Are Reasoning and Refining Really Helpful for LLM-based Machine Translation in Low-Resource Scenarios?

Dezhi Wang[1], Mieradilijiang Maimaiti[1]*, Yi Chen[1], Shibo Zhang[1], Zhengyi Guo[1], Zhexin Li[1], Wushouer Silamu[1]

[1]School of Computer Science and Technology, Xinjiang University, China

{kyriezhizhi, chenyi, zhangshibo, guozhengyi, lizhexin}@stu.xju.edu.cn, {miradeljan51, wushour}@xju.edu.cn

Abstract

Large language models (LLMs) perform well on high-resource languages but struggle with low-resource ones. Direct prompting often produces translations in a single step, without leveraging relevant linguistic information or reasoning. We propose KnT (Knowledge-to-Translation), a refined chain-of-thought (CoT) approach in which the model first generates translation-relevant linguistic knowledge, including a demonstration, topic information, a dictionary chain, and an explanation of a difficult word, then incorporates it into different prompts to guide translation. Multiple candidates are generated, with the best selected via a reference-free quality estimator. Experiments on FLORES-200, IWSLT, and CCMatrix, among others, show consistent improvements over strong baselines on common translation metrics such as XCOMET-XL and BLEURT-20, demonstrating that reasoning over extracted linguistic knowledge effectively improves low-resource translation. The code and data are publicly available[1].

1 Introduction

Large language models (LLMs) show strong machine translation (MT) capabilities (Marmonier et al., 2025; Pei et al., 2025), yet often perform translations implicitly without explicit linguistic reasoning. While chain-of-thought (CoT) prompting addresses this by decomposing tasks into interpretable steps, its application in MT remains limited, especially for low-resource scenarios (Wei et al., 2022). Exploring this direction is crucial, as explicit reasoning can elicit rich intermediate linguistic knowledge to mitigate data scarcity and fully unlock the latent multilingual potential of LLMs.

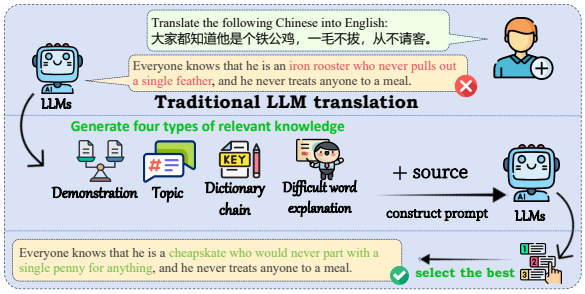


Figure 1: Multi-source reasoning for low-resource language translation.

While LLMs excel in high-resource settings, their performance declines in low-resource scenarios because of limited parallel data. This limitation hinders cross-lingual generalization and leads one-shot generation to fail in lexical choice and coherence. Although dictionary-augmented or example-augmented prompting shows promise (Merx et al., 2024), limited performance on human evaluations suggests that low-resource translation necessitates structured reasoning beyond surface-level prompting. Recent work explores reasoning and knowledge-augmented prompting for low-resource MT. A multi-aspect framework was proposed to extract keywords, demonstrations, and topics, yet it operates at a coarse level without fine-grained reasoning (He et al., 2024). For specific lexical challenges, multi-step prompting has been shown to improve idiom translation through explicit interpretation, although its scope remains limited to figurative language (Rezaeimanesh et al., 2025). Similarly, chain-of-dictionary (CoD) (Lu et al., 2024) enhances accuracy for rare words via

*Corresponding author.

[1]https://github.com/kyriezhizhi-ops/knt

multilingual dictionary chains but relies on external knowledge, including multi-grained retrieval (Pei et al., 2026), rather than dynamic generation.

In this work, we propose KnT, a reasoning-driven framework for low-resource machine translation built on a refined chain-of-thought approach. As illustrated in Figure 1, instead of translating the source sentence in a single step, our approach first prompts the model to generate multiple types of translation-relevant knowledge. These include topic-level analysis, an illustrative translation demonstration, a cross-lingual dictionary chain, and an explanation of difficult words. Each type of knowledge is then independently incorporated into the translation prompt, yielding one candidate translation per knowledge type. To determine which knowledge is most effective for a given input, we introduce a knowledge selection stage that scores all candidates with the reference-free estimator CometKiwi-XL. We output the highest-scoring candidate as the final translation, so no predefined weights or heuristics are needed. This framework yields a fully self-contained, reasoning-based translation process in which the model generates, uses, and selects knowledge, enabling robust translation under low-resource conditions. We evaluate our approach on six benchmarks, namely FLORES-200, IWSLT, CCMatrix, NTREX-128, TICO-19, and CCMT-2025. Our method demonstrates competitive performance against strong baselines, achieving superior results on XCOMET-XL and BLEURT-20 and leading on chrF++ in most directions.

Our contributions are threefold:

- We propose KnT, a multi-source reasoning framework for low-resource MT that elicits linguistic knowledge to guide the translation process.
- We introduce a fine-grained linguistic guidance strategy that provides explicit word-level interpretations as semantic anchors to resolve lexical ambiguity.
- We propose an integrated CoD mechanism that adaptively elicits lexical reasoning paths to enhance cross-language generalizability.

2 Related Work

2.1 Prompting Strategies for Translation

A major line of work focuses on designing effective prompts to elicit translation capabilities from LLMs. Early studies systematically evaluated basic prompting templates (Zhang et al., 2023; Vilar et al., 2023). Subsequent efforts have introduced various sophisticated schemes. These include decomposing translation via lexical anchors (Lu et al., 2024), dynamically selecting in-context examples (Tang et al., 2025), formalizing example retrieval criteria (Agrawal et al., 2023), and grounding prompts in bilingual phrase dictionaries (Ghazvininejad et al., 2023). To scale prompt design, extensive evaluations have been conducted across open-source LLMs (Leiter and Eger, 2024). Similarly, meta-learning frameworks (Pan et al., 2024) now employ probability-driven meta-graphs to facilitate unsupervised translation. Meanwhile, researchers have investigated interactions between explicit instruction and implicit learning (Marmonier et al., 2025), alongside decomposition-based frameworks tailored for low-resource scenarios (Zebaze et al., 2025). Sensitivity to prompt structure has been further revealed through a deep case study on Manchu (Pei et al., 2025), while multi-turn refinement and human-in-the-loop editing have been explored (Chen et al., 2024; Raunak et al., 2023), signaling a shift toward interactive paradigms.

2.2 Cultural and Stylistic Translation

Translating non-literal expressions remains a critical challenge in machine translation. Recent work has focused on idiomatic translation: a curated Persian-English idiom dataset has been introduced alongside evaluations of prompting and hybrid LLM-NMT pipelines (Rezaeimanesh et al., 2025); similarly, failure modes in Chinese idiom translation contexts have been analyzed (Yang et al., 2025b). Beyond idioms, GPT-4's reasoning has been leveraged for the culturally grounded rendering of East Asian expressions (Tang et al., 2024), while stylistic transfer has been modeled to preserve discourse norms (Havaldar et al., 2025). Cultural adequacy has likewise become an evaluation target, from anno-

tated culture-specific items (Yao et al., 2024) to cultural effectiveness on social media UGC (Wu et al., 2026) and SNS-specific benchmarks (Guo et al., 2025). Further research demonstrates that activation-level prompting can effectively control stylistic traits during translation (Wang et al., 2023). Additionally, a cognitive-theoretic agent collaboration framework has been proposed to enhance contextual coherence across translation tasks (Li et al., 2025b). More broadly, linguistically diverse prompting is advocated as a way to bridge the English-dominance gap (Nguyen et al., 2024).

2.3 Knowledge Integration for Translation

To address LLMs' limitations in structured linguistic knowledge, recent works integrate external resources during translation. Code-augmented grammar books guide translation in extremely low-resource settings (Hus and Anastasopoulos, 2024; Zhang et al., 2025). Another approach embeds grammatical constraints directly into in-context learning to ensure syntactic correctness (Ramos et al., 2025). For domain adaptation, retrieved domain-specific phrases provide the necessary conditioning for generation (Sun et al., 2022), while a systematic comparison exists between retrieval-augmented and generative knowledge integration approaches (Li et al., 2025a), with retrieval followed by refinement for low-resource translation (Zhang et al., 2026). Earlier, low-resource NMT compensated for scarce supervision through both transfer from high-resource languages and data augmentation (Maimaiti et al., 2019, 2020, 2021, 2022). However, significant dual failures in retrieval and semantic understanding persist (Elsner et al., 2024), highlighting gaps in robustness and generalization for truly low-resource languages.

Together, these works illustrate the rapid evolution of LLM-based machine translation, from simple prompting to knowledge-augmented, iterative, and culturally aware systems, and to benchmarks for extremely low-resource and multimodal settings (Hou et al., 2026). Yet the knowledge guiding translation is still either coarse, such as keywords or topics, or drawn from external resources such as dictionaries or grammar books. Eliciting fine-grained knowledge from the model itself therefore remains underexplored, which is the setting we address in this work.

Algorithm 1 Multi-Knowledge Integration

Require: Source sentence s; LLM f_θ; quality estimator CometKiwi-XL$(\cdot)$
Ensure: Final translation y^*
Generate translation-relevant knowledge
Generate knowledge set $\mathcal{K}(s) = \{K^{\text{demonstration}}(s), K^{\text{topic}}(s), K^{\text{chain-of-dictionary}}(s), K^{\text{difficult word}}(s)\}$ using f_θ
Generate translation candidates
$y_{\text{Direct}} \leftarrow f_\theta(s, \text{Prompt}_{\text{Direct}}(s))$
Initialize $\mathcal{Y}(s) \leftarrow \{y_{\text{Direct}}\}$
for all $K \in \mathcal{K}(s)$ **do**
  Construct prompt $P_K = \text{Prompt}(s, K)$
  $y_K \leftarrow f_\theta(s, P_K)$
  $\mathcal{Y}(s) \leftarrow \mathcal{Y}(s) \cup \{y_K\}$
end for
Quality estimation and selection
$y^* \leftarrow \arg\max_{y \in \mathcal{Y}(s)} \text{CometKiwi-XL}(s, y)$
return y^*

3 Method

3.1 Construction of CoT Prompts

As shown in Figure 2, the first step is to construct four initial chain-of-thought (CoT) prompts that guide LLMs to generate structured intermediate analysis before producing a translation. As detailed in Figure 6 in Appendix E, these templates prompt the model to produce explicit intermediate reasoning as structured linguistic knowledge.

Formally, each CoT prompt structures the input to elicit explicit intermediate reasoning from the model. Under such a prompt, the LLM performs step-wise reasoning:

$$R_t = f_\theta(R_{t-1}, s), \quad t = 1, \ldots, T$$

where R_t denotes the intermediate reasoning state at step t, and f_θ represents the LLM parameterized by θ. As a special case of this paradigm, our framework simplifies this process into a single step where R_0 is empty and $T = 1$.

This formulation follows the standard chain-of-thought paradigm (Wei et al., 2022), in which explicit intermediate reasoning supports generation tasks, rather than relying on direct decoding without intermediate reasoning.

3.2 Refinement of CoT

To operationalize the four parallel R_1 outputs, our framework uses each knowledge type as

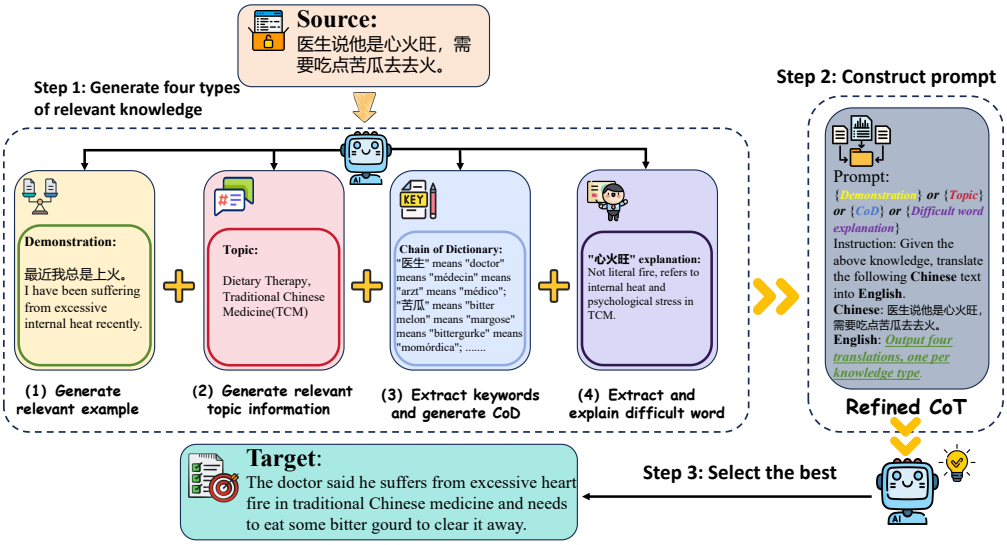


Figure 2: Construction process of refined CoT.

an additional condition to generate multiple translation candidates, and then indirectly refines the CoT output by selecting the highest-scoring candidate via CometKiwi-XL, as summarized in Algorithm 1.

The four intermediate R_1 outputs yield four translation-relevant knowledge types, $\mathcal{K}(s) = \{K^{\text{demonstration}}(s),\ K^{\text{difficult word}}(s),\ K^{\text{topic}}(s),\ K^{\text{chain-of-dictionary}}(s)\}$.

These correspond to a demonstration, an explanation of a difficult word, topic information, and a chain-of-dictionary, respectively.

Each knowledge type $k \in \mathcal{K}(s)$ is independently incorporated into the translation prompt. Formally, a translation prompt is constructed as:

$$P_k = \text{Prompt}(s, k), \quad k \in \mathcal{K}(s).$$

Each prompt conditions the model on the source sentence and a type of knowledge. Here, $P_\emptyset$ represents the standard direct translation prompt.

Applying these different prompts results in five translation candidates:

$$\mathcal{Y}(s) = \{f_\theta(P_\emptyset)\} \cup \{f_\theta(P_k) \mid k \in \mathcal{K}(s)\}.$$

Finally, we score all candidates using CometKiwi-XL, a large reference-free metric to evaluate translation quality, and select the highest-scoring final output as:

$$y^* = \arg\max_{y \in \mathcal{Y}(s)} \text{CometKiwi-XL}(s, y).$$

4 Experiments

4.1 Dataset and Evaluation

We evaluate our method on six diverse benchmarks covering various linguistic styles: 1) FLORES-200 provides high-quality translations to assess general performance on En↔Fa, Ur, Lo, Uz[2]; 2) IWSLT (TED2014) features spoken-language transcripts. We align Zh→Fa, He, Id, Vi subtitles to test conversational robustness[3]; 3) CCMatrix is a noisy web-crawled corpus used to evaluate performance on real-world data across Zh→Fa, He, Id, Vi, Bn[4]; 4) NTREX-128 covers news-domain topics to challenge low-resource scenarios in En→He,Lo,Ur,Bn[5]; 5) TICO-19 focuses on the biomedical domain, assessing cross-domain portability for En, Zh→Bn, Ur[6]; 6) CCMT-2025 is a news-domain dataset covering translation from Bo, Ug, En→ Zh, designed for real-world news translation scenarios.

We report BLEU (Post, 2018), chrF++ (Popović, 2017), COMET (Rei et al., 2022), XCOMET-XL (Guerreiro et al., 2024), and BLEURT-20 (Sellam et al., 2020) scores, cov-

[2] https://huggingface.co/datasets/openlanguagedata/flores_plus

[3] https://aclanthology.org/2014.iwslt-evaluation.1/

[4] https://opus.nlpl.eu/CCMatrix/corpus/version/CCMatrix

[5] https://github.com/MicrosoftTranslator/NTREX

[6] https://huggingface.co/datasets/gmnlp/tico19

ering both lexical similarity and neural semantic evaluation. We also utilize CometKiwi-XL (Rei et al., 2023) as the reference-free scorer for candidate selection in our framework. The exact checkpoints and repositories are listed in Appendix F.

In the result tables, **bold** and underlined values denote the best and second-best performance, respectively; XCOMET and BLEURT stand for XCOMET-XL and BLEURT-20. We test significance against the strongest baseline using paired Wilcoxon signed-rank tests at the sentence level, where † and ‡ indicate $p < 0.05$ and $p < 0.01$, respectively.

4.2 Implementation Details

We instantiate the difficult word prompt by translation direction. For EN→X, it elicits an explanation in the target language; for X→EN, it instead instructs the model to attend to lexical details in the source, since LLMs already handle English well and such explanations can be ineffective when relied on too heavily.

All experiments are run on a single NVIDIA H100 GPU using the vLLM inference engine. To ensure reproducibility, we employ greedy decoding.

4.3 Baseline

We evaluate our method against the following baselines using multiple complementary metrics.

- **Madlad** (Kudugunta et al., 2023): A specialized 10B-parameter multilingual NMT model, MADLAD-400-10B, employed as a strong non-LLM baseline to evaluate low-resource translation performance against our proposed framework.

- **Direct** (Yang et al., 2025a): A vanilla baseline that prompts the LLMs to translate directly without any auxiliary knowledge or reasoning.

- **Rerank** (Fernandes et al., 2022): A pure reranking baseline that uses the Direct prompt. We sample five candidates: four at temperature 0.3 (Moslem et al., 2023) and one by greedy decoding, and select the best using CometKiwi-XL, without any knowledge guidance.

- **MAPS** (He et al., 2024): A method that prompts LLMs with multiple types of translation-related knowledge and selects the optimal output through knowledge-based filtering.

- **CoD** (Lu et al., 2024): A prompting method that incorporates multilingual dictionary chains as explicit prior knowledge into the prompt, thereby enhancing translation of rare words.

- **TEaR** (Feng et al., 2025): A framework that improves translation quality through systematic self-refinement, where LLMs iteratively identify potential errors and refine the translation based on targeted feedback.

- **CompTra** (Zebaze et al., 2025): A strategy that decomposes the source sentence into phrases and leverages their translations as compositional guidance for the final sentence-level translation.

All prompting-based methods share the same LLM backbone within each benchmark. FLORES-200, NTREX-128, and TICO-19 use Qwen3-30B-A3B-Instruct-2507, while IWSLT, CCMatrix, and CCMT-2025 use Qwen3-32B.

4.4 Main Results

Analysis on FLORES-200 Our framework demonstrates broad advantages over direct and competitive baselines across the evaluated datasets. As shown in Table 1, under XCOMET-XL, KnT dominates most regions, generally surpassing both prompting-based methods and specialized NMT models such as Madlad. Specifically, KnT yields notable XCOMET-XL gains in linguistically challenging Middle Eastern, South Asian, and Central Asian directions. Consistent trends hold under BLEURT-20, where KnT attains the best scores on six of the eight pairs (tying with MAPS on Fa-En) and trails Madlad only in Southeast Asia. While Madlad retains an edge in Southeast Asian languages, our method surpasses all other prompting-based baselines on every pair under both metrics, confirming its robustness across diverse writing systems and morphological variations.

Methods	Middle East				South Asia			
	En-Fa		Fa-En		En-Ur		Ur-En	
	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT
Madlad	84.38	65.38	92.93	76.33	66.59	56.97	88.75	74.38
Direct	82.96	65.86	93.62	76.59	66.81	56.49	90.33	74.83
Rerank	86.74	67.24	94.15	76.85	71.36	57.43	91.09	75.23
MAPS	86.69	67.43	94.28	**77.30**	71.05	57.56	91.45	75.50
CoD	81.74	64.98	93.68	76.53	65.66	56.05	90.01	74.52
TEaR	83.85	66.06	92.81	75.93	67.82	56.57	90.03	74.37
CompTra	80.33	64.01	92.39	75.45	62.59	55.32	88.42	73.50
Ours	**88.47**‡	**67.76**‡	**94.94**‡	**77.30**	**73.50**‡	**58.06**‡	**92.30**‡	**75.80**‡
Methods	**Middle Asia**				**Southeast Asia**			
	En-Uz		**Uz-En**		**En-Lo**		**Lo-En**	
	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT
Madlad	-	-	79.75	68.64	**73.34**	**74.73**	**89.77**	**75.35**
Direct	63.43	61.05	88.14	73.53	49.97	63.17	86.45	72.66
Rerank	70.15	64.02	89.51	74.01	57.37	67.13	87.68	73.08
MAPS	69.64	64.19	90.24	74.59	57.18	67.70	81.49	64.78
CoD	62.88	61.20	88.16	73.09	44.62	48.75	86.34	72.41
TEaR	66.14	62.40	86.75	72.54	51.42	63.63	85.69	71.87
CompTra	63.13	62.47	85.93	72.74	49.83	52.57	82.20	69.87
Ours	**72.95**‡	**65.11**‡	**91.28**‡	**74.91**‡	59.75	68.17	89.24	73.73

Table 1: Main results on FLORES-200 for En↔X translation under different methods, where all LLM-based methods use Qwen3-30B-A3B-Instruct-2507, evaluated with XCOMET-XL and BLEURT-20. "-" denotes unsupported language directions for MADLAD-400-10B. Bold and underlined values denote the best and second-best performance, respectively; $^{\dagger}p < 0.05$ and $^{\ddagger}p < 0.01$ against the strongest baseline.

We provide more comprehensive results, including BLEU, chrF++, and COMET, in Table 10 in Appendix D. KnT also achieves the best chrF++ scores on most directions, and COMET shows a similar pattern, with KnT leading on all X→En directions and only marginally behind MAPS in En→X. BLEU is the metric with the largest gap; this is in line with the finding that lexical overlap metrics such as BLEU agree less with human judgment than neural metrics (Freitag et al., 2022). Finally, results across different Qwen3 model scales are provided in Table 8 in Appendix B.

Analysis on IWSLT Table 2 summarizes the performance on the IWSLT dataset for Zh-X translation. Our method (KnT) consistently outperforms the direct baseline across all pairs and achieves the highest XCOMET scores on Zh-Fa, Zh-He, and Zh-Id, validating its generalizability in cross-lingual transfer across diverse language families. Under BLEURT, KnT achieves the best score on Zh-Fa and Zh-Id and the highest average BLEURT among all prompting-based methods, indicating that its advantage is consistent rather than driven by a single metric. While MAPS leads on Zh-Vi under both metrics, our framework maintains more robust consistency overall and delivers significant gains elsewhere. These results indicate that our multi-source framework provides more stable linguistic support, effectively bridging the gap in Chinese-centric low-resource translation.

Analysis on CCMatrix Table 3 presents the translation results from Chinese to five low-resource languages on CCMatrix. In this scenario, our method demonstrates clear superiority, consistently surpassing the direct

Method	Zh-Fa		Zh-He		Zh-Id		Zh-Vi	
	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT
Madlad	70.04	40.79	71.74	**56.80**	81.68	65.83	76.53	53.25
Direct	75.02	47.08	71.24	50.93	84.93	68.18	81.60	56.21
Rerank	77.40	48.08	75.12	53.68	<u>85.64</u>	<u>68.70</u>	82.55	56.67
MAPS	<u>78.03</u>	<u>48.42</u>	<u>75.94</u>	55.03	83.23	56.32	**86.46**	**68.39**
CoD	73.56	45.34	72.58	52.26	79.44	52.32	83.91	<u>66.55</u>
TEaR	75.72	47.86	72.64	51.75	84.87	68.07	81.11	55.66
CompTra	74.33	46.82	72.29	52.61	81.11	54.59	<u>84.54</u>	63.90
Ours	**78.73**†	**48.65**	**76.56**‡	<u>55.23</u>	**85.66**‡	**68.78**	83.16	57.03

Table 2: Main results on IWSLT for Zh→X translation under different methods, where all LLM-based methods use Qwen3-32B, evaluated with XCOMET-XL and BLEURT-20. Bold and underlined values denote the best and second-best performance, respectively; $^{\dagger}p < 0.05$ and $^{\ddagger}p < 0.01$ against the strongest baseline.

Method	Zh-Fa		Zh-He		Zh-Id		Zh-Bn		Zh-Vi	
	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT
Madlad	64.82	37.09	<u>70.18</u>	**57.11**	80.30	71.00	57.38	57.36	80.97	60.23
Direct	68.16	46.34	64.94	49.14	83.02	72.94	57.12	55.04	85.74	66.26
Rerank	70.37	47.34	69.28	51.96	83.77	73.42	<u>59.50</u>	56.74	<u>86.66</u>	66.62
MAPS	<u>70.90</u>	<u>47.69</u>	69.43	53.66	<u>83.82</u>	<u>73.76</u>	59.49	<u>57.71</u>	86.27	<u>66.71</u>
CoD	66.57	44.25	66.39	49.91	81.58	71.62	56.55	54.31	82.07	61.54
TEaR	69.02	46.93	67.05	51.18	82.96	73.13	57.56	56.04	85.53	65.80
CompTra	67.56	46.40	64.94	52.33	82.33	72.94	57.16	55.21	84.39	65.31
Ours	**71.32**	**47.81**	**70.76**	<u>54.00</u>	**84.33**†	**73.85**	**60.47**†	**57.83**	**87.26**‡	**67.06**†

Table 3: Main results across multiple Zh→X translation directions on the CCMatrix dataset under different methods, where all LLM-based methods use Qwen3-32B, evaluated with XCOMET-XL and BLEURT-20. Bold and underlined values denote the best and second-best performance, respectively; $^{\dagger}p < 0.05$ and $^{\ddagger}p < 0.01$ against the strongest baseline.

baseline and all prompting-based competitors across all language pairs on both XCOMET-XL and BLEURT. Notably, KnT yields substantial XCOMET-XL gains in highly challenging directions such as Zh-Bn and Zh-He compared to the strongest baseline, MAPS. Consistent trends hold under BLEURT, where KnT attains the best BLEURT scores on four of the five pairs and outperforms MAPS on all five pairs; on Zh-He, Madlad ranks first overall. The performance gap between KnT and prompting-based baselines such as CoD and CompTra is even more pronounced across the board on both metrics. These results provide strong evidence that our multi-source reasoning via refined CoT is highly robust across domains, consistently enhancing the quality of Chinese-centric low-resource translation.

Results on Other Benchmarks Due to space constraints, the results for NTREX-128, TICO-19, and CCMT-2025 are reported in Table 5, Table 6, and Table 7 in Appendix A, respectively.

Across all six benchmarks, Ours generally outperforms Rerank on both XCOMET-XL and BLEURT-20, even though both use the same number of candidates and the same reference-free selector; since knowledge guidance is their only difference, the gains come from knowledge rather than sampling and selection.

4.5 Ablation Study

To evaluate the contribution of each module, we compare KnT against four ablated variants, each obtained by removing one knowledge type from the framework. In each variant, the candidate translation that would be guided by the removed knowledge is instead randomly sampled at temperature 0.3. For robustness, every variant is evaluated four times with different random seeds, and the reported scores are averaged over the four runs.

Figure 3 confirms that, on average, all

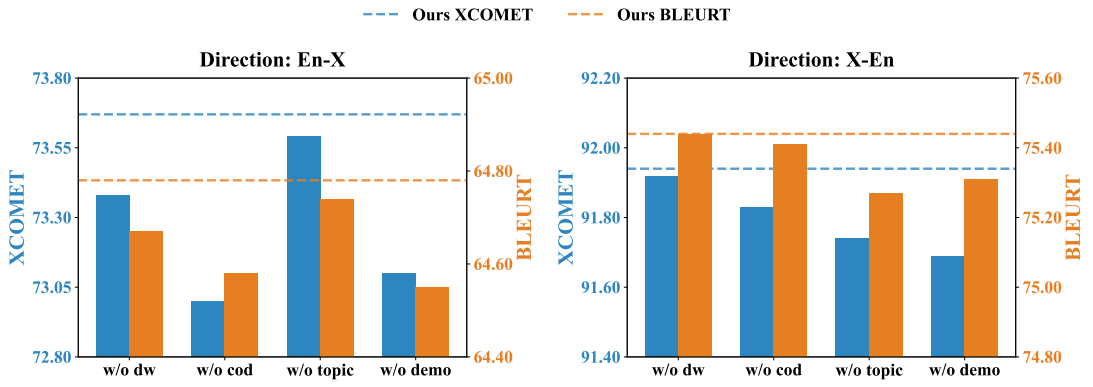


Figure 3: Module performance across translation directions. Dashed lines denote the full KnT model, and the bars in each panel correspond, from left to right, to removing the difficult word module, the chain-of-dictionary module, the topic module, and the demonstration module, respectively.

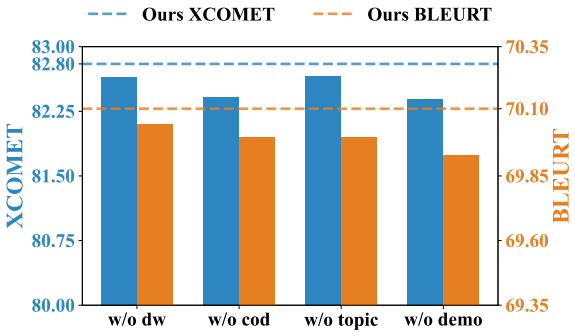


Figure 4: Module performance across all language pairs. Dashed lines and bar labels follow Figure 3.

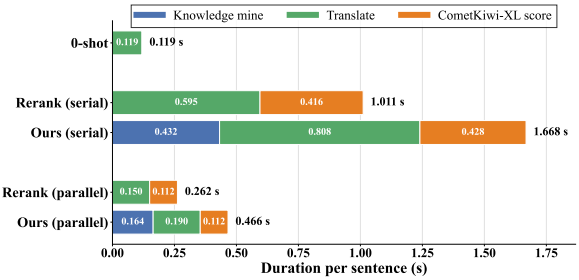


Figure 5: Per-sentence inference time breakdown for Direct, Rerank, and Ours under serial and parallel processing.

knowledge modules are beneficial. In En-X, removing chain-of-dictionary causes the largest drop in XCOMET, while removing demonstration causes the largest drop in BLEURT, indicating that structured reasoning is essential for low-resource target languages. In X-En, demonstration leads on XCOMET and topic on BLEURT, whereas the difficult word module has only a marginal effect. The topic and difficult-word modules provide finer refinements: topic knowledge resolves source ambiguity in X-En, while difficult-word explanations disambiguate lexical choices in the source. The stronger reliance on knowledge in En-X than in X-En indicates that explicit knowledge guidance compensates for the scarcity of target-side pre-training data.

Figure 4 shows the ablation results averaged over all language pairs. Removing any module degrades performance on average, with demonstration and chain-of-dictionary causing the largest drops on XCOMET and demonstration causing the largest drop on BLEURT, confirming that these two modules are the most critical. The topic and difficult-word modules incur only minor losses, offering finer refinements. These trends match the per-direction results in Figure 3.

Full results for each language pair are provided in Table 9 in Appendix C.

4.6 Inference Time

We measure the inference latency on 100 sentences randomly sampled from FLORES-200, averaged over 5 independent trials using vLLM on a single H100 GPU. Figure 5 reports the per-sentence duration breakdown of Ours (KnT) across three stages: Translate, CometKiwi-XL scoring, and Knowledge Mining.

For serial processing, the total time for Rerank and KnT is 1.011 s and 1.668 s per sentence, approximately 8.5× and 14× that of the Direct baseline (0.119 s), respectively. Since all candidate translations and knowledge

Model / Method	English Generation
Source (UZ)	Karri giyohlar hamda ziravorlar bilan birgalikda go'sht yoxud sabzavotlarga asoslangan ovqatdir.
Ref (EN)	A curry is a dish based on herbs and spices, together with either meat or vegetables.
Madlad	They are either covered with a thick layer of meat or are not.
Direct	Carryover dishes are meals based on meat or vegetables, along with herbs and spices.
Rerank	Carryover dishes are meals based on meat or vegetables, along with herbs and spices.
TEaR	Curry leaves and spices, along with meat or vegetable-based dishes.
CompTra	Carryover meals are dishes based on meat or vegetables, along with spices.
MAPS	A curry is a dish based on meat or vegetables, flavored with herbs and spices.
CoD	A dish based on meat or vegetables, combined with carrot greens and spices.
Difficult word	Meat or vegetable-based dishes are served along with side dishes and condiments.
Demonstration	Meals based on meat or vegetables, combined with herbs and spices, are typically prepared in a light and healthy manner.
Topic	A curry is a dish based on meat or vegetables, along with herbs and spices.
Ours	A curry is a dish based on meat or vegetables, along with herbs and spices.

Table 4: Case study from Uzbek to English.

types are independent, we can theoretically parallelize them by taking the maximum duration among parallel branches, yielding a per-sentence latency of 0.262 s for Rerank and 0.466 s for KnT. This corresponds to a latency reduction of 74% and 72% over their serial counterparts, respectively. These results confirm that parallel execution significantly reduces overhead, making our approach practical for real-world deployment.

4.7 Case Study

Table 4 presents a case study from Uzbek to English, comparing baseline translations with those of individual knowledge modules. Red indicates hallucinations, noise, or lexical errors; blue indicates correctly translated core concepts. Without knowledge anchoring, Direct, Rerank, and CompTra all mistranslate the pivotal concept "karri" as "carryover", while Madlad and TEaR produce hallucinated or distorted sentence structures. Notably, Rerank still fails despite selecting among candidates with CometKiwi-XL, indicating that candidate selection alone cannot compensate for the missing lexical knowledge.

The individual knowledge modules are similarly brittle: difficult word and demonstration produce substitutions or irrelevant expansions, and CoD substitutes the key ingredient with an incorrect term. In contrast, the topic knowledge correctly identifies "curry" and preserves the key ingredients, and KnT selects this candidate as the final output, yielding a translation that is both lexically accurate and syntactically faithful to the source.

5 Conclusion

We propose KnT, a multi-source reasoning framework that enhances low-resource machine translation through refined chain-of-thought and linguistic knowledge. By consistently outperforming strong baselines, KnT shows that explicit reasoning and refinement are helpful, robust, and generalizable for addressing translation data scarcity.

6 Acknowledgment

We sincerely thank all collaborators and co-authors who worked with us throughout this project and offered valuable support at various stages of the work. We also thank the anonymous reviewers for their constructive comments and insightful suggestions, which helped improve the quality and clarity of this paper. This work was supported by the National Natural Science Foundation of China (Grant Nos. 62406316, 201704041014, and 201404041254) and the Xinjiang "Tianchi Talent" Recruitment and Introduction Program (awarded to Mieradilijiang Maimaiti).

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Com-*

putational Linguistics: ACL 2023, pages 8857–8873.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190.

Micha Elsner et al. 2024. Shortcomings of llms for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. TEaR: Improving LLM-based machine translation with systematic self-refinement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3922–3938, Albuquerque, New Mexico. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chikiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Hongcheng Guo, Fei Zhao, Shaosheng Cao, Xinze Lyu, Ziyan Liu, Yue Wang, Boyang Wang, Zhoujun Li, Chonggang Lu, Zhe Xu, and Yao Hu. 2025. Redefining machine translation on social network services with large language models. *Preprint*, arXiv:2504.07901.

Shreya Havaldar, Adam Stein, Eric Wong, and Lyle Ungar. 2025. Towards style alignment in cross-cultural translation. *arXiv preprint arXiv:2507.00216*.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Shuoshuo Hou, Mieradilijiang Maimaiti, Zhexin Li, Jiaxin Wang, Ahmad Hassan, Kaishaer Jiapaer, Nilufar Abdurakhmonova, Roza Urinbayeva, Madina Mansurova, Shormakova Assem, Gulnar Murat, Wu Le, and Wushour Silamu. 2026. Vision-to-text: Benchmarking multimodal LLMs on extremely low-resource languages. In *2026 International Joint Conference on Neural Networks (IJCNN)*.

Jonathan Hus and Antonios Anastasopoulos. 2024. Back to school: Translation using grammar books. *arXiv preprint arXiv:2410.15263*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 67284–67296. Curran Associates, Inc.

Christoph Leiter and Steffen Eger. 2024. Prexme! large scale prompt exploration of open source llms for machine translation and summarization evaluation. *arXiv preprint arXiv:2406.18528*.

Bryan Li, Jiaming Luo, Eleftheria Briakou, and Colin Cherry. 2025a. Leveraging domain knowledge at inference time for LLM translation: Retrieval versus generation. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 91–106, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Weiya Li, Junjie Chen, Bei Li, Boyang Liu, Zichen Wen, Nuanqiao Shan, Xiaoqian Liu, Anping Liu, Huajie Liu, Youyan Wang, et al. 2025b. Tactic: Translation agents with cognitive-theoretic interactive collaboration. *arXiv preprint arXiv:2506.08403*.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, Zegao Pan, and Maosong Sun. 2021. Improving data augmentation for low-resource NMT guided by pos-tagging and paraphrase embedding. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(6):107:1–107:21.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-round transfer learning for low-resource NMT using multiple high-resource languages. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 18(4):38:1–38:26.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2022. Data augmentation for low-resource languages NMT guided by constrained sampling. *Int. J. Intell. Syst.*, 37(1):30–51.

Mieradilijiang Maimaiti, Yang Liu, Maosong Sun, and Huanbo Luan. 2020. Enriching the transfer learning with pre-trained lexicon embedding for low-resource neural machine translation. *Tsinghua Science & Technology*, page 1.

Malik Marmonier, Rachel Bawden, and Benoît Sagot. 2025. Explicit learning and the LLM in machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31360–31410, Suzhou, China. Association for Computational Linguistics.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. *Preprint*, arXiv:2301.13294.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516.

Shilong Pan, Zhiliang Tian, Liang Ding, Haoqi Zheng, Zhen Huang, Zhihua Wen, and Dongsheng Li. 2024. Pomp: Probability-driven meta-graph prompter for llms in low-resource unsupervised neural machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9976–9992.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 2025. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics.

Yu Pei, Mieradilijiang Maimaiti, Yi Chen, Wu Le, Zhuofei Xie, and Jiawei Chen. 2026. Dmg-rag: Dynamic multi-grained retrieval augmented generation for multi-hop qa. In *2026 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rita Ramos, Everlyn Asiko Chimoto, Maartje Ter Hoeve, and Natalie Schluter. 2025. Grammamt: Improving machine translation with grammar-informed in-context learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29920–29940.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024.

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. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Sara Rezaeimanesh, Faezeh Hosseini, and Yadollah Yaghoobzadeh. 2025. Large language models for Persian-English idiom translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7974–7985, Albuquerque, New Mexico. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Zewei Sun, Qingnan Jiang, Shujian Huang, Jun Cao, Shanbo Cheng, and Mingxuan Wang. 2022. Zero-shot domain adaptation for neural machine translation with retrieved phrase-level prompts. *arXiv preprint arXiv:2209.11409*.

Kenan Tang, Peiyang Song, Yao Qin, and Xifeng Yan. 2024. Creative and context-aware translation of east asian idioms with gpt-4. *arXiv preprint arXiv:2410.00988*.

Lei Tang, Jinghui Qin, Wenxuan Ye, Hao Tan, and Zhijing Yang. 2025. Adaptive few-shot prompting for machine translation with pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25255–25263.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. 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)*, pages 15406–15427.

Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2023. Controlling styles in neural machine translation with activation prompt. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2606–2620.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Linjuan Wu, Ruiqi Zhang, Xinze Lyu, Ye Guo, Daoxin Zhang, Zhe Xu, Yao Hu, Yixin Cao, Yongliang Shen, and Weiming Lu. 2026. Beyond literal translation: Evaluating cultural effectiveness in social media ugc. *Preprint*, arXiv:2605.25626.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Cai Yang, Yao Dou, David Heineman, Xiaofeng Wu, and Wei Xu. 2025b. Evaluating llms on chinese idiom translation. *arXiv preprint arXiv:2508.10421*.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. Compositional translation: A novel LLM-based approach for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22328–22357, Suzhou, China. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Chen Zhang, Jiuheng Lin, Xiao Liu, Zekai Zhang, and Yansong Feng. 2025. Read it in two steps: Translating extremely low-resource languages with code-augmented grammar books. *arXiv preprint arXiv:2506.01796*.

Shibo Zhang, Mieradilijiang Maimaiti, Zhengyi Guo, Dezhi Wang, Wu Le, Zhuofei Xie, Jiawei Chen, and Wushouer Silamu. 2026. Retrieve, refine, and translate: Llm-based translation for low-resource languages. In *2026 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

A Results on Other Benchmarks

Methods	En-He		En-Lo		En-Ur		En-Bn	
	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT
Direct	62.30	59.99	46.02	57.33	62.37	55.22	63.01	69.48
Rerank	<u>70.90</u>	<u>64.99</u>	<u>53.05</u>	63.16	<u>67.16</u>	56.42	<u>69.43</u>	72.20
CoD	64.47	61.36	48.56	57.96	58.44	54.01	62.63	69.46
MAPS	69.03	64.45	52.92	<u>63.77</u>	67.03	<u>56.80</u>	68.19	<u>72.30</u>
TEaR	66.00	62.28	48.35	59.88	63.93	55.27	65.72	70.63
CompTra	65.26	62.01	48.95	52.27	59.87	54.61	52.95	58.87
Ours	**72.37**‡	**65.23**‡	**55.88**‡	**65.00**‡	**69.46**‡	**57.07**	**71.22**‡	**73.07**‡

Table 5: Results on the NTREX-128 dataset (**English→XX**) evaluated by XCOMET-XL and BLEURT-20 using the Qwen3-30B-A3B-Instruct-2507 model. Bold and underlined values denote the best and second-best performance, respectively; $^{\dagger}p < 0.05$ and $^{\ddagger}p < 0.01$ against the strongest baseline.

Methods	En-Bn		En-Ur		Zh-Bn		Zh-Ur	
	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT
Direct	62.22	72.37	63.44	52.09	58.18	69.30	59.70	46.87
Rerank	<u>69.11</u>	<u>75.93</u>	<u>67.69</u>	<u>53.40</u>	<u>63.62</u>	72.85	62.28	47.27
CoD	61.43	72.38	59.70	52.23	57.83	69.32	58.20	46.82
MAPS	68.25	75.89	67.55	52.91	62.85	<u>73.01</u>	<u>63.23</u>	47.24
TEaR	65.75	74.66	65.63	52.53	60.54	72.00	61.49	**47.44**
CompTra	52.02	61.48	62.25	**54.78**	49.52	57.76	56.25	<u>47.32</u>
Ours	**70.92**‡	**76.58**‡	**69.15**‡	53.38	**64.22**‡	**73.33**†	**63.92**†	47.21

Table 6: Results on the TICO-19 dataset (**English→XX** and **Chinese→XX**) evaluated by XCOMET-XL and BLEURT-20 using the Qwen3-30B-A3B-Instruct-2507 model. Bold and underlined values denote the best and second-best performance, respectively; $^{\dagger}p < 0.05$ and $^{\ddagger}p < 0.01$ against the strongest baseline.

Method	En-Zh		Ug-Zh		Bo-Zh	
	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT
Direct	90.74	73.85	68.54	66.72	55.76	47.37
Rerank	<u>91.82</u>	**74.18**	72.29	68.00	<u>57.76</u>	47.86
CoD	89.18	73.04	69.09	66.88	56.19	<u>48.01</u>
MAPS	91.19	<u>74.03</u>	<u>73.75</u>	<u>68.77</u>	55.16	44.35
TEaR	90.02	73.17	67.79	66.14	55.19	47.33
CompTra	90.18	73.61	67.33	64.43	56.40	**48.11**
Ours	**91.95**‡	73.93	**76.51**‡	**69.49**‡	**58.09**‡	46.53

Table 7: Results on the CCMT-2025 dataset (**XX→Chinese**) evaluated by XCOMET-XL and BLEURT-20 using the Qwen3-32B model. Bold and underlined values denote the best and second-best performance, respectively; $^{\dagger}p < 0.05$ and $^{\ddagger}p < 0.01$ against the strongest baseline.

B Results Across Model Scales

	En-Fa		En-Ur		En-Lo		En-Uz		Fa-En		Ur-En		Lo-En		Uz-En	
	XC	**BL**	**XC**	**BL**	**XC**	**BL**	**XC**	**BL**	**XC**	**BL**	**XC**	**BL**	**XC**	**BL**	**XC**	**BL**
Qwen3-4B																
Direct	53.00	47.55	31.77	38.05	20.24	30.14	28.88	37.83	88.72	72.73	81.90	68.89	59.12	59.17	74.72	65.65
Ours	**64.78**	**52.37**	**40.87**	**42.43**	**24.06**	**40.22**	**38.42**	**43.15**	**91.52**	**74.06**	**86.99**	**71.45**	**68.06**	**62.46**	**81.13**	**68.49**
Qwen3-8B																
Direct	71.24	57.68	48.76	47.42	34.23	51.41	43.19	48.92	91.66	74.77	87.81	72.56	80.19	68.72	82.24	69.53
Ours	**79.59**	**61.45**	**58.63**	**50.58**	**45.13**	**60.35**	**55.68**	**54.21**	**93.76**	**75.86**	**90.63**	**74.04**	**85.17**	**70.93**	**87.49**	**72.12**
Qwen3-14B																
Direct	78.33	62.60	58.95	51.95	43.79	59.03	54.94	55.65	93.20	76.04	89.91	74.47	86.81	72.37	87.14	72.84
Ours	**84.69**	**64.85**	**67.23**	**54.00**	**55.73**	**65.99**	**66.62**	**60.06**	**94.49**	**76.73**	**91.94**	**75.56**	**89.51**	**73.69**	**90.33**	**74.23**
Qwen3-32B																
Direct	80.34	63.60	60.12	52.88	49.74	62.59	55.14	56.15	93.74	76.71	90.87	75.50	88.43	73.93	88.24	73.93
Ours	**86.80**	**66.55**	**69.64**	**55.74**	**61.54**	**68.92**	**68.55**	**61.72**	**94.84**	**77.16**	**92.81**	**76.46**	**90.84**	**74.94**	**91.72**	**75.74**
Qwen3-30B-A3B-Instruct-2507																
Direct	82.96	65.86	66.81	56.49	49.97	63.17	63.43	61.05	93.62	76.59	90.33	74.83	86.45	72.66	88.14	73.53
Ours	**88.47**	**67.76**	**73.50**	**58.06**	**59.75**	**68.17**	**72.95**	**65.11**	**94.94**	**77.30**	**92.30**	**75.80**	**89.24**	**73.73**	**91.28**	**74.91**

Table 8: Performance of KnT across Qwen3 variants (4B-32B). KnT consistently outperforms the direct baseline, with gains growing alongside model capacity, indicating larger models exploit multi-source reasoning better. Robust 4B gains confirm KnT as a general low-resource MT enhancer where XC and BL denote XCOMET-XL and BLEURT-20.

C Detailed Ablation Results

Method	**En-Fa**		**En-Ur**		**En-Lo**		**En-Uz**	
	XCOMET	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**
Ours	**88.47**	**67.76**	**73.50**	**58.06**	**59.75**	**68.17**	**72.95**	**65.11**
– w/o dw	88.09 ↓0.38	67.68 ↓0.08	73.16 ↓0.34	57.93 ↓0.13	59.66 ↓0.09	68.14 ↓0.03	72.61 ↓0.34	64.93 ↓0.18
– w/o cod	88.04 ↓0.43	67.46 ↓0.30	73.00 ↓0.50	57.83 ↓0.23	58.75 ↓1.00	68.04 ↓0.13	72.20 ↓0.75	64.98 ↓0.13
– w/o topic	88.37 ↓0.10	67.68 ↓0.08	73.24 ↓0.26	57.94 ↓0.12	59.75 ↓0.00	68.15 ↓0.02	73.00 ↑0.05	65.17 ↑0.06
– w/o demo	87.99 ↓0.48	67.61 ↓0.15	72.94 ↓0.56	58.02 ↓0.04	58.99 ↓0.76	67.79 ↓0.38	72.46 ↓0.49	64.76 ↓0.35
Method	**Fa-En**		**Ur-En**		**Lo-En**		**Uz-En**	
	XCOMET	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**	**XCOMET**	**BLEURT**
Ours	**94.94**	**77.30**	**92.30**	**75.80**	**89.24**	**73.73**	**91.28**	**74.91**
– w/o dw	94.90 ↓0.04	77.29 ↓0.01	92.28 ↓0.02	75.80 ↓0.00	89.26 ↑0.02	73.71 ↓0.02	91.22 ↓0.06	74.94 ↑0.03
– w/o cod	94.90 ↓0.04	77.25 ↓0.05	92.23 ↓0.07	75.89 ↑0.09	88.98 ↓0.26	73.60 ↓0.13	91.20 ↓0.08	74.91 ↓0.00
– w/o topic	94.79 ↓0.15	77.19 ↓0.11	92.17 ↓0.13	75.68 ↓0.12	89.19 ↓0.05	73.62 ↓0.11	90.79 ↓0.49	74.60 ↓0.31
– w/o demo	94.81 ↓0.13	77.23 ↓0.07	92.13 ↓0.17	75.77 ↓0.03	88.98 ↓0.26	73.55 ↓0.18	90.82 ↓0.46	74.69 ↓0.22

Table 9: Ablation study of KnT components across multiple low-resource language pairs on FLORES-200 using Qwen3-30B-A3B-Instruct-2507, evaluated by XCOMET-XL and BLEURT-20.

D Comprehensive Evaluation

Lang	Method	BLEU	chrF++	COMET	XCOMET-XL	BLEURT-20
EN→FA	Direct	18.62	47.23	85.63	82.96	65.86
	Rerank	18.99	47.70	86.99	86.74	67.24
	MAPS	27.70	47.56	**87.59**	86.69	67.43
	CoD	19.41	46.52	85.26	81.74	64.98
	TEaR	19.42	46.77	85.76	83.85	66.06
	CompTra	**28.67**	**47.83**	86.72	80.33	64.01
	Ours	20.03	47.65	87.37	**88.47**‡	**67.76**‡
FA→EN	Direct	30.79	59.22	87.93	93.62	76.59
	Rerank	30.88	59.48	88.10	94.15	76.85
	MAPS	**36.00**	59.48	88.32	94.28	**77.30**
	CoD	31.98	58.91	87.78	93.68	76.53
	TEaR	31.48	58.02	87.51	92.81	75.93
	CompTra	33.49	58.09	87.89	92.39	75.45
	Ours	33.34	**59.92**‡	**88.36**	**94.94**‡	**77.30**
EN→UR	Direct	15.61	41.89	80.07	66.81	56.49
	Rerank	15.83	42.48	81.52	71.36	57.43
	MAPS	**22.80**	42.22	**82.24**	71.05	57.56
	CoD	15.67	40.69	79.38	65.66	56.05
	TEaR	16.17	41.51	80.54	67.82	56.57
	CompTra	22.53	41.55	81.26	62.59	55.32
	Ours	17.16	**42.68**	82.13	**73.50**‡	**58.06**‡
UR→EN	Direct	28.21	57.06	87.13	90.33	74.83
	Rerank	28.44	57.41	87.31	91.09	75.23
	MAPS	**32.93**	57.24	**87.61**	91.45	75.50
	CoD	28.53	56.20	86.91	90.01	74.52
	TEaR	29.02	55.95	86.83	90.03	74.37
	CompTra	31.28	56.38	86.78	88.42	73.50
	Ours	30.91	**58.14**‡	**87.61**	**92.30**‡	**75.80**‡
EN→LO	Direct	10.03	35.33	74.82	49.97	63.17
	Rerank	10.00	36.18	78.64	57.37	67.13
	MAPS	**18.37**	36.73	**79.86**	57.18	67.70
	CoD	4.63	27.90	75.27	44.62	48.75
	TEaR	7.07	34.80	75.30	51.42	63.63
	CompTra	16.49	32.11	76.32	49.83	52.57
	Ours	7.51	**36.86**	79.68	**59.75**	**68.17**
LO→EN	Direct	26.43	54.41	85.94	86.45	72.66
	Rerank	27.00	54.96	86.25	87.68	73.08
	MAPS	28.67	49.62	80.54	81.49	64.78
	CoD	27.59	54.08	85.56	86.34	72.41
	TEaR	27.86	53.51	85.68	85.69	71.87
	CompTra	28.40	53.43	86.32	82.20	69.87
	Ours	**29.00**	**55.71**‡	**86.60**	**89.24**	**73.73**
EN→UZ	Direct	10.87	41.24	83.09	63.43	61.05
	Rerank	11.52	42.96	86.34	70.15	64.02
	MAPS	20.86	43.30	**87.25**	69.64	64.19
	CoD	11.15	41.85	82.76	62.88	61.20
	TEaR	9.77	41.38	84.00	66.14	62.40
	CompTra	**21.89**	**44.34**	85.24	63.13	62.47
	Ours	11.66	43.73	87.08	**72.95**‡	**65.11**‡
UZ→EN	Direct	25.80	54.50	86.33	88.14	73.53
	Rerank	26.39	55.02	86.58	89.51	74.01
	MAPS	**31.31**	54.99	87.11	90.24	74.59
	CoD	26.87	54.08	85.74	88.16	73.09
	TEaR	26.02	52.83	85.72	86.75	72.54
	CompTra	30.35	54.40	86.32	85.93	72.74
	Ours	28.32	**55.48**	**87.13**	**91.28**‡	**74.91**‡

Table 10: Experimental results of translation performance across four language pairs on the FLORES-200 test set using Qwen3-30B-A3B-Instruct-2507. All compared systems are prompting-based methods, and we evaluate all results with BLEU, chrF++, COMET, XCOMET-XL, and BLEURT-20. Bold and underlined values denote the best and second-best performance, respectively; $^{\dagger}p < 0.05$ and $^{\ddagger}p < 0.01$ against the strongest baseline.

E Prompt Templates

CHAIN OF DICTIONARY

(Case 1) Extract the keywords from the given **English** sentence and provide their translations into **Uzbek**, French, German and Portuguese.Only output the result, using 'means' to separate translations across languages and enclosing each word in double quotes.

English: *On Monday, scientists from the Stanford University School.....*

Chain-of-Dictionary: *"Stanford University" means "Stanford Universiteti" means "Université Stanford" means "Stanford Universität" means "Universidade de Stanford",.....*

.....(3-shot)

Extract the keywords from the given **English** sentence and provide their translations into **Uzbek**, French, German and Portuguese.Only output the result, using 'means' to separate translations across languages and enclosing each word in double quotes.

English: "We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.

Chain-of-Dictionary: *"mice" means "sichqonlar" means "souris" means "Mäuse" means "camundongos", ...*

DEMONSTRATION:

(Case 1) Let's write an **English** sentence related to but different from the input **English** sentence and translate it into **Uzbek**.

Input **English** sentence: *The JAS 39C Gripen crashed onto a runway....*

Output **English-Uzbek** sentence pair: *Libyan security officials say the Afriqiyah Airways..... Liviya xavfsizlik rasmiylari Afriqiyah....*

.....(5-shot)

Let's write an **English** sentence related to but different from the input **English** sentence and translate it into **Uzbek**.

Input **English** sentence: "We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.

Output **English-Uzbek** sentence pair: *"Currently, we have mice that are four months old .... Hozir bizda 4 oylik bo'lgan .....*

TOPIC:

(Case 1) Use a few words to describe the topics of the following input sentence.

Input: *The protest started around 11:00 local time (UTC+1) ....*

Topics: *Politics*

.....(5-shot)

Use a few words to describe the topics of the following input sentence.

Input: "We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.

Topics: *Health, diabetes, medical research*

DIFFICULT WORD (EN-X):

(Case 1) Extract difficult word or phrase from the **English** sentence and provide its interpretation in **Uzbek**.

English: *The agreement was reached after prolonged negotiations amid growing international pressure.*

Output: *"prolonged" explanation:kutilganidan uzoq davom etgan*

.....(3-shot)

Extract difficult word or phrase from the **English** sentence and provide its interpretation in **Uzbek**.

English: "We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.

Output: *"non-diabetic" explanation:Diabet bo'lmagan*

DIFFICULT WORD (X-EN):

Instruction: Translate the following **Uzbek** text into **English**.

Preserve the original meaning and facts; avoid paraphrase or invention, especially for names and fixed expressions.Use standard, correct English spelling and conventional forms for names and abbreviations.

Uzbek: "Hozir bizda diabetik bo'lishi kerak bo'lgan, diabetik bo'lmagan 4 oylik sichqon mavjud", deb qo'shib qo'ydi u.

English: *"Currently, we have a 4-month-old rat that needs to become diabetic, but isn't diabetic yet," he added.*

Use Knowledge To Generate Translation Results

Chain of dictionary: *"mice" means "sichqonlar" means "souris" means "Mäuse" means "camundongos", ... or*

Related **English-Uzbek** sentence pairs: *"Currently, we have mice that are four months old .... Hozir bizda 4 oylik bo'lgan ..... or*

Difficult word: *"non-diabetic" explanation:Diabet bo'lmagan or*

Topics: *Health, diabetes, medical research*

Instruction: Given the above knowledge, translate the following **English** text into **Uzbek**.

English: "We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.

Uzbek: *Output Target Text. (One Translation from Each Knowledge Type, Four Total)*

Figure 6: Illustration of the KnT prompting strategy (**English→Uzbek** example). Red text denotes few-shot demonstrations (underlined for current input), blue indicates the four knowledge types generated separately, and green represents the final translation output. A direct variant simply omits the knowledge and prompts the model to translate directly. For difficult words, EN→X extracts an explanation in the target language, while X→EN directly instructs the model to translate, emphasizing attention to source-language details, as LLMs already possess strong English ability and over-relying on difficult word explanations can be ineffective.

F Checkpoints and Repositories

We use sacrebleu[7], chrF++[8], COMET, XCOMET-XL[9], and BLEURT-20[10] for evaluation, together with CometKiwi-XL as the reference-free scorer for candidate selection.

For COMET we use the wmt22-comet-da checkpoint[11], and for CometKiwi-XL the wmt23-cometkiwi-da-xl checkpoint[12].

[7] https://github.com/mjpost/sacrebleu

[8] https://github.com/m-popovic/chrF

[9] https://huggingface.co/Unbabel/XCOMET-XL

[10] https://github.com/google-research/bleurt

[11] https://huggingface.co/Unbabel/wmt22-comet-da

[12] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl