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

CCMT 2026
Kunming, China
13–15 November

Dezhi Wang[1], Mieradilijiang Maimait[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, Urumqi, China

BACKGROUND

Context: LLMs generalize well across NLP tasks including machine translation, and prompting with in-context knowledge has become the dominant paradigm.
Problem: Low-resource languages lack dictionaries and grammar books, so knowledge-augmented translation cannot scale.
Motivation: LLMs already hold translation knowledge; refined chain-of-thought elicits it without external resources.

INTRODUCTION

LLMs have shown promise in machine translation, yet low-resource directions still suffer from word-level errors. We present KnT, a knowledge-to-translation framework that requires no external resources:

- **Knowledge Generation**: A refined chain-of-thought elicits four knowledge types from the model itself.
- **Knowledge Selection**: Each type builds one prompt; CometKiwi-XL picks the best of five candidates.



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

METHOD

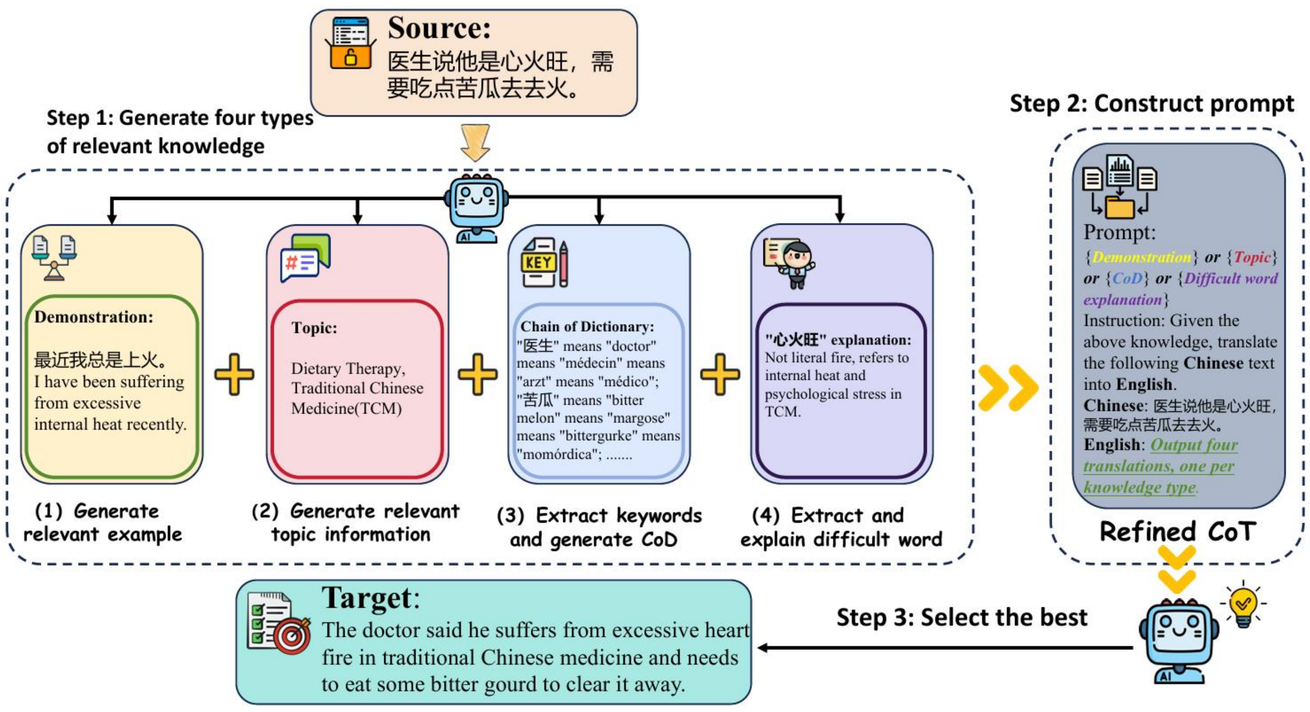


Figure 2: Construction process of refined CoT.

Knowledge Elicitation

- **CoT Prompting**: Four refined prompts elicit explicit reasoning before translating, so knowledge comes from the model itself.
- **Knowledge Types**: Demonstration, topic, dictionary chain and difficult word, each constrains a different granularity.
- **Prompt Construction**: Each type builds one prompt, forming a knowledge set:

$$\mathcal{K}(s) = \left\{K^{demonstration}(s),\ K^{topic}(s),\ K^{chain-of-dictionary}(s),\ K^{difficult\ word}(s)\right\}$$

Knowledge Selection

- **Candidate Scoring**: CometKiwi-XL, a reference-free estimator, scores all five candidates and returns the best.
- **Implicit Refinement**: Choosing the best candidate fixes word-level errors without another generation round.

Implementation Setup

- **Unified Backbone**: One Qwen3 model acts as both knowledge generator and translator.

Experiment

Main Results:

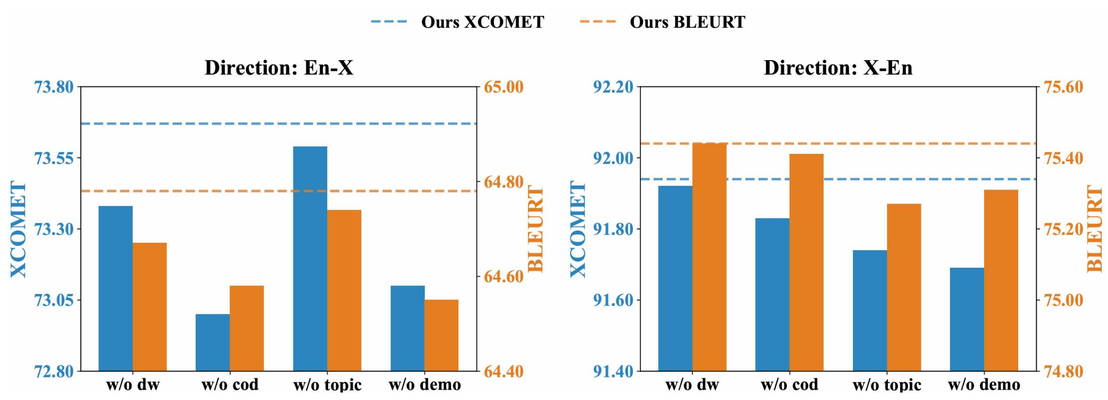
Method	En-Fa		Fa-En		En-Ur		Ur-En	
	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT	XCOMET	BLEURT
Madlad	84.38	65.38	92.93	76.33	66.59	56.97	88.75	74.38
Direct	82.96	65.86	93.62	76.59	66.81	56.49	90.33	74.83
Rerank	86.74	67.24	94.15	76.85	71.36	57.43	91.09	75.23
MAPS	86.69	67.43	94.28	77.30	71.05	57.56	91.45	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**

Method	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

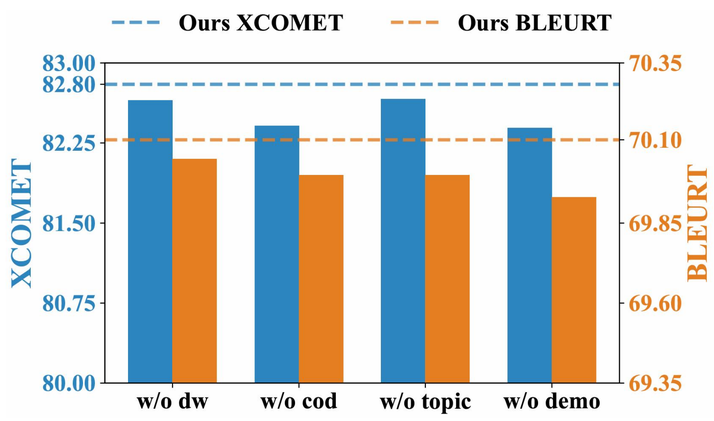
Table 1: Result on FLORES-200 (English ↔ XX)

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	85.64	68.70	82.55	56.67
MAPS	78.03	48.42	75.94	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	66.55
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	84.54	63.90
Ours	**78.73**	**48.65**	**76.56**	55.23	**85.66**	**68.78**	83.16	57.03

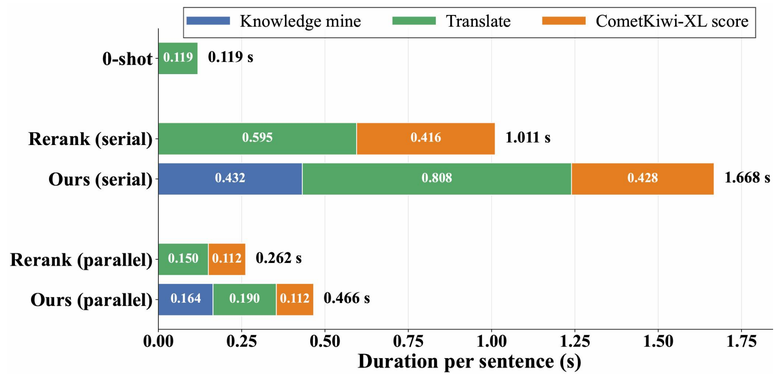
Table 2: Result on IWSLT (Chinese → XX)



Ablation by Direction:
Removing any of the four knowledge types lowers both metrics, and the loss is largest for demonstration and chain-of-dictionary.English-to-X directions rely more on the dictionary chain, while X-to-English relies more on demonstration and topic.



Average Contribution:
Averaged over all language pairs, demonstration and chain-of-dictionary contribute most, topic is moderate, and difficult word is marginal, yet every type proves necessary.



Inference Cost:
Five candidates cost more than direct prompting, but the knowledge branches are independent. Running them in parallel brings per-sentence latency down to a level practical for deployment.

CONCLUSION

We introduce KnT, a knowledge-to-translation framework for low-resource MT that elicits four types of knowledge from the model itself, removing the dependence on external dictionaries and grammar books. Experiments on multiple benchmarks show consistent gains over strong LLM-based baselines, and ablations confirm that all four knowledge types contribute. Future work will explore adaptive knowledge selection to further improve efficiency.

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