Retrieval-Augmented Machine Translation with Unstructured Knowledge

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) introduces additional information to enhance large language models (LLMs). In machine translation (MT), previous work typically retrieves in-context examples from paired MT corpora, or domain-specific knowledge from knowledge graphs, to enhance MT models. However, a large amount of world knowledge is organized in unstructured documents, and might not be fully paired across different languages. In this paper, we study retrieval-augmented MT using unstructured documents. Specifically, we build RAGtrans, the first benchmark to train and evaluate LLMs' retrieval-augmented MT ability. RAGtrans contains 79K MT samples collected via GPT-4o and human translators. Besides, documents from different languages are also provided to supply the knowledge to these samples. Based on RAGtrans, we further propose a multi-task training method to teach LLMs how to use information from multilingual documents during their translation. The method uses existing multilingual corpora to create auxiliary training objectives without additional labeling requirements. Extensive experiments show that the method improves LLMs by 1.58∼3.09 BLEU and 1.00∼2.03 COMET scores. We also conclude the critical difficulties that current LLMs face with this task.[1]

1 Introduction

Retrieval-augmented generation (RAG) has grown into a practical paradigm in the development of large language models (LLMs). With the help of retrieved information, LLMs could generate more accurate and knowledge-enrich responses (Li et al., 2022; Gao et al., 2023).

Previous work brought RAG into machine translation (MT), and could be mainly classified into the following two streams: (1) *Retrieving in-context examples* (also known as "translation memory"): for a source sentence, a few studies retrieve the relevant paired sentences from bilingual corpora to enhance MT models (Zhang et al., 2018; Bulte and Tezcan, 2019; He et al., 2021; Hoang et al., 2023). Further, Cai et al. (2021) relax the bilingualism limitation, and try to directly retrieve similar target-language translations to enhance models. (2) *Retrieving knowledge triplets*: the others retrieve relevant information from knowledge graphs to let the models know domain or cultural knowledge w.r.t. the source sentences (Conia et al., 2024; Chen et al., 2024b). Despite the great success that has been achieved, a large amount of world knowledge is organized in unstructured documents, and might not be fully paired across different languages. This unstructured knowledge is neglected by previous work. For example, Wikipedia is the largest public wiki (Viégas et al., 2004), serving as an encyclopedia of world knowledge. Most of its information is listed in documents. Besides, for a piece of specific knowledge, Wikipedia does not always provide it in all languages. Though multilingual information of some general knowledge is provided, their content might be differentiated among different languages (Perez-Beltrachini and Lapata, 2021).

In this paper, we study retrieval-augmented MT using unstructured documents. Since we are the first to study this topic and previous datasets do not support the research, we first build a benchmark dataset, named **RAGtrans**. In detail, RAGtrans is collected based on Wikipedia with three key features: (i) *Knowledge-intensive sentences*: RAGtrans randomly selects 79K English sentences from Wikipedia as the source sentences, which generally come from the lead paragraphs of different Wikipedia pages, containing knowledge-intensive semantics. Thus, understanding these source sentences tends to require additional knowledge. (ii) *Useful relevant documents*: To achieve retrieval-augmented MT, for each source sentence, its following content on the Wikipedia page (in English)

[1] The codes and dataset will be released upon publication.

could serve as its relevant document. (iii) *Transferability to multilingual RAG*: Wikipedia also provides multilingual parallel content. Therefore, for a source sentence, its relevant knowledge in different languages can also serve as the relevant documents. As a result, MT models can leverage knowledge from multilingual documents beyond the source and the target languages. In this work, we choose Chinese, German, French and Czech. After collecting the source English sentences and relevant documents, 79K samples are randomly split into training, validation and testing sets with 74.5K, 2.5K and 2K samples. For training and validation samples, we employ GPT-4o (OpenAI, 2024) to collect the Chinese translation; while we employ professional human translators to perform the same process for the testing samples. Finally, RAGtrans involves 79K retrieval-augment MT samples, each of which contains an English source sentence, a document in English, Chinese, German, French or Czech, and the corresponding Chinese translation.

Based on RAGtrans, we train LLMs and evaluate their retrieval-augmented MT performance from the following settings: (1) *Golden evaluation*: providing LLMs with the golden relevant documents during data collection, and testing the translation performance. (2) *Robustness evaluation*: providing irrelevant documents to test the LLMs' robustness. (3) *Full Wiki evaluation*: Equipping LLMs with a (multilingual) retriever to first retrieve relevant documents from the whole Wikipedia, and then evaluate their retrieval-augmented MT ability.

Furthermore, during the application phase of a retrieval-augmented MT model, the model might receive multiple documents from various languages. These multilingual documents are not restricted to parallel documents and can convey diverse meanings. In light of this, we propose a multi-task training method to enhance LLMs' ability to leverage multilingual knowledge. Specifically, we design three training objectives, *i.e.*, cross-lingual information completion, self-knowledge-enhanced translation and cross-lingual relevance discrimination. Among them, cross-lingual information completion and cross-lingual relevance discrimination train LLMs to refine and judge information from multilingual documents. Self-knowledge-enhanced translation lets LLMs generate relevant knowledge in various languages for the source sentences, and then perform MT with the help of its multilingual self-knowledge. The multi-task training samples of these objectives can be automatically created from existing multilingual corpora, and do not need any additional labeling costs. Experiments on RAGtrans show that the multi-task training method improves LLMs' ability to leverage relevant knowledge. Using Qwen-2.5-7B (Yang et al., 2024) as the backbone, the retrieval-augmented MT performance is improved by 1.58∼3.09 BLEU and 1.00∼2.03 COMET scores compared with simply instruction-tuning on RAGtrans. Finally, we discuss specific challenges that current approaches faced with this task and give multiple promising directions for future research.

Our main contributions are concluded as follows:

- To the best of our knowledge, we are the first to study retrieval-augmented MT using unstructured knowledge. To this end, we construct the first corresponding benchmark dataset, *i.e.*, RAGtrans, containing 79K translation samples collected via GPT-4o and human translators.
- We propose a multi-task training method with three designed training objectives to improve LLMs' retrieval-augmented MT ability. The multi-task training samples are low-cost, and do not require additional labeling costs. Experimental results show the effectiveness of the method.
- In-depth analyses of the retrieval-augmented MT results on automatic evaluation and human evaluation provide a deeper understanding of this research direction.

2 RAGtrans

In this section, we first discuss how we select English source sentences and their relevant documents from Wikipedia (§ 2.1). Then, we introduce the details of the data translation via GPT-4o and human translators (§ 2.2). Finally, we give statistical analyses of RAGtrans (§ 2.3), and provide the details of benchmark settings (§ 2.4).

2.1 Data Selection

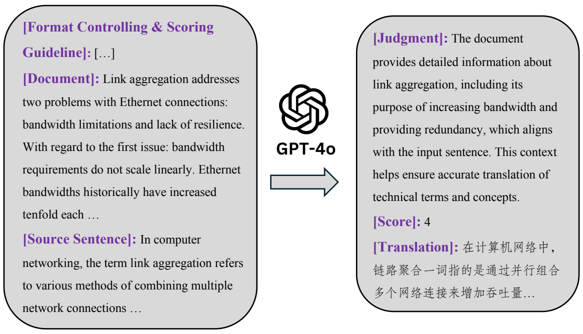
When deciding the source sentences we focus on, there are three requirements that should be met: (1) The source sentences should involve knowledge-intensive semantics, otherwise, they might be trivial to translate and do not need additional knowledge. (2) It should be convenient to collect their relevant documents from existing resources, otherwise, annotating relevant documents is labor-intensive. (3) It should also be possible to collect relevant documents in other languages. This is because world knowledge is recorded in multilingual form. If we restrict the language of the retrieved documents, the practicality will decrease.

After carefully comparing existing open-source resources, we decide to select both source sentences and relevant documents from Wikipedia. Formally, we denote an English document on a Wikipedia page as $D^{\text{en}} = \{p_1^{\text{en}}, p_2^{\text{en}}, ..., p_{|D|}^{\text{en}}\}$, where p_i^{en} indicates the i-th paragraph in D^{en}. Inspired by Perez-Beltrachini and Lapata (2021), the lead paragraph of a Wikipedia page contains knowledge-intensive semantics. Thus, we use p_1^{en} from each randomly selected Wikipedia page as a source sentence to meet the requirement (1). In view of the paragraphs on the same Wikipedia page are generally highly related, to meet the requirement (2) for p_1^{en}, we randomly select its consecutive paragraphs, *i.e.*, $D^{\text{en}} \setminus p_1^{\text{en}}$, as its relevant document. To further collect relevant documents beyond English, *i.e.*, the requirement (3), we exploit the parallel documents in other languages of D^{en} provided by Wikipedia. In this work, we choose Chinese, German, French and Czech, and denote the corresponding parallel documents as D^{zh}, D^{de}, D^{fr} and D^{cs}, respectively. Given this, the consecutive paragraphs from $D^l \setminus p_1^l (l \in \{\text{zh}, \text{de}, \text{fr}, \text{cs}\})$ form as the relevant document in other languages.

To ensure robustness, for a small number of source sentences, we randomly select documents from the whole Wikipedia to serve as noisy documents. After the above process, we obtain 79K English source sentences, which are further split into training, validation and testing sets with 74.5K, 2.5K and 2K sentences. For each sentence, a (relevant or noisy) document in English, Chinese, German, French or Czech is also provided.

2.2 Translation Annotation

For a given source sentence, we next collect its translation in the target language conditioned on the corresponding document. In this work, we focus on English-to-Chinese translation, and we collect the Chinese translation for the 79K English sentences. Since the source sentences are lead paragraphs of Wikipedia's English pages (*i.e.*, p_1^{en}), one straightforward way is to directly use the counterparts of the Chinese parallel Wikipedia pages (*i.e.*, p_1^{zh}) as the translation. However, in Wikipedia, the parallel documents are not fully paired across different languages (Perez-Beltrachini and Lapata, 2021). Therefore, p_1^{zh} cannot be regarded as the translation of p_1^{en}. We also conduct a preliminary experiment to calculate the CometKiwi score (Rei et al., 2022) between the English lead paragraphs and the parallel Chinese ones, resulting in a low score (<60.0). Thus, we should annotate the Chinese translations of the 79K source sentences. Considering the trade-off between quality and cost, we decide to translate the source sentences of the training and validation sets via GPT-4o (OpenAI, 2024), while those of the testing set are translated via human translators.



[Format Controlling & Scoring Guideline]: […]

[Document]: Link aggregation addresses two problems with Ethernet connections: bandwidth limitations and lack of resilience. With regard to the first issue: bandwidth requirements do not scale linearly. Ethernet bandwidths historically have increased tenfold each …

[Source Sentence]: In computer networking, the term link aggregation refers to various methods of combining multiple network connections …

[Judgment]: The document provides detailed information about link aggregation, including its purpose of increasing bandwidth and providing redundancy, which aligns with the input sentence. This context helps ensure accurate translation of technical terms and concepts.

[Score]: 4

[Translation]: 在计算机网络中，链路聚合一词指的是通过并行组合多个网络连接来增加吞吐量…

Figure 1: The overview of GPT-4o translation.

GPT-4o translation. Given a source sentence and the corresponding document, we prompt GPT-4o to perform retrieval-augmented MT to collect the Chinese translation. To achieve better translation, we let GPT-4o first judge the relevance of the given document to respond with a judgment and a 5-point rating, and then translate the sentence to Chinese in a chain-of-thought (CoT) manner. Figure 1 gives a brief overview of the process. We also provide an example of the complete prompt, quality analysis and other details in Appendix A.

Human translation. For source sentences in the testing set, we employ 10 professional human translators to collect the Chinese translations. All translators are native Chinese, majoring in English, and have passed the translator qualification. We only provide the source English sentences to the annotators, and encourage them to search for the information they need from Wikipedia. In addition, there are three data reviewers with rich experience in checking translation quality, and 20% of the sentences translated by each translator are checked by a reviewer. If the translation accuracy is lower than 95%, the translator needs to modify all his/her translations under the guidance of the reviewer.

Finally, we obtain 79K retrieval-augmented MT samples. Among them, 77K samples from the training and validation sets are translated by the GPT-4o translator. Each sample can be formulated as a triplet $\langle s, d^l, t \rangle$, where s and t indicate the source English sentence and its Chinese translation, re-

Document		Training	Validation	Testing
Type	Lang.			
Relevant	En	19,500	500	2,000
	Zh	19,500	500	
	De	9,700	300	
	Fr	9,700	300	
	Cs	9,700	300	
Noisy	En	1,850	150	
	Zh	1,850	150	
	De	900	100	
	Fr	900	100	
	Cs	900	100	
Total		74,500	2,500	2,000

Table 1: The number of retrieval-augment MT samples in RAGtrans w.r.t. different types and different languages (Lang.) of documents.

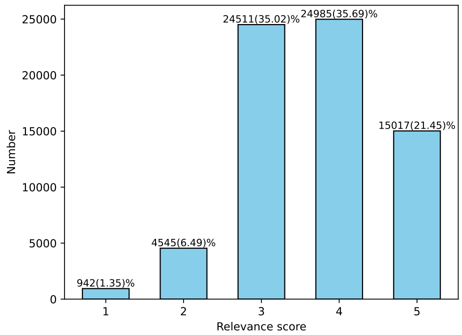


Figure 2: The distribution of relevance scores.

spectively. d^l indicates the given document for s, and $l \in \{\text{en}, \text{zh}, \text{de}, \text{fr}, \text{cs}\}$ represents its language. In addition, 2K samples from the testing set are translated by the human translators. Though we do not provide human translators with relevant documents, we provide the relevant English, Chinese and German documents (derived from the corresponding and parallel Wikipedia documents) in RAGtrans.[2] Thus, a testing sample could be formulated as a quintuple $\langle s, d^{\text{en}}, d^{\text{zh}}, d^{\text{de}}, t\rangle$.

2.3 Data Statistics

Table 1 shows the number of samples w.r.t. different types (relevant or noisy) and different languages of the given documents. In the training and validation sets, 8.59% and 24% of samples are associated with noisy documents. We emphasize the ratio of noisy documents in the validation set since robustness is vital in real applications. Moreover, for the training and validation samples, GPT-4o also outputs a 5-point rating (named relevance score) w.r.t. the given documents (c.f. right middle section in Figure 1). For samples with relevant documents, we also calculate the distribution of their relevance scores. As shown in Figure 2, more than 92% of the documents are regarded as "relevant" ($\geq$ 3) by GPT-4o. For samples with noisy documents, 99.93% (6,995/7,000) of samples are judged as "1", while the remaining 0.07% (5/7000) are "2".

	Min.	Max.	Avg.	95th ptcl.
Source (En)	5	526	85.83	173
Target (Zh)	6	669	100.26	202
Document (En)	2	3,254	326.08	874
Document (Zh)	2	4,456	349.65	925
Document (De)	33	1,065	367.43	791
Document (Fr)	29	3,481	365.09	902
Document (Cs)	38	962	369.84	769

Table 2: The minimum (Min.), maximum (Max.), average (Avg.) and 95th percentile (ptcl.) of tokens in the source sentence, target translation, and documents.

As for the length of source sentences, target sentences and documents, we use tiktoken[3] to calculate their token-level length. Table 2 shows the minimum, maximum, and average length. 95th percentile of length is also provided. We find that an extremely small number of documents only have single-digit tokens, which should be considered as noises, and we reserve these samples under the robustness consideration.

2.4 Benchmark Settings

We design three benchmark settings to evaluate the retrieval-augmented MT models: (1) **Golden Evaluation**: For each testing sample $\langle s, d^{\text{en}}, d^{\text{zh}}, d^{\text{de}}, t\rangle$, we give the source sentence (s) and a golden relevant document (d^{en}/d^{zh}/d^{de}) to the model, and evaluate models' translation. (2) **Robustness Evaluation**: We give s and an irrelevant document (randomly selected from Wikipedia) to the model, and evaluate its translation. (3) **Full Wiki Evaluation**: This setting equips the MT models with a retriever, and truly tests models' retrieval-augmented MT ability. For a given s, a retriever should first retrieve relevant documents from the whole Wikipedia, and then input both s and retrieved documents to the MT model to get translation.

3 Multi-Task Training

To further enhance LLMs' retrieval-augmented MT ability, we propose a multi-task training method, named **CSC**, which involves three designed train-

[2] French and Czech documents are not provided since only a small number of samples have French and Czech parallel documents in Wikipedia.

[3] https://github.com/openai/tiktoken

ing objectives, *i.e.*, **C**ross-lingual information completion, **S**elf-knowledge-enhanced translation and **C**ross-lingual relevance discrimination. In this section, we first introduce these objectives (§ 3.1) and then discuss how to create their training samples from existing corpora (§ 3.2).

3.1 Multi-Task Training Objectives

When developing a retrieval-augment MT model in real applications, it is possible to retrieve information from multilingual knowledge bases for a given source sentence. As a result, the model might receive multiple documents from various languages, extending beyond both the source and target languages. In such a situation, the challenge of effectively refining knowledge from these multilingual documents becomes increasingly significant. To this end, we design three training objectives:

(1) *Cross-lingual information completion.* Given a multilingual document d^{mix} whose paragraphs might be in different languages, and its truncated summary $\hat{y}$ in one language (*e.g.*, English), we require LLMs to expand $\hat{y}$ to a complete summary y. Formally, this objective can be formulated as $\Theta(y|d^{\text{mix}}, \hat{y})$, where Θ denotes the LLMs.

(2) *Self-knowledge-enhanced translation.* As revealed by recent RAG studies (Wang et al., 2023b; Liu et al., 2024; Asai et al., 2024), RAG models can achieve better performance with the help of their own knowledge. Inspired by this idea, we design self-knowledge-enhanced translation. Specifically, given a source sentence s, LLMs first generate its relevant document $\tilde{d}^l$ in a specific language $l \in \{\text{en}, \text{zh}, \text{de}, \text{fr}, \text{cs}\}$ and then incorporate the document to translate s to t, denoted as $\Theta(t|\tilde{d}^l|s)$.

(3) *Cross-lingual relevance discrimination.* Given that the retrieved documents may be in various languages, a crucial capability is to assess the relevance between two texts in different languages. To this end, given a document pair $\langle d^{l_1}, d^{l_2} \rangle$ ($l_1 \neq l_2$), l_1 and l_2 denote the languages of the documents, the model is required to generate the relevance between d^{l_1} and d^{l_2}, denotes as $\text{r}(d^{l_1}, d^{l_2})$. The object can be formulated as $\Theta(r|d^{l_1}, d^{l_2})$

3.2 Multi-Task Training Samples

To create the samples for these training objectives, a principle is to reformulate existing corpora instead of labeling new data to ensure scalability.

(1) *Cross-lingual information completion.* To create the multilingual document d^{mix} and its summary y, we reformulate the Wikipedia corpus. As revealed by Perez-Beltrachini and Lapata (2021), the lead paragraph in a Wikipedia page could be regarded as its summary. Given this, for an English Wikipedia page D^{en}, we extract its lead paragraph (*i.e.*, p_1^{en}) as y, and randomly truncate y to $\hat{y}$. We next construct d^{mix} from the remaining paragraphs $\hat{D}^{\text{en}} = \{p_i^{\text{en}} | i \geq 2\}$, and the parallel counterparts in other languages, *i.e.*, $\hat{D}^{\text{zh}}$, $\hat{D}^{\text{de}}$, $\hat{D}^{\text{fr}}$ and $\hat{D}^{\text{cs}}$ in this work. Since there might be redundant information across parallel paragraphs, we use MMR algorithm (Carbonell and Goldstein, 1998) to select paragraphs from these multilingual paragraphs, *i.e.*, $\bigcup_l \hat{D}^l$, to form d^{mix}. MMR is a statistical algorithm that iteratively selects key paragraphs from the given document, at each selection step, it evaluates the relevance and redundancy of the unselected paragraphs in relation to the selected ones to determine which paragraph to select in that step.

(2) *Self-knowledge-enhanced translation.* We reformulate previous multilingual MT corpora to create samples. In detail, we use TED talk corpus (Aharoni et al., 2019), where each sentence is provided with multilingual parallel sentences. For an English sentence s^{en}, we input the sentence or its parallel sentences in other languages (*i.e.*, s^l) to a LLM Θ, and prompt Θ to generate its relevant knowledge in the corresponding languages, *i.e.*, $\tilde{d}^l$. In this way, $\tilde{d}^l$ could be used as a relevant document to translate s^{en} to other languages.

(3) *Cross-lingual relevance discrimination.* We reformulate the parallel Wikipedia documents to create the samples. Intuitively, randomly selected paragraphs from two parallel documents are relevant; while those from different documents are irrelevant. In this way, we create the document pair and the corresponding boolean relevance.

4 Experiments

4.1 Experimental Setup

Metrics. Following previous work, we adopt *BLEU* (Papineni et al., 2002) and reference-based *COMET* score (Rei et al., 2022). BLEU evaluates n-grams overlap between the generated translations and corresponding references, while COMET evaluates the semantic similarity of translations against references. Besides, recent studies (Kocmi and Federmann, 2023; Wang et al., 2023a) also show the strong ability of LLMs in NLP evaluation. Thus, we use evaluators implemented using GPT-4o in reference-based and reference-free styles, which we refer to as *GRB* and *GRF*, respectively.

		Zero-Shot LLMs				SFT LLMs				SFT+CSC LLMs			
		BLEU	COMET	GRB	GRF	BLEU	COMET	GRB	GRF	BLEU	COMET	GRB	GRF
w/ Empty Document													
1	Qwen2.5-7B	50.11	84.63	85.48	86.97	54.64	86.72	88.50	89.07	56.82	87.74	91.16	91.81
2	Qwen2.5-14B	51.34	84.71	86.20	87.65	55.43	86.93	89.12	89.42	57.59	87.88	91.62	92.04
3	LLama-3-8B	40.94	77.54	80.39	78.08	54.35	86.60	88.06	88.24	56.09	87.60	90.30	91.12
4	Mistral-7B	38.12	76.40	79.75	77.60	53.58	86.32	87.69	88.08	54.67	87.16	90.06	90.48
w/ Noisy Document													
5	Qwen2.5-7B	48.62	83.36	84.57	86.05	54.66	86.73	88.50	89.11	56.48	87.66	91.09	91.75
6	Qwen2.5-14B	49.40	83.66	84.90	86.49	55.47	86.97	89.17	89.38	57.74	87.91	91.64	92.09
7	LLama-3-8B	26.33	74.64	78.20	78.92	54.25	86.59	87.99	88.17	56.12	87.60	90.26	91.09
8	Mistral-7B	24.29	73.86	77.69	78.04	53.49	86.30	87.72	88.11	54.46	87.14	89.95	90.45
w/ Golden English Document													
9	Qwen2.5-7B	49.55	84.29	85.05	86.48	57.10	87.72	90.57	91.34	58.68	88.72	92.48	93.14
10	Qwen2.5-14B	50.58	84.47	85.47	87.10	56.95	87.90	91.10	91.76	59.43	88.85	92.71	93.26
11	LLama-3-8B	29.44	75.29	80.23	80.85	55.84	87.58	89.74	90.35	58.00	88.59	91.39	92.03
12	Mistral-7B	26.56	74.75	79.75	80.27	54.90	87.34	89.49	89.82	56.63	88.14	91.14	91.87
w/ Golden Chinese Document													
13	Qwen2.5-7B	49.93	84.40	85.39	86.72	57.25	87.07	91.19	91.50	59.97	89.10	93.06	93.57
14	Qwen2.5-14B	50.65	84.68	85.72	86.95	58.09	87.23	91.67	91.95	60.47	89.21	93.29	93.73
15	LLama-3-8B	35.74	76.39	80.82	81.04	57.09	87.04	90.20	90.81	59.47	88.99	92.84	93.15
16	Mistral-7B	34.86	75.21	79.82	80.43	56.36	87.76	89.93	90.33	58.78	88.75	92.58	93.02
w/ Golden German Document													
17	Qwen2.5-7B	44.17	83.15	84.25	85.07	55.59	87.20	88.93	89.57	58.68	88.74	92.39	93.07
18	Qwen2.5-14B	45.20	82.89	84.82	85.64	56.28	87.46	89.36	89.71	59.41	88.89	92.65	93.13
19	LLama-3-8B	35.27	74.83	78.46	78.93	55.13	87.08	88.62	89.13	57.72	88.56	91.27	91.73
20	Mistral-7B	34.98	73.29	77.32	76.45	54.34	86.81	88.28	88.79	56.62	88.14	91.02	91.67

Table 3: Experimental results of golden evaluation and robustness evaluation on RAGtrans. "SFT LLMs" denotes the LLMs are instruction-tuned on the training data of RAGtrans, while "SFT+CSC LLMs" denotes the LLMs are instruction-tuned on both RAGtrans and CSC multi-task training.

Backbones. We adopt four LLMs in the experiments: (1) Qwen2.5-7B-Instruct and (2) Qwen2.5-14B-Instruct (Yang et al., 2024) are two cutting-edge Qwen-series LLMs. (3) Llama-3-8B (Dubey et al., 2024) is the latest llama-series LLM. (4) Mistral-7B (Jiang et al., 2023) also shows great performance among the same-scale LLMs.

Retriever. To support the full Wiki evaluation in RAGtrans, we implement two retrievers in the experiments: (1) BM25 (Robertson et al., 2009) is a traditional lexical search method that matches keywords efficiently with an inverted index. For a given source sentence, BM25 can retrieve its relevant documents only in the same language. (2) BGE-m3 (Chen et al., 2024a) is a multilingual sentence embedding model that supports dense retrievals across different languages.

Implementation Details. Llama-Factory (Zheng et al., 2024) is used to instruct-tune LLMs. All LLMs are tuned on 8×NVIDIA A100 GPUs (40G) with 1e-5 learning rate and 32 (8×4) batch size. We use the DeepSpeed optimization (Rasley et al., 2020), and set ZeRO-2 optimization for Qwen2.5-7B-Instruct and Mistral-7B, while ZeRo-3 for Qwen2.5-14B-Instruct and Llama-3-8B. During tuning, documents are also truncated to ensure the input length is within 2K tokens. For more details about SFT prompts, model checkpoints, training hours, CSC multi-task training samples and metric implementation, please refer to Appendix B.

4.2 Main Results

Table 3 shows the main results of the golden and robustness evaluation settings. For each LLM, we evaluate its retrieval-augmented MT performance when giving empty, noisy or golden documents.

Zero-Shot Performance. Among all backbones, Qwen2.5-14B typically performs best in terms of all metrics followed by Qwen2.5-7B. When giving noisy documents to LLMs, the MT performance of all LLMs decreases compared with those of giving empty documents. For example, Qwen2.5-7B (w/ empty document) achieves 50.11 BLEU and 84.63 COMET, while the counterparts of Qwen2.5-7B (w/ noisy document) are 48.62 and 83.36. This observation indicates the *low robustness of zero-shot LLMs when faced with irrelevant documents*. Moreover, when zero-shot LLMs use golden relevant documents as inputs, their MT performances do not increase (compared with those using empty documents) as expected. Specifically, the performance slightly decreases with golden English or Chinese documents (rows 9-12 or rows 13-16 vs. rows 1-4); while significantly decreasing with golden German documents (rows 17-20 vs. rows 1-4). Thus, *the retrieval-augmented MT ability is limited in zero-shot LLMs, especially when the retrieved documents are in a language beyond the source and the target languages (named a third language).*

Instruction-Tuning Performance. After we tune LLMs on RAGtrans, their MT performance generally increases by a large margin. For example, when giving empty documents, SFT Qwen2.5-7B outperforms zero-shot Qwen2.5-7B by 4.53 BLEU, 2.09 COMET, 3.02 GRB and 2.10 GRF. In addition, we find that when using golden documents, SFT LLMs achieve better performance than those using empty documents, indicating that *instruct-tuning on RAGtrans improves LLMs' retrieval-augmented MT ability*. Even when giving the relevant documents in a third language, it can bring improvement. *The model robustness is also enhanced*, and the given noisy documents do not significantly perturb the model performance. This is because a small number of training samples in RAGtrans consist of irrelevant documents as inputs, thus, LLMs can learn to translate source sentences conditioned on both relevant and irrelevant documents.

CSC Training Performance. After instruction-tuning on RAGtrans and CSC multi-task training, the model performance is further improved. When giving empty documents, CSC brings 1.09∼2.18 BLEU and 0.84∼1.02 COMET improvements compared with SFT LLMs (rows 1-4). This observation verifies the effectiveness of CSC, and *LLMs' retrieval-augmented MT ability can be enhanced by the designed training objectives*. Besides, when giving the relevant documents in a third language (*i.e.*, German) to SFT LLMs, it brings 0.76∼0.95 BLEU and 0.48∼0.53 COMET improvements compared with when giving empty documents (rows 17-20 vs. rows 1-4); while the counterpart improvements in SFT+CSC LLMs are 1.63∼1.95 BLEU and 0.96∼1.01 COMET. Therefore, *CSC enhances LLMs' ability to leverage relevant knowledge in a third language*. Moreover, we discuss the scalability of CSC in Appendix C.

4.3 Full Wiki Evaluation

Table 4 shows the experimental results on full Wiki evaluation. We use the Wikipedia dumps in different languages as the knowledge sources for retrievers to retrieve relevant documents, and then leverage SFT+CSC LLMs to translate the source sentences (please refer to Appendix D for more details). Compared with using empty documents, retrieving documents from knowledge sources generally brings improvement, indicating the usability of retrieved knowledge. Compared with the BM25 retriever, BGEm3 retriever helps LLMs achieve better performance (rows 5-6 vs. rows 3-4). Besides, BGEm3 could retrieve knowledge from other languages, and the retrieved knowledge from a third language could also enhance model performance, verifying the SFT+CSC LLMs could leverage multilingual knowledge in retrieval-augmented MT.

#	Knowledge	Method	BLEU	COMET
1	Empty	Qwen2.5-7B (+BM25)	56.82	87.74
2	Document	Qwen2.5-14B (+BM25)	57.59	87.88
3	English	Qwen2.5-7B (+BM25)	57.39	87.95
4	Wikipedia	Qwen2.5-14B (+BM25)	58.04	88.16
5	English	Qwen2.5-7B (+BGEm3)	57.91	88.12
6	Wikipedia	Qwen2.5-14B (+BGEm3)	58.74	88.40
7	Chinese	Qwen2.5-7B (+BGEm3)	58.33	88.29
8	Wikipedia	Qwen2.5-14B (+BGEm3)	59.02	88.51
9	German	Qwen2.5-7B (+BGEm3)	57.85	88.11
10	Wikipedia	Qwen2.5-14B (+BGEm3)	58.69	88.34
11	French	Qwen2.5-7B (+BGEm3)	57.63	87.98
12	Wikipedia	Qwen2.5-14B (+BGEm3)	58.34	88.25
13	Czech	Qwen2.5-7B (+BGEm3)	57.29	87.86
14	Wikipedia	Qwen2.5-14B (+BGEm3)	57.87	88.06

Table 4: Experimental results of full wiki evaluation on SFT+CSC LLMs.

	BLEU	COMET	BLEU	COMET
	w/ Empty Doc.		w/ Golden En.	
Qwen2.5-7B (SFT+CSC)	56.82	87.74	58.68	88.72
- CLIC	56.49	87.62	57.92	88.34
- SKET	56.10	87.52	57.45	88.11
- CLRD	56.54	87.68	58.25	88.51
	w/ Golden Zh.		w/ Golden De.	
Qwen2.5-7B (SFT+CSC)	59.97	89.10	58.68	88.74
- CLIC	59.48	88.89	56.95	88.15
- SKET	58.74	88.46	57.88	88.49
- CLRD	59.70	89.02	56.67	87.94

Table 5: Ablation study on golden evaluation. Doc.: Document; "En.", "Zh." and "De." indicate English, Chinese and German documents, respectively.

4.4 Ablations

As shown in Table 5, we conduct ablation studies on RAGtrans to evaluate the contributions of each training objective in CSC. Specifically, we remove each objective, and evaluate the model performance accordingly. In each case, the performance is lower than using all training objectives, indicating the effectiveness of every objective. When giving empty documents or documents in source/target language, the most important objective is self-knowledge-enhanced translation (SKET) since it enhances models' MT ability. When giving documents in a third language, cross-lingual information comple-

	Method		Error Type (%)				
#	Model	Document	Ref.	Word	Phrase	Fluency	Other
1	Qwen2.5-7B (SFT)	Empty	0.50	7.00	3.33	1.50	0.50
2	Qwen2.5-14B (SFT)	Empty	0.33	5.50	2.67	1.33	0.50
3	Qwen2.5-7B (SFT+CSC)	Empty	0.50	5.50	2.67	1.50	0.50
4	Qwen2.5-14B (SFT+CSC)	Empty	0.50	4.67	2.33	1.00	0.17
5	Qwen2.5-7B (SFT)	Golden Zh.	0.33	5.17	2.50	1.33	0.50
6	Qwen2.5-14B (SFT)	Golden Zh.	0.17	4.67	2.33	1.33	0.00
7	Qwen2.5-7B (SFT+CSC)	Golden Zh.	0.17	5.00	2.17	1.50	0.33
8	Qwen2.5-14B (SFT+CSC)	Golden Zh.	0.17	3.50	1.50	1.33	0.33

Table 6: Human evaluation of the retrieval-augmented MT ability. Ref.: Reference.

#	Document	Error Type (%)				
		Ref.	Word	Phrase	Fluency	Other
1	Empty	0.50	5.50	2.67	1.50	0.50
2	Noisy	1.83	6.67	5.33	1.83	1.50
3	Golden Chinese	0.17	5.00	2.17	1.50	0.33
4	Golden English	0.17	3.17	1.33	0.83	0.17
5	Golden German	0.50	8.17	1.67	1.33	0.33

Table 7: Human evaluation of the effects of documents, using Qwen2.5-7B (SFT+CSC) as the MT model.

tion (CLIC) and cross-lingual relevance discrimination (CLRD) become more important than SKET, since these two objectives train LLMs to refine and judge information from multilingual knowledge.

4.5 Human Evaluation

Retrieval-Augmented MT Ability. We employ human evaluation to further study the MT performance of SFT LLMs and SFT+CSC LLMs. Specifically, human evaluators judge whether the translations include the following flaws: reference errors, word-level errors, phrase-level errors, fluency flaws and other errors (more details are given in Appendix E). As shown in Table 6, the two most common error types are word-level and phrase-level errors. In these two types, the CSC multi-task training method and the golden documents enhance the model performance, verifying the effectiveness of CSC and the usability of the golden documents.

The Effects of Documents. To study the effects of documents, we provide the different documents to Qwen2.5-7B (SFT+CSC), including empty, noisy, golden English, golden Chinese and golden German documents, to evaluate if the corresponding translations involve flaws. As shown in Table 7, the noisy documents still increase the number of translation flaws (row 2 vs. row 1). Robustness is a crucial factor for the deployment of LLMs in real applications, thus *future work could pay more attention to model robustness*. Besides, we also find that when providing the model with golden Chinese or English documents, the number of translation flaws typically decreases. However, when providing golden German documents, the number of word-level errors significantly increases (row 5 vs. row 1). We further observe the cases, and find this is because *the German documents might encourage MT models to translate some entities to German if the entities listed in the documents, thus raising word-level errors*. This issue should also be noticed in future work, since the retrieval-argument MT models might receive documents from various languages in real applications.

5 Related Work

To leverage additional knowledge to enhance MT performance, previous literature typically explores paired sentences (also known as "translation memory") or structured knowledge graphs as the knowledge sources: (1) *Paired Sentences*: Zhang et al. (2018) utilize a search engine to retrieve sentence pairs whose source sides are similar to the input sentences. Bulte and Tezcan (2019) design a fuzzy retriever to enhance the model performance. He et al. (2021) design a fast and accurate method to improve the robustness of pair-sentence-enhanced MT models. Cai et al. (2021) relax the bilingualism limitation in retrieving paired sentences, and they try to retrieve similar target-language sentences to enhance MT models. (2) *Knowledge Graphs*: A few studies leverage relevant information from structured knowledge graphs to enhance MT models. Conia et al. (2024) use Wikidata (Vrandečić and Krötzsch, 2014), a multilingual knowledge graphs, to enhance MT models. Chen et al. (2024b) build an internal knowledge graph based on context, and then use it to enhance translation.

Different from previous work, we aim to utilize unstructured documents to provide supplementary knowledge to MT models. Furthermore, these documents can be in various languages and do not require any alignment across languages.

6 Conclusion

In this paper, we explore the retrieval-augmented MT with unstructured knowledge. To this end, we build RAGtrans dataset with 79K retrieval-augment MT samples to train and evaluate LLMs' retrieval-augmented MT ability. Further, we propose CSC multi-task training method with three designed objectives to teach LLMs to leverage multilingual knowledge in retrieval-augmented MT. Extensive experiments demonstrate the usability of RAGtrans and the effectiveness of CSC.

Limitations

While we show LLMs' retrieval-augmented MT ability and the effectiveness of CSC multi-task training method, there are some limitations worth noting: (1) We focus on English-to-Chinese translation in this work, and future work could extend the dataset and the method to other translation directions. (2) For multilingual knowledge bases, we use Wikipedia in some specific languages (*e.g.*, Chinese, English, German, French, and Czech). Future work could extend the multilingual sources to other languages or other sources. (3) During data collection of RAGtrans, a CoT prompt is used in the GPT-4o translation (c.f. Figure 1). However, in the SFT process, we do not use the CoT prompt to train LLMs, and future work could explore the effect of CoT in retrieval-augmented MT.

Ethical Considerations

We discuss the main ethical considerations of RAGtrans as follows: (1) *Licenses*. The source sentences and documents are derived from Wikipedia, whose texts are under CC BY-SA and GFDL licenses. We will release the RAGtrans dataset under CC-BY-SA 4.0 license. (2) *Compensation*. During the translation annotation, the salary for translating each sentence is determined by the average time of annotation and local labor compensation standard.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7307–7318, Online. Association for Computational Linguistics.

Jaime G Carbonell and Jade Goldstein. 1998. The use of mmr and diversity-based reranking for reodering documents and producing summaries.(1998). *Citado*, 4:24–42.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Meiqi Chen, Fandong Meng, Yingxue Zhang, Yan Zhang, and Jie Zhou. 2024b. Crat: A multi-agent framework for causality-enhanced reflective and retrieval-augmented translation with large language models. *arXiv preprint arXiv:2410.21067*.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3170–3180, Online. Association for Computational Linguistics.

Cuong Hoang, Devendra Sachan, Prashant Mathur, Brian Thompson, and Marcello Federico. 2023. Improving retrieval augmented neural machine translation by controlling source and fuzzy-match interactions. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 289–295, Dubrovnik, Croatia. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *24th Annual Conference of the European Association for Machine Translation*, page 193.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.

Yanming Liu, Xinyue Peng, Xuhong Zhang, Weihao Liu, Jianwei Yin, Jiannan Cao, and Tianyu Du. 2024. Ra-isf: Learning to answer and understand from retrieval augmentation via iterative self-feedback. *arXiv preprint arXiv:2403.06840*.

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Fernanda B Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023b. Self-knowledge guided retrieval augmentation for large language models. *arXiv preprint arXiv:2310.05002*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

A Details of GPT-4o translation

Complete Prompt. We provide the system prompt as follows: "*You are a professional translator, and your task is to translate a given input sentence from English to Chinese. In addition to the input sentence, you will be provided with a document that may contain relevant information to aid in the*

PROMPT

```
The document is provided as follows:
<document>
【Course à la direction du Parti québécois de 2005】Le, Bernard Landry, chef du Parti québécois, annonça son intention de démissionner de son poste de chef après avoir reçu un vote de confiance de 76,2 % de la part des délégués présents au Conseil national du parti se déroulant alors. Le Parti québécois désigna Louise Harel comme chef par intérim. La forte candidature de Gilles Duceppe, chef du Bloc québécois, le parti souverainiste au niveau fédéral, était attendue. Toutefois, malgré une importante pression, il y renonça en date du pour plutôt disputer les prochaines élections fédérales à la tête du Bloc québécois.
</document>

The sentence that needs to be translated is provided as follows:
<input sentence>
The Parti Québécois leadership election of 2005 was held from November 13 to November 15, 2005 to elect the new leader of the Parti Québécois, the main sovereigntist and social democratic political party in Quebec, Canada.
</input sentence>

Please response in the following JSON format:
{
  "is_helpful": {
    "reason": str,
    "score": int
  },
  "translation": str
}
Please carefully read the explanation of these keys:
- "is_helpful": Dict type. Whether the given document is helpful for translating the sentence.
  - "reason": Str type. The reason of whether the given document is helpful.
  - "score": Int type. 1 - 5 scores. 1 indicates the given document is totally irrelevant to the sentence. 3 indicates that though the document is relevant to the sentence, but it not helpful during the translation, you can still reach the translation result without seeing the document. 5 indicates the given document is very helpful during translation, you may make mistakes during translation without seeing the document.
- "translation": Str type. The translated result.
```

GPT-4o Response

```
{
  "is_helpful": {
    "reason": "The document provides context about the Parti Québécois leadership election in 2005, including key figures and events leading up to the election. However, it does not provide specific details about the dates or the nature of the election itself.",
    "score": 3
  },
  "translation": "2005年魁北克党领导人选举于2005年11月13日至11月15日举行，旨在选举魁北克党这一加拿大魁北克省主要的主权主义和社会民主主义政党的新领导人。"
}
```

Figure 3: An example of the complete prompt in GPT-4o translator.

translation. However, be aware that some documents may contain irrelevant or noisy information". An example of user prompt and model response is shown in Figure 3, where both a (French) document and an English source sentence are provided in the user prompt. We also define a 5-point rating breakdown to align the scoring value between GPT-4o and humans. In the model response, GPT-4o first judges the relevance between the given document and the source sentence, and then provides the corresponding translation.

Quality Analysis. To figure out the quality of GPT-4o translations, we calculate the reference-free CometKiwi score between the source English sentences and GPT-4o translations. As a result, the average score is 84.48, indicating high translation quality (Rei et al., 2022).

Other Details. The version of GPT-4o used in this work is *GPT-4o-2024-08-06*. When calling the official APIs, we set the temperature to 0.1, and set default values for other hyper-parameters.

B Implementation Details

SFT prompt. The system prompt in SFT is the same as the GPT-4o translator (c.f. Appendix A). The user prompt in SFT is provided as follows: "`<document>`[doc]`</document><input sentence>`[sent]`</input sentence>`", where `<document>`, `</document>`, `<input sentence>` and `</input sentence>` are special tokens to indicate the boundaries of the given document (denoted as "[doc]") and the source sentence ("[sent]").

Model Checkpoints. We use four LLM backbones in experiments, *i.e.*, Qwen2.5-7B-Instruct[4], Qwen2.5-14B-Instruct[5], Llama-3-8B[6] and Mistral-7B[7]. All model checkpoints are available at Huggingface.co community.

Training Hours. All experiments are conducted on NVIDIA A100 GPUs with 40G memory, and we use its GPU hours to denote the consumption of computing resources. We SFT LLMs on the training data of RAGtrans with 2 epochs, and each epoch costs 9.1 GPU hours, 54.0 GPU hours, 33.5 GPU hours, and 9.3 GPU hours for Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Llama-3-8B and Mistral-7B, respectively. For SFT+CSC LLMs, more GPU hours are costed. For example, to SFT Qwen2.5-14B-Instruct on both the RAGtrans training samples and CSC samples, each epoch costs

[4] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[5] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

[6] https://huggingface.co/hfl/llama-3-chinese-8b-instruct-v3

[7] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

```
PROMPT:

Please continue the English sentence based on the information provided in the given document.
<document>
一直以来,K联赛都没有引入升降级制度。2007年韩国足球协会试图将K联赛最后一名跟韩国全国联赛第一名之间实行升降级,但韩国全国联赛第一名以参加K联赛经费太高为由拒绝升级,韩国足球协会让K联赛跟韩国全国联赛之间实行升降级尝试失败。 The K League then took steps to create the new second division, mainly with the addition of a split-system during the 2012 K-League season in which the bottom clubs are placed in a competition for safety with the last placed club being relegated to the new second division (originally it was going to be two clubs relegated but the withdrawal of Sangju Sangmu FC meant only one would be relegated). Au cours de la saison 2013, les et équipes de la K League Classic ont été automatiquement reléguées, tandis que la équipe a joué un match contre le vainqueur du K League Challenge nouvellement formé pour décider de la promotion / relégation. 2012 wurden erstmals die neuen Mitglieder der 2. The promotion playoffs are as follows: the fourth-placed team plays against the third-placed team, then the winner of this match plays with the second-placed team. Si le match est à égalité, l'équipe la mieux classée avance. Le 3 janvier 2013, le nom officiel de la deuxième division a été annoncé comme la K League tandis que la K League originale a été changée en K League Classic avec le nouveau logo. This name change caused some degree of confusion and controversy, and on 11 March 2013 the official name was changed to K League Challenge. On 22 January 2018 the official name was changed to K League 2.
</document>
<English sentence>
The K League 2 is a South Korean professional association football league.
</English sentence>

OUTPUT:

Being the second tier of the South Korean football league system, it is currently contested by ten clubs, and operates on a system of promotion and relegation with the K League 1.
```

(a) Cross-lingual Information Completion

```
PROMPT:

You are a professional translator, and your task is to translate an given input sentence from English to Chinese. In addition to the input sentence, you will be provided with a document that may contain relevant information to aid in the translation. However, be aware that some documents may contain irrelevant or noisy information.

<document>
1. **牡蛎的生态作用**:
  - 牡蛎是一种重要的海洋生物，不仅为人类提供食物资源，还在生态系统中扮演着关键角色。它们能够通过过滤水中的浮游植物和有机物来净化水质，有助于维持海洋环境的健康。
  - 牡蛎还能够形成天然的礁石结构，这些礁石为许多其他海洋生物提供了栖息地，增加了生物多样性。
2. **牡蛎的生理特性**:
  - 牡蛎的身体具有一定的适应性，可以在特定条件下改变体液的浓度，使身体变得稀薄，从而更容易与其他牡蛎粘合在一起，形成坚固的礁石结构。
  - 这种粘合能力是牡蛎生存策略的一部分，有助于它们在潮间带等环境中抵御水流和波浪的冲击。
3. **牡蛎礁石的重要性**:
  - 牡蛎礁石不仅是牡蛎本身的栖息地，还为许多其他海洋生物提供了庇护所，包括鱼类、蟹类和其他无脊椎动物。
  - 这些礁石还能减少海岸侵蚀，保护沿海地区免受风暴和海浪的侵袭，对
</document>
<input sentence>
Finally, oysters can attenuate and agglomerate onto each other and form these amazing natural reef structures.
</input sentence>

OUTPUT:

最后，牡蛎可以减弱并聚集在一起，形成这些令人惊叹的自然礁石结构。
```

(b) Self-knowledge-enhanced Translation

```
PROMPT:

You are a professional translator responsible for assessing the relevance of a provided document in relation to a source input sentence. Please response “relevant” or “irrelevant”.

<document>
软件的RAM disk使用一般的RAM的一部分来模拟一个硬盘。通常都可以使用由操作系统内核支持的机制来实现,但是通过用户空间(相对于“内核空间”)的应用程序进程来进行这种的模拟也是有可能的。 因为存贮在RAM disk中文件一般都是临时性的,所以也就不需要使用电池备份,或者UPS等设备。某些RAM disk使用压缩的文件系统例如:cramfs。这是因为一般的RAM disk都容量一般都较小, 且RAM的价格比硬盘要高,所以这样做是很合理的。
</document>

<input sentence>
A RAM drive is a block of random-access memory that a computer's software is treating as if the memory were a disk drive . It is sometimes referred to as a virtual RAM drive or software RAM drive to distinguish it from a hardware RAM drive that uses separate hardware containing RAM, which is a type of battery-backed solid-state drive.
</input sentence>

OUTPUT:

relevant
```

(c) Cross-lingual Relevance Discrimination

Figure 4: Examples of CSC training objectives. Different colors in (a) means different langauges, including Chinese, English, French and German.

```
Score the following translation from English to Chinese with respect to the human reference on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

English source: [src]
Chinese human reference: [ref]
English translation: [hyp]

Score:
```

```
Score the following translation from English to Chinese on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

English source: [src]
English translation: [hyp]

Score:
```

Figure 5: The prompts in GRB (upper part) and GRF (lower part). “[src]”, “[ref]” and “[hyp]” denote the source sentence, human translation and model translation, respectively.

208 GPU hours; while the counterparts of Qwen2.5-7B-Instruct, Llama-3-8B and Mistral-7B are 34.1, 128.0 and 33.9 GPU hours, respectively.

Multi-Task Training Samples. As we introduce in Section 3, there are three training objectives in CSC multi-task training method. To provide a deeper understanding of these objectives, here we give some example samples in Figure 4. In our main experiments, we create 40K samples for each training objective.

Metric Implementation. To calculate the reference-based COMET score (Rei et al., 2022), we leverage the official codes[8] and the official model[9]. To calculate the BLEU score, we use the *sacrebleu* toolkit[10] to calulate the corpus-level BLEU. For GRB and GRF, we prompt GPT-4o (2024-08-06 version) as the MT evaluator in the reference-based and reference-free manners, respectively. The corresponding prompts borrow from Kocmi and Federmann (2023), and are illustrated in Figure 5. Since GRB and GRF need the API costs, we randomly select 200 samples from the RAGtrans testing set, and conduct the GRB and GRF evaluation. All experimental results listed in this paper are the average of 3 runs.

C Scalability of CSC

As we demonstrate the effectiveness of CSC multi-task training method in experiments, we wonder the upper limit of the improvement brought by CSC. To this end, we use Qwen2.5-7B-Instruct as the backbone, and systematically vary the number of CSC samples during the instruction tuning to examine the resulting performance (w/ golden Chinese document) changes. As shown in Figure 6, when the number of CSC samples exceeds 120K, the improvement brought by CSC begins to plateau. When the number of CSC samples increases from

[8] https://github.com/Unbabel/COMET

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

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

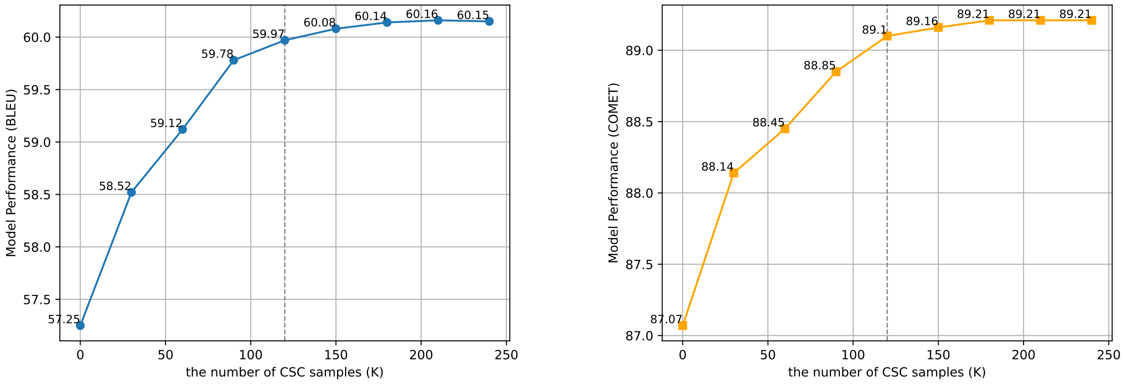


Figure 6: The experimental results of CSC scalability.

210K to 240K, the model performance does not improve accordingly.

D Details of Full Wiki Testing

Retriever. For BM25 retriever, we use the implementation of *elasticsearch*[11] toolkit to retrieve top-3 documents for each source sentence. For BGE-m3 retriever, we first use BGE-m3 sentence embedding model[12] to calculate the embedding of all documents in knowledge sources, and then use the embedding of source sentence to retrieve top-3 relevant documents via FAISS (Johnson et al., 2019).

Knowledge Sources. We use Wikipedia dumps (20241001 version) as the knowledge sources, and leverage *wikiextractor*[13] toolkit to extract articles from Wikipedia dumps. Following Karpukhin et al. (2020), we split each article into multiple, disjoint text blocks of 100 words as passages, serving as our basic retrieval units. In the full Wiki evaluation, we build the knowledge sources based on English, Chinese, German, French, Czech, Russian, Korean and Japanese Wikipedia dumps, resulting in tens of millions of retrieval units.

E Details of Human Evaluation

Evaluators. Three master students are recruited in our human evaluation, and they are fluent in both Chinese and English.

Instruction. The human evaluators are provided with the instructions for each translation error type: (1) Reference Errors: Are there any mistakes in pronoun or reference usage that could cause confusion about what or whom is being referred to? (2) Word-Level Errors: Are there incorrect translations, omissions, or additions of individual words that alter the meaning of the text? (3) Phrase-Level Errors: Are there incorrect translations, omissions, or additions of phrases that affect the overall coherence and accuracy of the translation? (4) Fluency Issues: Does the translation flow smoothly, or are there awkward phrases or constructions that impede comprehension? (5) Other Errors: Are there any additional errors present in the translation that do not fit into the categories above?

Evaluation Samples. Since human evaluation is labor-intensive, we randomly select 200 samples from the testing set of RAGtrans to conduct the human evaluation.

Inter-agreement. The Fleiss' Kappa scores (Fleiss, 1971) of the five error types are 0.63, 0.57, 0.68, 0.75 and 0.66 in our human evaluation, respectively, indicating a good inter-agreement among our evaluators.

[11] https://github.com/elastic/elasticsearch

[12] https://huggingface.co/BAAI/bge-m3

[13] https://github.com/attardi/wikiextractor