CCMT 2026
Kunming, China
13–15 November

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

A Refined Chain-of-Thought Framework for Generating and Selecting Translation Knowledge

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

* 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

Low-resource MT needs external knowledge, yet dictionaries and grammar books are scarce, so knowledge-augmented methods cannot scale.

Challenges

Existing methods inject keyword or topic level knowledge, which is too coarse to resolve the rare and ambiguous words that decide translation quality, and multiple knowledge sources are usually merged into one prompt.

Our goal

Elicit translation knowledge from the model itself with a refined chain-of-thought, and select among candidates so that neither external resources nor reference translations are needed.



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

Research Goal and Contributions

Goal **Improve LLM-based translation quality in low-resource settings through structured, self-generated multi-source knowledge.**

01 **Knowledge-to-Translation**

Present a translation framework built entirely on refined chain-of-thought prompting, without retrieval or external resources.

02 **Multi-Source Knowledge**

Elicit demonstration, topic, chain-of-dictionary and difficult word knowledge from the model itself.

03 **Reference-Free Selection**

Score the five candidates with CometKiwi-XL and return the best, requiring no reference translation.

04 **Benchmark Validation**

Validate the framework on multiple benchmarks covering eight low-resource directions, showing consistent improvements.

Methodology

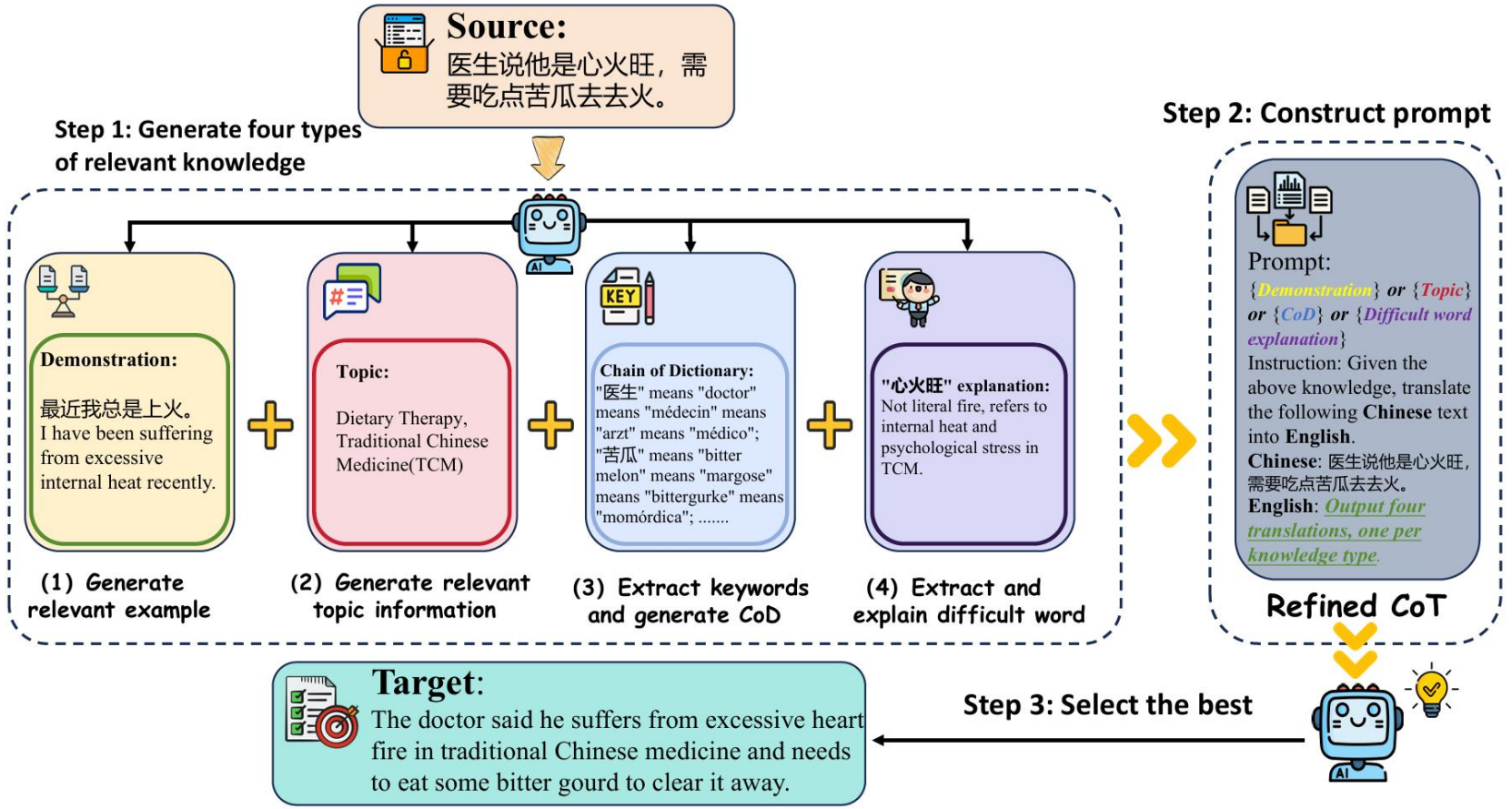


Figure 2: Construction process of refined CoT.

Three-stage workflow

1 Knowledge Elicitation

Four refined prompts make the model reason explicitly before translating, so knowledge is elicited rather than retrieved.

2 Prompt Construction

Each knowledge type builds one prompt, producing four knowledge-guided candidates plus one direct draft.

3 Candidate Selection

CometKiwi-XL scores all five candidates and the highest-scoring one is returned as the output.

Experimental Setup

Benchmarks

FLORES-200

General-domain professional translations, covering eight low-resource directions. It serves as the primary benchmark for all main comparisons.

IWSLT

Spoken TED transcripts, Chinese to Persian, Hebrew, Indonesian and Vietnamese. It tests transfer to a non-English source and a spoken register.

TICO-19

Biomedical texts, English and Chinese to Bengali and Urdu, with dense terminology. Also NTREX , CCMatrix and CCMT.

Target languages

Persian, Urdu, Lao, Uzbek, Hebrew, Bengali, Uyghur, Tibetan

Metrics

XCOMET-XL: Semantic Adequacy
BLEURT: Semantic Equivalence

Implementation details

Backbone: Qwen3 series, shared by all prompting-based methods within each benchmark.
Selection model: CometKiwi-XL, which needs no reference translation.

Experiments: 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
KnT(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
KnT(Ours)	**72.95**	**65.11**	**91.28**	**74.91**	59.75	68.17	89.24	73.73

Table 1: Results on FLORES-200.

Experiments: Main Results

Key findings

KnT outperforms every competitive LLM-based baseline across all eight directions, including neural MT systems and prior knowledge-injection or refinement methods. The margin widens where single-source prompting degrades, showing that combining knowledge sources beats relying on any one alone.

Takeaway

These gains require no retrieval corpus, no external lexicon and no reference translation, and they hold across general, news and biomedical data in both translation directions.

Experiments: Generalization

Spoken Register Adaptability

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
KnT(Ours)	**78.73**	**48.65**	**76.56**	55.23	**85.66**	**68.78**	83.16	57.03

Table 2: Results on IWSLT.

Results

IWSLT changes both the source language and the register. KnT still leads in most directions, so the prompts stay effective outside the original setting.

Experiments: Generalization

Medical Domain Adaptability

Method	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	69.11	75.93	67.69	53.40	63.62	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	73.01	63.23	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	47.32
KnT(Ours)	**70.92**	**76.58**	**69.15**	53.38	**64.22**	**73.33**	**63.92**	47.21

Table 3: Results on TICO-19.

Results

TICO-19 is biomedical, with dense terminology and little in-domain data. KnT stays best or near-best on all four targets, so the knowledge transfers to a specialized domain.

Analysis and Evidence

Ablation Study

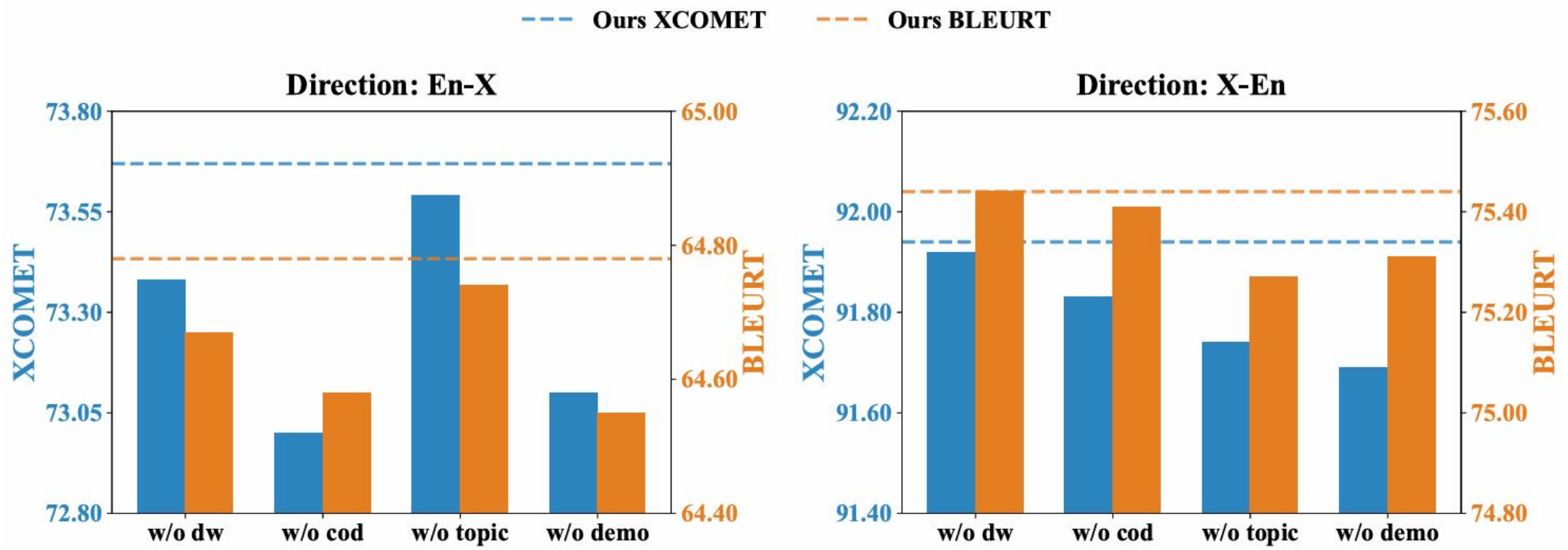


Figure 3: Removing any of the four knowledge types lowers both metrics, and the loss is largest for demonstration and chain-of-dictionary. English-to-X directions rely more on the dictionary chain, while X-to-English relies more on demonstration and topic.

Analysis and Evidence

Ablation Study

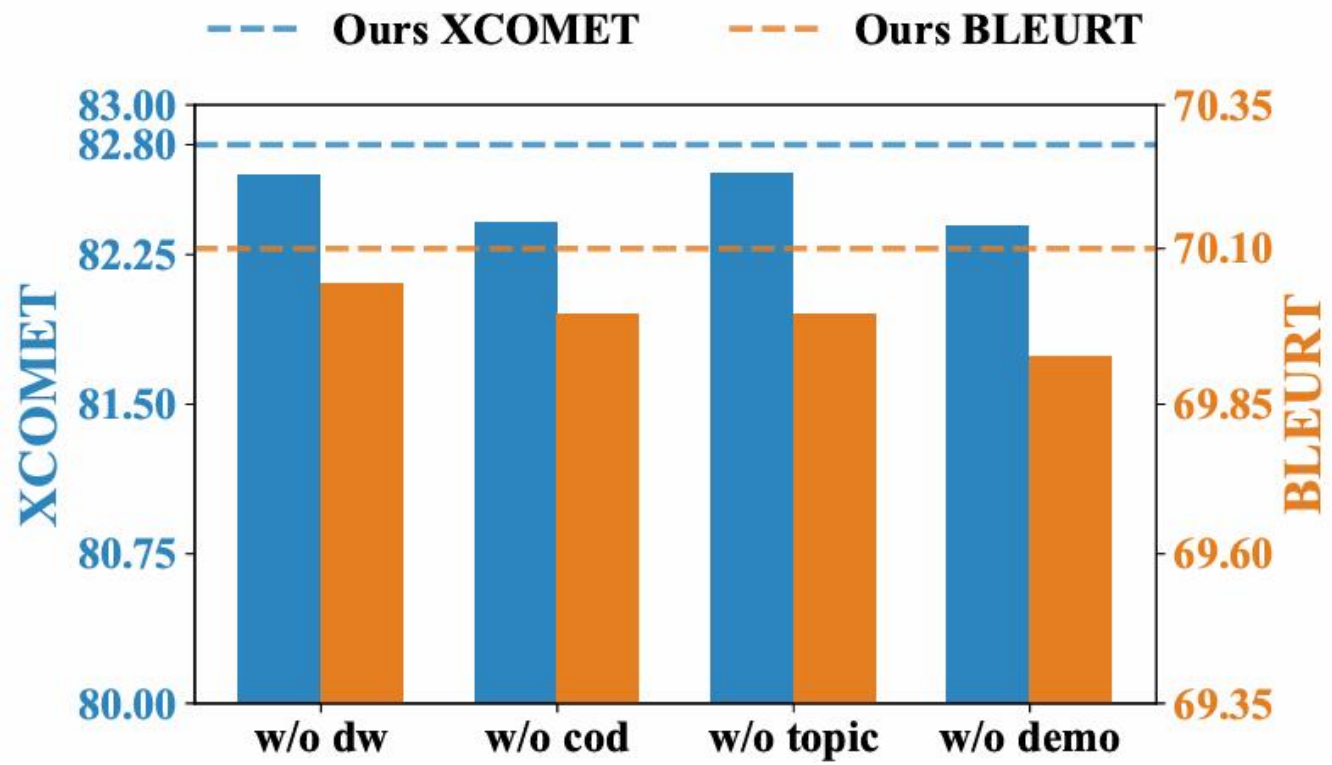


Figure 4: Averaged over all language pairs, demonstration and chain-of-dictionary contribute most, topic is moderate, and difficult word is marginal, yet every type proves necessary.

Analysis and Evidence

Inference Cost

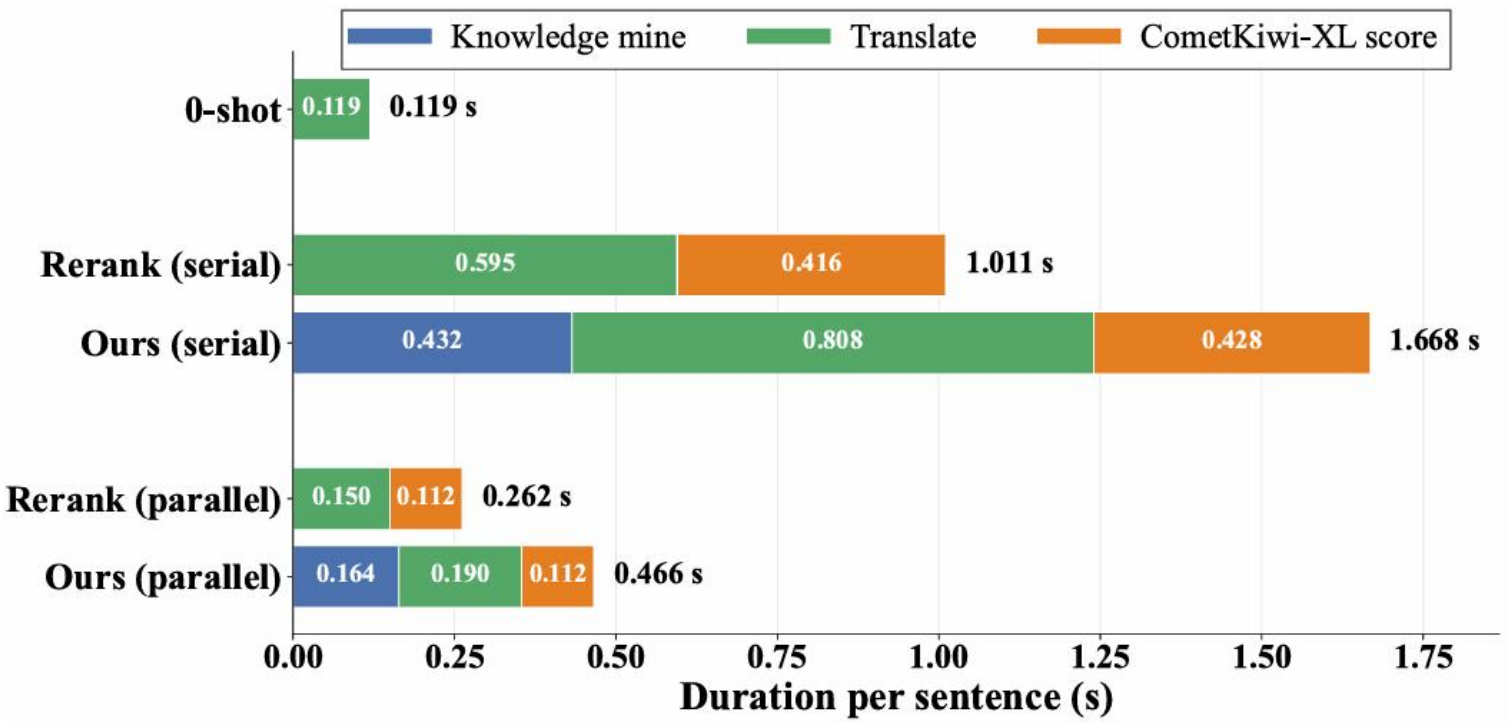


Figure 5: Five candidates cost more than direct prompting, but the knowledge branches are independent. Running them in parallel brings per-sentence latency down to a level practical for deployment.

Analysis and Evidence

Case study

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: Baselines mistranslated the core concept, rendering “curry” as “carryover” or an unrelated term. KnT selected the topic-guided candidate and recovered the correct word, preserving both main ingredients and the cooking method.

Conclusion and Q&A

Takeaways

- We propose KnT, a knowledge-to-translation framework that elicits four types of translation knowledge from the model itself.
- A refined chain-of-thought turns a single sentence into multi-source context, removing the dependence on external dictionaries and grammar books.
- Reference-free selection lets the model choose among candidates, and KnT outperforms competitive baselines on multiple benchmarks.
- Future work will explore adaptive knowledge selection to reduce redundant accumulation and further improve efficiency.

Thank you!

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

Conclusion and Q&A

Code & Dataset

Corresponding author:
miradeljan51@xju.edu.cn