסטיקה, וכן ספורט אחרים, ליותר – יש להם את אותו מטרה américain גימנסטיקה של ארצות הברית וההועד האולימפי כדי שיתאמצו למשהו שמחפשים בסביבה בטוחה, חיובית ומעודדת, אתletes 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 וההועד האולימפי כדי שיתאמצו לרצונותיהם בסביבה בטוחה, חיובית ומעודדת, אתletes 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 וההועד האולימפי האמריקאי מחפשים את אותה מטרה בסביבה בטוחה, חיובית ומאפשרת חלומותיהם אחרי כדי שיתאימו, אתלטים ככל האפשר עבור. (Corrected terminology "athletes" & improved idiomatic phrasing)

Stage 1 implicitly fixed a hallucinated named entity (French "américain"), but a mixed-script morphological error ("letes") persisted. Stage 2 successfully resolved it into correct Hebrew ("athletim") while optimizing idioms.

Conclusion and Q&A

Takeaways

- We propose a two-stage refinement framework (Implicit + Explicit) to address the challenges of low-resource machine translation.
- Our approach effectively synergizes retrieval-augmented contextual guidance with the model's intrinsic MQM-based self-correction reasoning.
- Extensive experiments show consistent outperformance over competitive baselines across multi-domain, news, and specialized medical datasets.
- Future work will prioritize optimizing efficiency and adaptive strategies to balance translation quality with inference costs.

Thank you! Questions are welcome.

Code & Dataset

Corresponding author:
miradeljan51@xju.edu.cn

Supported by NSFC (Grant Nos. 62406316, 201704041014, and 201404041254) and the Xinjiang “Tianchi Talent” Recruitment and Introduction Program.