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

Shibo Zhang[1,2,3,4], Mieradilijiang Maimait[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 Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi, China
5 Integrated Laboratory for Space, Air, and Ground Systems

## BACKGROUND

**Context:** LLMs exhibit strong generalization in NLP tasks, including MT. RAG enhances this capability by using retrieved external examples.
**Problem:** In low-resource scenarios, empirical studies indicate that generated translations still exhibit diverse and persistent errors.
**Motivation:** Retrieved parallel pairs contain implicit correction signals. Comparing model-generated outputs with reference translations can implicitly guide the model to refine its own errors.

## INTRODUCTION

LLMs have shown promise in machine translation, but in low-resource scenarios, generated translations still exhibit persistent errors. We present a two-stage refinement framework:

- Implicit Refinement: Corrects major errors using retrieved parallel examples and auxiliary knowledge.
- Identifies and revises residual errors iteratively using quality-oriented feedback (MQM).

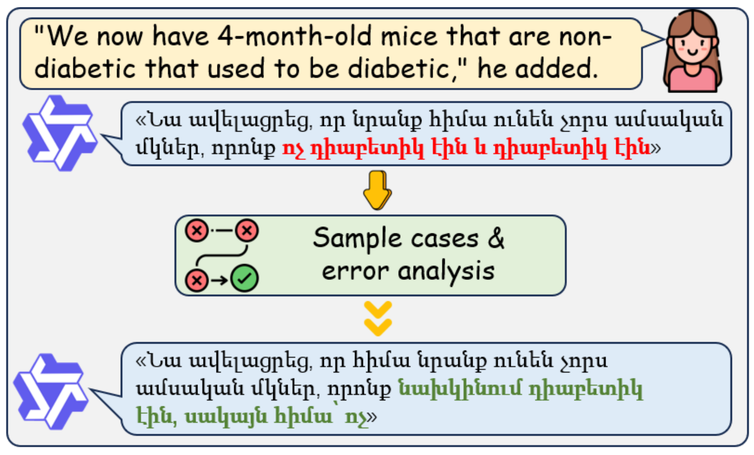



**Figure 1: Example of English to Armenian translation.**

## METHOD

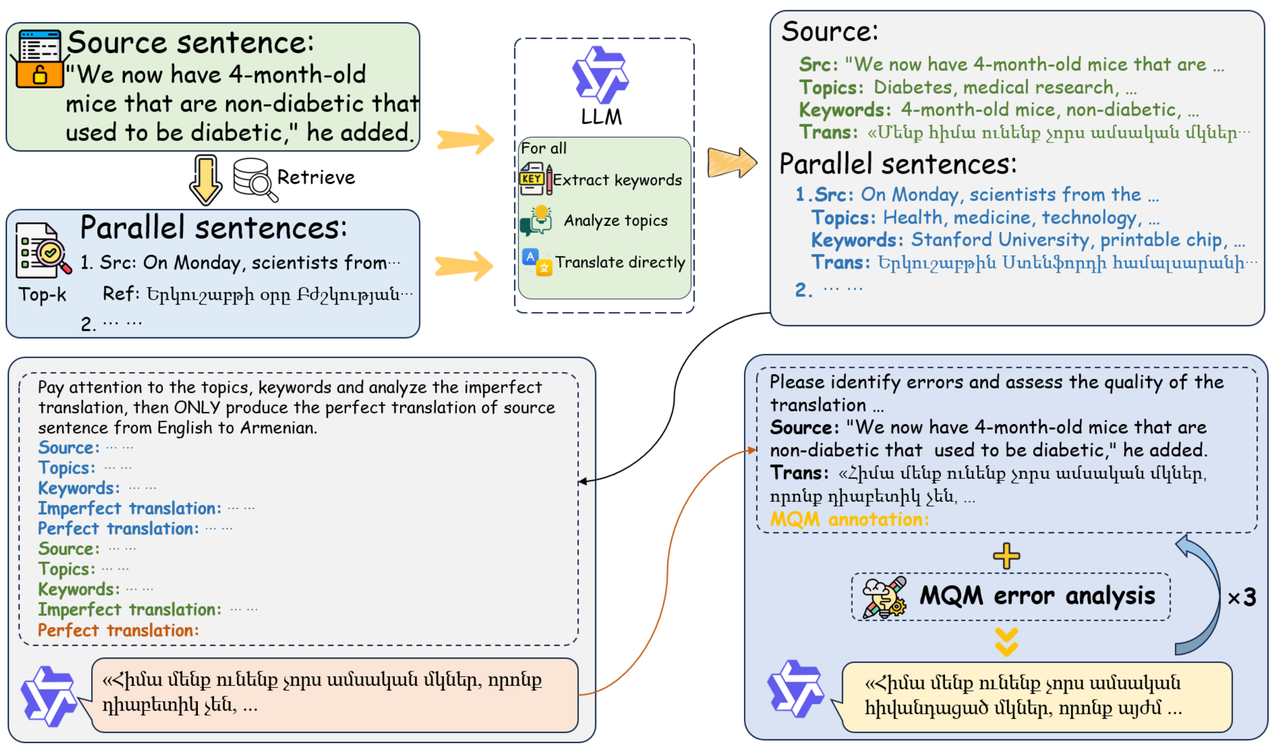



**Figure 2: Overview of pipeline**

### Implicit Refinement

- **Retrieval**: Use BM25 and Dense retrieval to fetch parallel pairs.
- **Knowledge Extraction**: Extracts auxiliary knowledge $\mathcal{K}(x)$, including topical information and keywords, to constrain the information structure.
- **Implicit Correction**: Constructs structured contrastive demonstrations by presenting the model's own imperfect translations alongside high-quality references. The refinement context is formulated as:

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

### Explicit Refinement

- **MQM Feedback**: Act as an automated critic to generate MQM annotations, which pinpoint precise error spans, classify error types, and provide natural language explanations for the corrections.
- **Iterative Revision**: Apply feedback to update the translation iteratively (T=3 iterations for optimal quality/efficiency).

### Implementation Setup

- **Unified Pipeline**: The entire framework is driven by a single base LLM.

## RESULTS

**Main Results:**

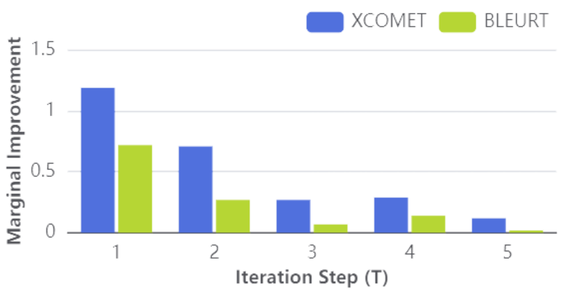
- Results on FLORES-200 consistently outperforms competitive LLM-based baselines across multi-domain texts, demonstrating superior semantic fidelity and robustness.
- Domain & Language Generalization: Results on Tico-19 effectively handles specialized medical texts (TICO-19) and the challenging Chinese-to-X language shift, securing the highest XCOMET scores across all target languages.

| 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 1: Result on FLORES-200 (English → XX)**

| 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 |

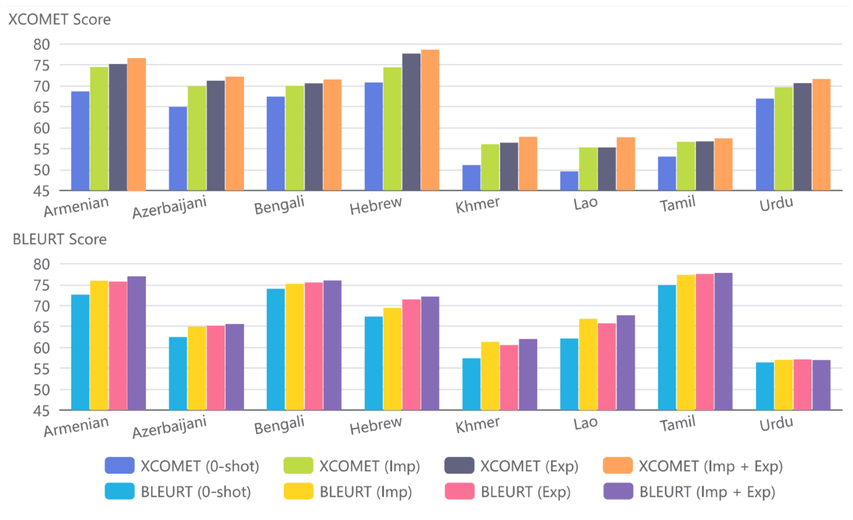
**Table 2: Result on TICO-19 (Chinese → XX)**




**Optimal Trade-off:** The average marginal improvement of XCOMET and BLEURT across 8 languages shows that translation quality significantly improves in the initial steps but converges after the **third** iteration.

| Settings | Average (8 langs) | |
|---|---|---|
| | XCOMET | BLEURT |
| 0-shot | 61.48 | 65.87 |
| Vanilla RAG | 64.27 | 67.93 |
| **Implicit Refinement** | **65.71** | **68.48** |
| - w/o knowledge | 65.36 | 68.25 |
| - w/o imperfect translation | 63.66 | 67.49 |

**Ablation on Implicit Refinement:** Removing the imperfect translation causes a pronounced performance drop, and omitting auxiliary knowledge degrades semantic fidelity show that both components are beneficial for effective implicit correction.




**Synergy of Two-Stage Refinement:** Both implicit and explicit refinement independently enhance translation quality over the 0-shot baseline. Combining both stages yields the highest XCOMET and BLEURT scores across all languages.

## CONCLUSION

We introduce a two-stage refinement framework for low-resource MT that synergizes RAG-based implicit correction with explicit iterative feedback. Experiments demonstrate consistent superiority over competitive baselines across three benchmarks, proving robust even in specialized medical domains. Future work will focus on optimizing computational efficiency to balance translation quality with inference cost.

**Acknowledgement**: Supported by NSFC (Grant Nos. 62406316, 201704041014, and 201404041254) and the Xinjiang "Tianchi Talent" Recruitment and Introduction Program.*
**Corresponding author:** miradeljan51@xju.edu.cn